

Gabriele Meiselwitz (Ed.)

LNCS 12194

# Social Computing and Social Media

Design, Ethics, User Behavior,  
and Social Network Analysis

12th International Conference, SCSM 2020

Held as Part of the 22nd HCI International Conference, HCII 2020  
Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I

1  
Part I



 Springer

## Founding Editors

Gerhard Goos

*Karlsruhe Institute of Technology, Karlsruhe, Germany*

Juris Hartmanis

*Cornell University, Ithaca, NY, USA*

## Editorial Board Members

Elisa Bertino

*Purdue University, West Lafayette, IN, USA*

Wen Gao

*Peking University, Beijing, China*

Bernhard Steffen 

*TU Dortmund University, Dortmund, Germany*

Gerhard Woeginger 

*RWTH Aachen, Aachen, Germany*

Moti Yung

*Columbia University, New York, NY, USA*

More information about this series at <http://www.springer.com/series/7409>

Gabriele Meiselwitz (Ed.)

# Social Computing and Social Media

Design, Ethics, User Behavior,  
and Social Network Analysis

12th International Conference, SCSM 2020

Held as Part of the 22nd HCI International Conference, HCII 2020

Copenhagen, Denmark, July 19–24, 2020

Proceedings, Part I

*Editor*  
Gabriele Meiselwitz  
Towson University  
Towson, MD, USA

ISSN 0302-9743                      ISSN 1611-3349 (electronic)  
Lecture Notes in Computer Science  
ISBN 978-3-030-49569-5              ISBN 978-3-030-49570-1 (eBook)  
<https://doi.org/10.1007/978-3-030-49570-1>

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

© Springer Nature Switzerland AG 2020

The chapter “Automatic Versus Manual Forwarding in Web Surveys - A Cognitive Load Perspective on Satisficing Responding” is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>). For further details see license information in the chapter.

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Foreword

The 22nd International Conference on Human-Computer Interaction, HCI International 2020 (HCII 2020), was planned to be held at the AC Bella Sky Hotel and Bella Center, Copenhagen, Denmark, during July 19–24, 2020. Due to the COVID-19 coronavirus pandemic and the resolution of the Danish government not to allow events larger than 500 people to be hosted until September 1, 2020, HCII 2020 had to be held virtually. It incorporated the 21 thematic areas and affiliated conferences listed on the following page.

A total of 6,326 individuals from academia, research institutes, industry, and governmental agencies from 97 countries submitted contributions, and 1,439 papers and 238 posters were included in the conference proceedings. These contributions address the latest research and development efforts and highlight the human aspects of design and use of computing systems. The contributions thoroughly cover the entire field of human-computer interaction, addressing major advances in knowledge and effective use of computers in a variety of application areas. The volumes constituting the full set of the conference proceedings are listed in the following pages.

The HCI International (HCII) conference also offers the option of “late-breaking work” which applies both for papers and posters and the corresponding volume(s) of the proceedings will be published just after the conference. Full papers will be included in the “HCII 2020 - Late Breaking Papers” volume of the proceedings to be published in the Springer LNCS series, while poster extended abstracts will be included as short papers in the “HCII 2020 - Late Breaking Posters” volume to be published in the Springer CCIS series.

I would like to thank the program board chairs and the members of the program boards of all thematic areas and affiliated conferences for their contribution to the highest scientific quality and the overall success of the HCI International 2020 conference.

This conference would not have been possible without the continuous and unwavering support and advice of the founder, Conference General Chair Emeritus and Conference Scientific Advisor Prof. Gavriel Salvendy. For his outstanding efforts, I would like to express my appreciation to the communications chair and editor of HCI International News, Dr. Abbas Moallem.

July 2020

Constantine Stephanidis

# **HCI International 2020 Thematic Areas and Affiliated Conferences**

Thematic areas:

- HCI 2020: Human-Computer Interaction
- HIMI 2020: Human Interface and the Management of Information

Affiliated conferences:

- EPCE: 17th International Conference on Engineering Psychology and Cognitive Ergonomics
- UAHCI: 14th International Conference on Universal Access in Human-Computer Interaction
- VAMR: 12th International Conference on Virtual, Augmented and Mixed Reality
- CCD: 12th International Conference on Cross-Cultural Design
- SCSM: 12th International Conference on Social Computing and Social Media
- AC: 14th International Conference on Augmented Cognition
- DHM: 11th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management
- DUXU: 9th International Conference on Design, User Experience and Usability
- DAPI: 8th International Conference on Distributed, Ambient and Pervasive Interactions
- HCIBGO: 7th International Conference on HCI in Business, Government and Organizations
- LCT: 7th International Conference on Learning and Collaboration Technologies
- ITAP: 6th International Conference on Human Aspects of IT for the Aged Population
- HCI-CPT: Second International Conference on HCI for Cybersecurity, Privacy and Trust
- HCI-Games: Second International Conference on HCI in Games
- MobiTAS: Second International Conference on HCI in Mobility, Transport and Automotive Systems
- AIS: Second International Conference on Adaptive Instructional Systems
- C&C: 8th International Conference on Culture and Computing
- MOBILE: First International Conference on Design, Operation and Evaluation of Mobile Communications
- AI-HCI: First International Conference on Artificial Intelligence in HCI

## Conference Proceedings Volumes Full List

1. LNCS 12181, Human-Computer Interaction: Design and User Experience (Part I), edited by Masaaki Kurosu
2. LNCS 12182, Human-Computer Interaction: Multimodal and Natural Interaction (Part II), edited by Masaaki Kurosu
3. LNCS 12183, Human-Computer Interaction: Human Values and Quality of Life (Part III), edited by Masaaki Kurosu
4. LNCS 12184, Human Interface and the Management of Information: Designing Information (Part I), edited by Sakae Yamamoto and Hirohiko Mori
5. LNCS 12185, Human Interface and the Management of Information: Interacting with Information (Part II), edited by Sakae Yamamoto and Hirohiko Mori
6. LNAI 12186, Engineering Psychology and Cognitive Ergonomics: Mental Workload, Human Physiology, and Human Energy (Part I), edited by Don Harris and Wen-Chin Li
7. LNAI 12187, Engineering Psychology and Cognitive Ergonomics: Cognition and Design (Part II), edited by Don Harris and Wen-Chin Li
8. LNCS 12188, Universal Access in Human-Computer Interaction: Design Approaches and Supporting Technologies (Part I), edited by Margherita Antona and Constantine Stephanidis
9. LNCS 12189, Universal Access in Human-Computer Interaction: Applications and Practice (Part II), edited by Margherita Antona and Constantine Stephanidis
10. LNCS 12190, Virtual, Augmented and Mixed Reality: Design and Interaction (Part I), edited by Jessie Y. C. Chen and Gino Fragomeni
11. LNCS 12191, Virtual, Augmented and Mixed Reality: Industrial and Everyday Life Applications (Part II), edited by Jessie Y. C. Chen and Gino Fragomeni
12. LNCS 12192, Cross-Cultural Design: User Experience of Products, Services, and Intelligent Environments (Part I), edited by P. L. Patrick Rau
13. LNCS 12193, Cross-Cultural Design: Applications in Health, Learning, Communication, and Creativity (Part II), edited by P. L. Patrick Rau
14. LNCS 12194, Social Computing and Social Media: Design, Ethics, User Behavior, and Social Network Analysis (Part I), edited by Gabriele Meiselwitz
15. LNCS 12195, Social Computing and Social Media: Participation, User Experience, Consumer Experience, and Applications of Social Computing (Part II), edited by Gabriele Meiselwitz
16. LNAI 12196, Augmented Cognition: Theoretical and Technological Approaches (Part I), edited by Dylan D. Schmorrow and Cali M. Fidopiastis
17. LNAI 12197, Augmented Cognition: Human Cognition and Behaviour (Part II), edited by Dylan D. Schmorrow and Cali M. Fidopiastis



18. LNCS 12198, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Posture, Motion and Health (Part I), edited by Vincent G. Duffy
19. LNCS 12199, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Human Communication, Organization and Work (Part II), edited by Vincent G. Duffy
20. LNCS 12200, Design, User Experience, and Usability: Interaction Design (Part I), edited by Aaron Marcus and Elizabeth Rosenzweig
21. LNCS 12201, Design, User Experience, and Usability: Design for Contemporary Interactive Environments (Part II), edited by Aaron Marcus and Elizabeth Rosenzweig
22. LNCS 12202, Design, User Experience, and Usability: Case Studies in Public and Personal Interactive Systems (Part III), edited by Aaron Marcus and Elizabeth Rosenzweig
23. LNCS 12203, Distributed, Ambient and Pervasive Interactions, edited by Norbert Streitz and Shin'ichi Konomi
24. LNCS 12204, HCI in Business, Government and Organizations, edited by Fiona Fui-Hoon Nah and Keng Siau
25. LNCS 12205, Learning and Collaboration Technologies: Designing, Developing and Deploying Learning Experiences (Part I), edited by Panayiotis Zaphiris and Andri Ioannou
26. LNCS 12206, Learning and Collaboration Technologies: Human and Technology Ecosystems (Part II), edited by Panayiotis Zaphiris and Andri Ioannou
27. LNCS 12207, Human Aspects of IT for the Aged Population: Technologies, Design and User Experience (Part I), edited by Qin Gao and Jia Zhou
28. LNCS 12208, Human Aspects of IT for the Aged Population: Healthy and Active Aging (Part II), edited by Qin Gao and Jia Zhou
29. LNCS 12209, Human Aspects of IT for the Aged Population: Technology and Society (Part III), edited by Qin Gao and Jia Zhou
30. LNCS 12210, HCI for Cybersecurity, Privacy and Trust, edited by Abbas Moallem
31. LNCS 12211, HCI in Games, edited by Xiaowen Fang
32. LNCS 12212, HCI in Mobility, Transport and Automotive Systems: Automated Driving and In-Vehicle Experience Design (Part I), edited by Heidi Krömker
33. LNCS 12213, HCI in Mobility, Transport and Automotive Systems: Driving Behavior, Urban and Smart Mobility (Part II), edited by Heidi Krömker
34. LNCS 12214, Adaptive Instructional Systems, edited by Robert A. Sottilare and Jessica Schwarz
35. LNCS 12215, Culture and Computing, edited by Matthias Rauterberg
36. LNCS 12216, Design, Operation and Evaluation of Mobile Communications, edited by Gavriel Salvendy and June Wei
37. LNCS 12217, Artificial Intelligence in HCI, edited by Helmut Degen and Lauren Reinerman-Jones

38. CCIS 1224, HCI International 2020 Posters - Part I, edited by Constantine Stephanidis and Margherita Antona
39. CCIS 1225, HCI International 2020 Posters - Part II, edited by Constantine Stephanidis and Margherita Antona
40. CCIS 1226, HCI International 2020 Posters - Part III, edited by Constantine Stephanidis and Margherita Antona

**<http://2020.hci.international/proceedings>**



# 12th International Conference on Social Computing and Social Media (SCSM 2020)

Program Board Chair: **Gabriele Meiselwitz, Towson University, USA**

- Sarah Alhumoud, Saudi Arabia
- Andria Andriuzzi, France
- Francisco Javier Álvarez Rodríguez, Mexico
- Karine Berthelot-Guiet, France
- James Braman, USA
- Adheesh Budree, South Africa
- Adela Coman, Romania
- Isabelle Dorsch, Germany
- Panagiotis Germanakos, Germany
- Tamara Heck, Germany
- Hung-Hsuan Huang, Japan
- Aylin Ilhan, Germany
- Carsten Kleiner, Germany
- Ana I. Molina Díaz, Spain
- Takashi Namatame, Japan
- Hoang D. Nguyen, Singapore
- Kohei Otake, Japan
- Carlos Alberto Peláez, Colombia
- Daniela Quiñones, Chile
- Cristian Rusu, Chile
- Christian W. Scheiner, Germany
- Simona Vasilache, Japan
- Giovanni Vincenti, USA
- Yuanqiong Wang, USA
- Brian Wentz, USA

The full list with the Program Board Chairs and the members of the Program Boards of all thematic areas and affiliated conferences is available online at:

<http://www.hci.international/board-members-2020.php>



# HCI International 2021

The 23rd International Conference on Human-Computer Interaction, HCI International 2021 (HCII 2021), will be held jointly with the affiliated conferences in Washington DC, USA, at the Washington Hilton Hotel, July 24–29, 2021. It will cover a broad spectrum of themes related to Human-Computer Interaction (HCI), including theoretical issues, methods, tools, processes, and case studies in HCI design, as well as novel interaction techniques, interfaces, and applications. The proceedings will be published by Springer. More information will be available on the conference website: <http://2021.hci.international/>.

General Chair

Prof. Constantine Stephanidis

University of Crete and ICS-FORTH

Heraklion, Crete, Greece

Email: [general\\_chair@hci2021.org](mailto:general_chair@hci2021.org)

<http://2021.hci.international/>



# Contents – Part I

## Design Issues in Social Computing

|  |     |
|--|-----|
| Image Strength and Identity Diffusion as Factors Influencing the Perception of Hospitals by Their Facebook Communities . . . . .                                 | 3   |
| <i>Michael Beier and Sebastian Früh</i>  |     |
| The Ideal Topic: Interdependence of Topic Interpretability and Other Quality Features in Topic Modelling for Short Texts . . . . .                               | 19  |
| <i>Ivan S. Blekanov, Svetlana S. Bodrunova, Nina Zhuravleva, Anna Smoliarova, and Nikita Tarasov</i>   |     |
| Making Reproducible Research Simple Using RMarkdown and the OSF . . . .  | 27  |
| <i>André Calero Valdez</i>   |     |
| Visual Saliency: How Text Influences . . . . .   | 45  |
| <i>Ying Fang, Liyu Zhu, Xueni Cao, Liqun Zhang, and Xiaodong Li</i>  |     |
| Improving the Web Accessibility of a University Library for People with Visual Disabilities Through a Mixed Evaluation Approach . . . . .                        | 56  |
| <i>Milda Galkute, Luis A. Rojas P., and Victor A. Sagal M.</i>   |     |
| Utilization of Vanity to Promote Energy Saving Activities . . . . .  | 72  |
| <i>Kyoko Ito, Yasutaka Kishi, and Shogo Nishida</i>  |     |
| Verification of the Effect of Presenting a Virtual Front Vehicle on Controlling Speed. . . . .   | 81  |
| <i>Tetsuma Konishi, Takayoshi Kitamura, Tomoko Izumi, and Yoshio Nakatani</i>  |     |
| Roles on Corporate and Public Innovation Communities: Understanding Personas to Reach New Frontiers . . . . .  | 95  |
| <i>Maximilian Rapp, Niclas Kröger, and Samira Scheerer</i>   |     |
| Qualitative Evaluation of the Usability of a Web-Based Survey Tool to Assess Reading Comprehension and Metacognitive Strategies of University Students . . . . . | 110 |
| <i>Luis A. Rojas P., Maria Elena Truyol, Juan Felipe Calderon Maureira, Mayron Orellana Quiñones, and Anibal Puente</i>  |     |
| Automatic Versus Manual Forwarding in Web Surveys - A Cognitive Load Perspective on Satisficing Responding. . . . .  | 130 |
| <i>Arto Selkälä, Mario Callegaro, and Mick P. Couper</i>   |     |

|  |     |
|--|-----|
| A New Information Theory Based Clustering Fusion Method<br>for Multi-view Representations of Text Documents . . . . .                          | 156 |
| <i>Juan Zamora and Jérémie Sublime</i>   |     |
| Application of Visual Saliency in the Background Image Cutting<br>for Layout Design . . . . .  | 168 |
| <i>Liyu Zhu, Xueni Cao, Ying Fang, Liqun Zhang, and Xiaodong Li</i>  |     |
| Gamification Elements on Social Live Streaming Service<br>Mobile Applications . . . . .  | 184 |
| <i>Franziska Zimmer, Katrin Scheibe, and Hantian Zhang</i>   |     |
| <b>Ethics and Misinformation in Social Media</b>   |     |
| A Two-Phase Framework for Detecting Manipulation Campaigns<br>in Social Media . . . . .  | 201 |
| <i>Dennis Assenmacher, Lena Clever, Janina Susanne Pohl,<br/>Heike Trautmann, and Christian Grimme</i>   |     |
| Filter Bubbles and Content Diversity? An Agent-Based<br>Modeling Approach . . . . .  | 215 |
| <i>Poornima Belavadi, Laura Burbach, Patrick Halbach,<br/>Johannes Nakayama, Nils Plettenberg, Martina Ziefle,<br/>and André Calero Valdez</i> |     |
| The Law of Live Streaming: A Systematic Literature Review<br>and Analysis of German Legal Framework . . . . .                                  | 227 |
| <i>Kaja J. Fietkiewicz</i>   |     |
| Social Media Use, Political Polarization, and Social Capital:<br>Is Social Media Tearing the U.S. Apart? . . . . .                             | 243 |
| <i>James Hawdon, Shyam Ranganathan, Scotland Leman,<br/>Shane Bookhultz, and Tanushree Mitra</i>   |     |
| Designing an Experiment on Recognition of Political Fake News<br>by Social Media Users: Factors of Dropout . . . . .                           | 261 |
| <i>Olessia Koltsova, Yadviga Sinyavskaya, and Maxim Terpilovskii</i>   |     |
| Illicit Drug Purchases via Social Media Among American Young People . . . .  | 278 |
| <i>Atte Oksanen, Bryan Lee Miller, Iina Savolainen, Anu Sirola,<br/>Jakob Demant, Markus Kaakinen, and Izabela Zych</i>                        |     |
| I Do It Because I Feel that...Moral Disengagement and Emotions<br>in Cyberbullying and Cybervictimisation . . . . .                            | 289 |
| <i>Oronzo Parlangei, Enrica Marchigiani, Stefano Guidi,<br/>Margherita Bracci, Alessandro Andreadis, and Riccardo Zambon</i>                   |     |

|   |     |
|---|-----|
| The Effects of Thinking Styles and News Domain on Fake News Recognition by Social Media Users: Evidence from Russia . . . . .               | 305 |
| <i>Alexander Porshnev and Alexandre Miltsov</i>   |     |
| Using Deep Learning to Detect Rumors in Twitter . . . . .   | 321 |
| <i>Eliana Providel and Marcelo Mendoza</i>  |     |
| The Role of Moral Receptors and Moral Disengagement in the Conduct of Unethical Behaviors on Social Media. . . . .                          | 335 |
| <i>Christian W. Scheiner</i>  |     |
| Catfishing: A Look into Online Dating and Impersonation . . . . .   | 349 |
| <i>Mariah Simmons and Joon Suk Lee</i>  |     |
| Riding the Wave of Misclassification: How We End up with Extreme YouTube Content . . . . .  | 359 |
| <i>Christian Stöcker and Mike Preuss</i>  |     |
| Characterizing Social Bots Spreading Financial Disinformation. . . . .  | 376 |
| <i>Serena Tardelli, Marco Avvenuti, Maurizio Tesconi, and Stefano Cresci</i>  |     |
| Cyber Risks in Social Media . . . . .   | 393 |
| <i>Linda R. Wilbanks</i>  |     |
| Misinformation in the Chinese Weibo . . . . .   | 407 |
| <i>Lu Xiao and Sijing Chen</i>  |     |
| Ethical, Legal and Security Implications of Digital Legacies on Social Media . . . . .  | 419 |
| <i>Paige Zaleppa and Alfreda Dudley</i>   |     |
| <b>User Behavior and Social Network Analysis</b>  |     |
| When Emotions Grow: Cross-Cultural Differences in the Role of Emotions in the Dynamics of Conflictual Discussions on Social Media . . . . . | 433 |
| <i>Svetlana S. Bodrunova, Kamilla Nigmatullina, Ivan S. Blekanov, Anna Smoliarova, Nina Zhuravleva, and Yulia Danilova</i>                  |     |
| The World of Museums and Web 2.0: Links Between Social Media and the Number of Visitors in Museums . . . . .                                | 442 |
| <i>Adela Coman, Ana-Maria Grigore, Andreea Ardelean, and Robert Maracine</i>  |     |
| Virtual Fitness Community: Online Behavior on a Croatian Fitness Forum. . .   | 459 |
| <i>Kristina Feldvari, Anita Dremel, and Snježana Stanarević Katavić</i>   |     |

|  |     |
|--|-----|
| An Examination of Gaze During Conversation for Designing<br>Culture-Based Robot Behavior . . . . .   | 475 |
| <i>Louisa Hardjasa and Atsushi Nakazawa</i>  |     |
| Investigation on the Fusion of Multi-modal and Multi-person<br>Features in RNNs for Detecting the Functional Roles of Group<br>Discussion Participants . . . . .         | 489 |
| <i>Hung-Hsuan Huang and Toyooki Nishida</i>  |     |
| Exploring Gaze Behaviour and Perceived Personality Traits . . . . .  | 504 |
| <i>Koki Ijuin and Kristiina Jokinen</i>  |     |
| Users of Fitbit Facebook Groups: A Gender- and Generation-Determined<br>Investigation of Their Motivation and Need . . . . .   | 513 |
| <i>Aylin Ilhan</i>   |     |
| Intelligent Auto Technologies Are Here, and Drivers Are Losing Control. . . .  | 534 |
| <i>Brian M. Jones</i>  |     |
| Emotions in Online Gambling Communities: A Multilevel<br>Sentiment Analysis . . . . .  | 542 |
| <i>Markus Kaakinen, Atte Oksanen, Anu Sirola, Iina Savolainen,<br/>and David Garcia</i>  |     |
| Analysis of the Exposing Media Pattern that Affect Accessing<br>Own Website . . . . .  | 551 |
| <i>Yuho Katagiri, Kohei Otake, and Takashi Namatame</i>  |     |
| Dynamic Properties of Information Diffusion Networks During<br>the 2019 Halle Terror Attack on Twitter . . . . .   | 568 |
| <i>Philipp Kessling, Bastian Kiessling, Steffen Burkhardt,<br/>and Christian Stöcker</i>   |     |
| Cultural Factors as Powerful Moderators of Romanian Students’<br>Adoption of Mobile Banking in Everyday Life . . . . .   | 583 |
| <i>Valentin Mihai Leoveanu, Mihaela Cornelia Sandu, and Adela Coman</i>  |     |
| Social Behaviour Understanding Using Deep Neural Networks:<br>Development of Social Intelligence Systems. . . . .  | 600 |
| <i>Ethan Lim Ding Feng, Zhi-Wei Neo, Aaron William De Silva, Kellie Sim,<br/>Hong-Ray Tan, Thi-Thanh Nguyen, Karen Wei Ling Koh, Wenru Wang,<br/>and Hoang D. Nguyen</i> |     |
| Materialism and Facebook Usage: Could Materialistic<br>and Non-materialistic Values Be Linked to Using Facebook Differently? . . . .                                     | 614 |
| <i>Roshan Rai, Jade Blocksidge, and Mei-I Cheng</i>  |     |



Analyzing #LasTesis Feminist Movement in Twitter Using Topic Models . . . 624  
*Sebastian Rodriguez, Héctor Allende-Cid, Cristian Gonzalez,  
Rodrigo Alfaro, Claudio Elortegui, Wenceslao Palma,  
and Pedro Santander*

User-Oriented Quality Estimation of Social News Systems and Its Content:  
Gender-Dependent Assessment of Reddit . . . . . 636  
*Katrin Scheibe and Franziska Zimmer*

Defining Network Borders on Instagram: The Case of Russian-Speaking  
Bloggers with Migration Background . . . . . 647  
*Anna Smoliarova, Konstantin Platonov, Ekaterina Sharkova,  
and Tamara Gromova*

Effects of Linguistic Proficiency and Conversation Topic  
on Listener’s Gaze in Triadic Conversation . . . . . 658  
*Ichiro Umata, Koki Ijuin, Tsuneo Kato, and Seiichi Yamamoto*

The Confidence in Social Media Platforms and Private Messaging . . . . . 669  
*Jukka Vuorinen, Aki Koivula, and Ilkka Koiranen*

**Author Index** . . . . . 683

## Contents – Part II

### Participation and Collaboration in Online Communities

|   |     |
|---|-----|
| Knowledge Sharing and Community Promotion in Online Health Communities: Examining the Relationship Between Social Support, Community Commitment, and Trust Transfer . . . . . | 3   |
| <i>Zaenal Abidin, Achmad Nizar Hidayanto, Dedi I. Inan, Amira Luthfia Fitriani, Atikah Zahrah Halim, M. Farhan Mardadi, and Rizkah Shalihah</i>                               |     |
| Compliment Rules or Compliments Rule? A Population-Level Study of Appearance Commenting Norms on Social Media . . . . .   | 16  |
| <i>Erica Åberg, Aki Koivula, Iida Kukkonen, Outi Sarpila, and Tero Pajunen</i>  |     |
| Understanding Open Collaboration of Wikipedia Good Articles . . . . .   | 29  |
| <i>Huichen Chou, Donghui Lin, Toru Ishida, and Naomi Yamashita</i>  |     |
| Federated Artificial Intelligence for Unified Credit Assessment . . . . .   | 44  |
| <i>Minh-Duc Hoang, Linh Le, Anh-Tuan Nguyen, Trang Le, and Hoang D. Nguyen</i>  |     |
| Exploring TikTok Use and Non-use Practices and Experiences in China . . . . .   | 57  |
| <i>Xing Lu, Zhicong Lu, and Changqing Liu</i>   |     |
| Building an Integrated Comment Moderation System – Towards a Semi-automatic Moderation Tool . . . . .   | 71  |
| <i>Dennis M. Riehle, Marco Niemann, Jens Brunk, Dennis Assenmacher, Heike Trautmann, and Jörg Becker</i>  |     |
| Understanding Moderation in Online Mental Health Communities . . . . .  | 87  |
| <i>Koustuv Saha, Sindhu Kiranmai Ernala, Sarmistha Dutta, Eva Sharma, and Munmun De Choudhury</i>   |     |
| User-Generated Short Video Content in Social Media. A Case Study of TikTok . . . . .  | 108 |
| <i>Aliaksandra Shutsko</i>  |     |
| Review of Electronic Word-of-Mouth Based on Bibliometrics . . . . .   | 126 |
| <i>Peihan Wen and Ruiquan Wang</i>  |     |

## Social Computing and User Experience

|  |     |
|--|-----|
| Identifying User Experiences for Decision-Making in Service Science . . . . .  | 147 |
| <i>Silvana Aciar, Mayela Coto, and Gabriela Aciar</i>  |     |
| Customer eXperience in e-Learning: A Systematic Mapping Study . . . . .  | 158 |
| <i>Iván Balmaceda Castro, Cristian Rusu, and Silvana Aciar</i>   |     |
| Customer eXperiences in Retail: Case Studies in Physical and Virtual Channels . . . . .  | 171 |
| <i>Camila Bascur, Cristian Rusu, and Daniela Quiñones</i>  |     |
| Evaluation of Customer eXperience and Behaviour: A Literature Review . . . . .   | 181 |
| <i>Sandra Cano, Cristian Rusu, and Daniela Quiñones</i>  |     |
| User eXperience Heuristics for National Park Websites . . . . .  | 193 |
| <i>Dania Delgado, Daniela Zamora, Daniela Quiñones, Cristian Rusu, Silvana Roncagliolo, and Virginica Rusu</i>   |     |
| Programmer eXperience: A Set of Heuristics for Programming Environments . . . . .  | 205 |
| <i>Jenny Morales, Cristian Rusu, Federico Botella, and Daniela Quiñones</i>  |     |
| Understanding User Needs and Customer eXperience in Tourism Area . . . . .   | 217 |
| <i>Luis Rojas, Daniela Quiñones, and Cristian Rusu</i>   |     |
| Customer eXperience in Valparaíso Hostels: Analyzing Tourists' Opinions . . . . .  | 226 |
| <i>Virginica Rusu, Cristian Rusu, Daniela Quiñones, Silvana Roncagliolo, Victoria Carvajal, and Martin Muñoz</i>   |     |
| Students' Perception on Customer eXperience: A Comparative Study . . . . .   | 236 |
| <i>Cristian Rusu, Virginica Rusu, Federico Botella, Daniela Quiñones, Bogdan Alexandru Urs, Ilie Urs, Jenny Morales, Sandra Cano, Silvana Aciar, and Iván Balmaceda Castro</i> |     |
| An Experimental Study on Promotion of Pro-Environmental Behavior Focusing on “Vanity” for Interactive Agent . . . . .  | 247 |
| <i>Mizuki Yamawaki, Kimi Ueda, Yoshiki Sakamoto, Hirotake Ishii, Hiroshi Shimoda, Kyoko Ito, Takuya Fujioka, Qinghua Sun, Yasuhiro Asa, and Takashi Numata</i>                 |     |

## Social Media Marketing and Consumer Experience

|   |     |
|---|-----|
| The Key Role of Social Media in Identifying Consumer Opinions for Building Sustainable Competitive Advantages . . . . . | 261 |
| <i>Armenia Androniceanu, Irina Georgescu, and Jani Kinnunen</i>   |     |

|  |     |
|--|-----|
| The Digital “Advertising Call”: An Archeology of Advertising Literacy. . . . .   | 278 |
| <i>Karine Berthelot-Guiet</i>  |     |
| Research on Computational Simulation of Advertising Posters<br>Visual Cognition. . . . .   | 295 |
| <i>Xueni Cao, Ying Fang, Liyu Zhu, Xiaodong Li, and Liqun Zhang</i>  |     |
| “Fail, Clickbait, Cringe, Cancel, Woke”: Vernacular Criticisms of Digital<br>Advertising in Social Media Platforms. . . . .  | 309 |
| <i>Gustavo Gomez-Mejia</i>   |     |
| A Study on the Similarity of Fashion Brands Using Consumer Relationship<br>and Consumer Sense . . . . .  | 325 |
| <i>Yuzuki Kitajima, Kohei Otake, and Takashi Namatame</i>  |     |
| Analysis of Consumer Community Structure and Characteristic<br>Within Social Media . . . . .   | 336 |
| <i>Shin Miyake, Kohei Otake, and Takashi Namatame</i>  |     |
| Exploring Advertising Literacy Digital Paths: Comparison Between Gender<br>Approaches Among Chilean Students . . . . .   | 355 |
| <i>Claudia Montero-Liberona, Gianluigi Pimentel Varas,<br/>and Gregorio Fernández Valdés</i>   |     |
| Comparison of the Purchasing Behavior for Oneself or Other Using Eye<br>Tracking Gaze Data. . . . .  | 374 |
| <i>Mei Nonaka, Kohei Otake, and Takashi Namatame</i>   |     |
| Analysis of Fashion Market Trend Using Advertising Data of Shopping<br>Information Site . . . . .  | 389 |
| <i>Retsuya Saito, Kohei Otake, and Takashi Namatame</i>  |     |
| The Power of Social Media Marketing on Young Consumers’<br>Travel-Related Co-creation Behavior . . . . .   | 401 |
| <i>Farzana Sharmin and Mohammad Tipu Sultan</i>  |     |
| An Exploratory Investigation of Facebook Live Marketing<br>by Women Entrepreneurs in Bangladesh . . . . .  | 415 |
| <i>Mohammad Tipu Sultan and Farzana Sharmin</i>  |     |
| A Study on Bilingual Superimposed Display Method on Digital Signage. . . . .   | 431 |
| <i>Takumi Uotani, Yoshiki Sakamoto, Yuki Takashima, Takashi Kurushima,<br/>Kimi Ueda, Hirotake Ishii, Hiroshi Shimoda, Rika Mochizuki,<br/>and Masahiro Watanabe</i> |     |

## Social Computing for Well-Being, Learning, and Entertainment

|   |     |
|---|-----|
| Zika Outbreak of 2016: Insights from Twitter. . . . .   | 447 |
| <i>Wasim Ahmed, Peter A. Bath, Laura Sbaffi, and Gianluca Demartini</i>   |     |
| An Analysis of the Current Policies for Social Media Use<br>in Saudi Higher Education . . . . .   | 459 |
| <i>Faowzia Alharthy, Yuanqiong Wang, and Alfreda Dudley</i>   |     |
| AMISA: A Pilot Study of an Emotional Supporting Device Between<br>Friends Over Long-Distance. . . . .   | 471 |
| <i>Yuanyuan Bian and Teng-Wen Chang</i>   |     |
| An Agile Product Design in a Smart City Context: A Use Case<br>for Air Pollution Awareness . . . . .  | 483 |
| <i>Jaime Díaz and Oscar Ancán</i>   |     |
| Instagram Stories . . . . .   | 501 |
| <i>Cristóbal Fernández-Robin, Scott McCoy, Diego Yáñez,<br/>and Luis Cardenas</i>   |     |
| Proposal to Enhance University Students' Motivation to Switch<br>to a Morning-Oriented Lifestyle with a Community Approach . . . . .                    | 511 |
| <i>Hidenori Fujino, Taiga Okunari, Yuko Kato, Honoka Kobashi,<br/>Tomoya Tarutani, Nao Miyano, and Soyoka Yagi</i>                                      |     |
| An Exploration of a Social Media Community: The Case of<br>#AcademicTwitter. . . . .  | 526 |
| <i>Lina Gomez-Vasquez and Enilda Romero-Hall</i>  |     |
| Does Delivery Method Matter for Multicultural Undergraduate Students?<br>A Case Study of an Australian University in the United Arab Emirates . . . . . | 538 |
| <i>Ajrina Hysaj and Doaa Hamam</i>  |     |
| Proposal of the Elderly Supporting System Based on the Perspective<br>of Local Community in Japan. . . . .  | 549 |
| <i>Ayaka Ito, Masaya Ando, Hitoshi Uchida, Muneo Takemoto,<br/>and Yuichi Murai</i>   |     |
| Proposal of the Onion Watch Application for Enjoying a Stroll . . . . .   | 559 |
| <i>Takayoshi Kitamura, Yu Gang, Tomoko Izumi, and Yoshio Nakatani</i>   |     |
| Online Gambling Activity in Finland 2006–2016 . . . . .   | 569 |
| <i>Aki Koivula, Pekka Räsänen, Ilkka Koiranen, and Teo Keipi</i>  |     |
| Being Together Apart: Does Communication via Social Media Help<br>or Harm Romantic Relationships? . . . . .   | 584 |
| <i>Mark Turner and Emma Prince</i>  |     |

|  |                   |
|--|-------------------|
| <p><b>Technology-Based Social Skills Learning for People with Autism<br/>Spectrum Disorder</b> . . . . .</p> <p style="padding-left: 20px;"><i>Katherine Valencia, Virginia Zaraza Rusu, Erick Jamet,<br/>Constanza Zúñiga, Eduardo Garrido, Cristian Rusu,<br/>and Daniela Quiñones</i></p> | <p><b>598</b></p> |
| <p><b>A Personalized and Context Aware Music Recommendation System</b> . . . . .</p> <p style="padding-left: 20px;"><i>Champika H. P. D. Wishwanath, Supuni N. Weerasinghe,<br/>Kanishka H. Illandara, A. S. T. M. R. D. S. Kadigamuwa,<br/>and Supunmali Ahangama</i></p>                   | <p><b>616</b></p> |
| <p><b>Author Index</b> . . . . .</p>   | <p><b>629</b></p> |

# **Design Issues in Social Computing**



# Image Strength and Identity Diffusion as Factors Influencing the Perception of Hospitals by Their Facebook Communities

Michael Beier<sup>(✉)</sup> and Sebastian Früh

University of Applied Sciences of the Grisons, 7000 Chur, Switzerland  
{michael.beier, sebastian.frueh}@fhgr.ch

**Abstract.** Facebook provides hospitals many potential benefits but also forces them to adapt the way they connect with their stakeholders in various fields of application. Within these fields of application hospitals can act in various roles. Hospitals might influence these roles, but they are also dependent on how they are perceived by their Facebook community. In this paper, we aim to find out how hospitals differ in the perception by their Facebook communities and how these differences can be measured. Furthermore, we develop hypotheses with image strength and identity diffusion as factors influencing how hospitals are perceived by their Facebook communities. We test our hypotheses with data of all hospitals in Switzerland. Our statistical analysis provides strong support for all of our hypotheses. Our findings might help hospitals better to assess their position in their Facebook community and to adapt their intended roles as well as their strategies, content, and behavior in accordance to that.

**Keywords:** Social media · Facebook · Hospitals · Organizational identity · Brand image · Content · Switzerland

## 1 Introduction

Social media (and Facebook in particular) provide hospitals many potential benefits but also force them to adapt the way they connect with their stakeholders in various fields of application, like marketing, public health information, enhanced service delivery, recruiting, or employer branding (Beier and Früh 2020; Carpentier et al. 2017; Diddi and Lundy 2017; Kordzadeh and Young 2018; O'Connor et al. 2016; Smith 2017; Wong et al. 2016). Within these fields of social media applications hospitals can act in various roles (e.g., information source, content curator, regional brand, or host of an online community). Hospitals might influence these roles by their strategy, content, or behavior. But regarding the ability to perform their intended role on social media to a certain extent they are just dependent on the perception of their social media community.

In this paper, we aim to answer two research questions. First, we want to find out how hospitals differ in the perception by their Facebook communities and how these differences can be measured (*RQ1*). Second, we want to answer the question what factors



influence this perception of hospitals by their Facebook communities (RQ2). To answer RQ1 we develop theory on users' motivations to like or follow Facebook fan pages. To answer RQ2 we develop a research model with two groups of hypotheses covering image strength and identity diffusion as influencing factors and test it with data of all hospitals in Switzerland.

## 2 Hospitals and Facebook

### 2.1 Hospitals and Social Media

Social media platforms provide important new channels for hospitals to reach the general public as well as to connect with specific stakeholder groups (Beier and Früh 2019a; Campbell et al. 2014; Moorhead et al. 2013). Social media allow hospitals to address their community with informational content as well as to offer social, emotional, and companionship support (Zhou et al. 2018). Therefore, social media in general are important for marketing and communication for hospitals (Smith 2017).

Many hospitals see in such digital channels significant potential to gain more patients for their services (Duymus et al. 2017; Huang and Chang 2012). On the one hand, this approach follows the logic of brand marketing by developing an individual brand image for a hospital, fostering trust and reputation, and to communicate it to its stakeholders (Wu 2011). On the other hand, this approach is often combined with the idea of digital word-of-mouth via social media (Bruhn et al. 2012). To this end, hospitals try to convert their stakeholders to "brand evangelists", who in turn further communicate the hospitals brand image to their other contacts (Medina et al. 2016). These stakeholder activities of further communicating and sharing content are also one of the basic ideas of virality in social media (De Bruyn and Lilien 2008; Heimbach et al. 2015).

Beside such marketing-oriented branding purposes, social media also allow hospitals to distribute relevant and reliable health information to their community, which is nowadays a very important service for society (Beier and Früh 2019a; Kordzadeh and Young 2018). On the one hand, more and more people rely on social media content for health information and their own health behavior (Zhou et al. 2018). On the other hand, many health related sources on social media are of questionable quality often spreading misleading information (Jindal and Liao 2018; Moorhead et al. 2013; Ventola 2014). Therefore, hospitals could act as a reliable source for health information in their communities and in many cases they have a societal mandate to do so (Beier and Früh 2019a; Hagg et al. 2018).

In summary, with today's social media platforms hospitals have the chance to promote their own brand marketing as well as to provide valuable and reliable information to their communities, if they manage to position their channels accordingly (Zhou et al. 2018). However, similar to organizations in other fields many hospitals still face considerable difficulties to apply social media platforms in a purposeful and strategic manner (Beier and Wagner 2016a; Glazer 2012; Medina et al. 2016; Vanzetta et al. 2014). Social media channels of many hospitals show low user engagement and the interaction with stakeholders is mainly unidirectional from the hospitals (Beier and Früh 2019b; Huang and Dunbar 2013; Vanzetta et al. 2014). Reasons of hospitals' challenges in that regard

are legal insecurities, ambiguous expectations of various stakeholder groups, a poor tradition in public communications but also difficulties of communicating scientific and health related content (Beier and Früh 2019a; Medina et al. 2016). Better knowledge of users' expectations in social media platforms and how the own hospital is perceived by its communities might be helpful to improve strategies, content, and behavior of hospitals in their own accounts on social media platforms.

## 2.2 Hospitals on Facebook

Facebook is currently the most popular social media platform in the world. In the third quarter 2018 Facebook had 2'271 million monthly active users (1'495 million daily active users) worldwide (Facebook 2019). Correspondingly, in many countries Facebook is the social media platform with the highest reach in the general public. In the USA and Canada Facebook has 242 million monthly users (185 million daily users), which is equivalent to shares of about 2/3 (monthly users) or 1/2 (daily users) of the population in the regions (Facebook 2019). With regard to hospitals, a study on social media observed already in 2014 a 99% adoption rate for Facebook in the USA (Griffis et al. 2014). In contrast, in Switzerland Facebook adoption has been slower and lower. In 2017, 45% of the population were Facebook users, 28% on a daily basis (IGEM 2017). A recent study on social media adoption of Swiss hospitals observed 58% of the hospitals having an own official Facebook fan page (Beier and Früh 2020). However, Facebook is the most used social media platform by hospitals in Switzerland.

Facebook is also the social media platform, which has been most often analyzed in empirical research on healthcare and social media (Hagg et al. 2018; Moorhead et al. 2013). Early research on hospitals and Facebook provided mainly descriptive results on adoption rates and usage intensities in various countries (e.g., Beier and Früh 2019b; Griffis et al. 2014; Richter et al. 2014; Thacker et al. 2011; Van de Belt et al. 2012; Vanzetta et al. 2014; Wong et al. 2016; Yang et al. 2018). Some of these studies also analyzed the engagement rates of Facebook users by simply counting page likes of hospital fan pages (e.g., Beier and Früh 2019b; Griffis et al. 2014; Richter et al. 2014; Yang et al. 2018). Other research analyzed content of Facebook posts and patterns of interaction to better understand how hospitals and Facebook users interact on the platform (e.g., Distaso et al. 2015; Kordzadeh and Young 2018; Yang et al. 2018).

Facebook is mainly applied by hospitals for brand marketing and content distribution purposes. For instance, content analyses of Facebook pages of US hospitals showed that the greatest proportion of posts (about 36%) consists of sharing of health information (Kordzadeh and Young 2018). Another study on US children hospitals' Facebook pages found balanced shares of promotional and informational posts (each about 35%) by the hospitals (Wong et al. 2016). Furthermore, researchers observed positive effects of Facebook applications for operational figures of hospitals. For instance, in a quantitative survey in Turkey six percent of the respondents said that they were influenced by Facebook in their hospital choice (Duyum et al. 2017). Another recent study observed positive effects of hospitals operating own Facebook pages on hospital patient revenues (only for hospitals located in rural areas but not for hospitals in urban regions) (Apenteng et al. 2020). However, research regarding organizational or corporate effects on hospitals resulting from their activities on Facebook is still in its infancy.

### 2.3 Facebook Page Engagement and Community Perception

As described above, strategies of organizations for Facebook are often driven by brand marketing purposes and word-of-mouth tactics. An important aspect of user engagement on Facebook (as well as word-of-mouth marketing) is the “Liking” of fan pages (Beukeboom et al. 2015; Halaszovich and Nel 2017; Mochon et al. 2017; Zaglia 2013). With liking a fan page, (depending on the privacy settings of the user and actual developments of the news feed algorithm) Facebook basically posts a respective “Like” story on the profile feed of the user and starts to display new posts of the fan page in the newsfeed of the user (Beukeboom et al. 2015; Halaszovich and Nel 2017). Furthermore, the fan page will be displayed in the info tab of the user profile under activities and interests as well as the user grants the owner of the fan page additional rights to communicate the connection on Facebook. Although Facebook likes are a rather weak form of customer or community engagement (Beukeboom et al. 2015; Zaglia 2013). However, they are a first step towards community engagement and a precondition for most of the interactive word-of-mouth functionalities on Facebook.

Research on the motivation for users to connect with fan pages on Facebook showed that there are two main groups of influencing factors (Brandtzaege et al. 2014; Halaszovich and Nel 2017). On the one hand, users are motivated by the content of a page (e.g., entertainment, information, curiosity). This content is provided by the posts on the fan page and can be followed by users in their own Facebook news feeds (Mochon et al. 2017). On the other hand, users are influenced by motives of brand attraction with a fan page mainly driven by a user’s wish to express and show a relationship to the organization or brand running the fan page (Nisar and Whitehead 2016).

Facebook provides its users functionalities to support them in both motivations to interact with fan pages. Facebook’s basic idea was that users can connect with brands and organizations like they can do it with people in the social network (Zaglia 2013). Therefore, they provided the “Like” button on fan pages. However, after this they recognized that the motivations to follow content posted on a page versus to publicly express a positive attitude and relation with a fan page are quite different. Therefore, Facebook additionally introduced the “Follow” button (initially called “Subscribe” button) in 2011 (Peters 2011; Hamburger 2012). Since then, users were able to differentiate if they are only interested in the content (“follow”) or to express their social connection with a fan page (“like”). However, technically, Facebook heavily promotes the application of the “like” button, which leads in a first step to both, a “like” and a “follow” of a fan page. In a second step, users however can define in the settings of their fan page connection, whether they really want both or if they want to switch of the “like” or the “follow” (Facebook Help). As described above, by adapting the settings of a fan page connection users can specify if they just want to get content of the page posted in their news feed (as “followers”), if they only want to express their connection with the organization or brand running the fan page (as “fans”), or if they want both. Table 1 shows a matrix how Facebook functionalities support users to express their preferences in dependence of their motivation to connect to a fan page.

These expressions of user preferences in Facebook pages also manifest an aggregate picture of how the community of a Facebook fan page perceives its operator. Users who perceive a hospital as an attractive brand, with which they want to express publicly their

**Table 1.** Motivation matrix for Facebook page connections

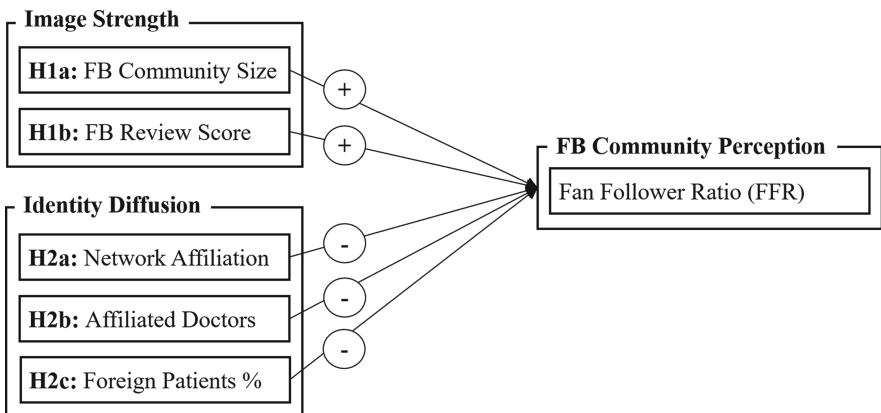
| Motivation for connection     | Content interest = <i>YES</i>  | Content interest = <i>NO</i>                                    |
|-------------------------------|--|---|
| Brand attraction = <i>YES</i> | “ <b>Fan</b> ” & “ <b>Follower</b> ”<br>Standard Setting after Page Like | “ <b>Fan</b> ”<br>Content Subscription Canceled after Page Like |
| Brand attraction = <i>NO</i>  | “ <b>Follower</b> ”<br>Visible Page Like switched off after initial Like | <b>No Connection</b>  |

connection on Facebook, connect as fans with the fan page. Users, who perceive the hospital as a source or provider of interesting content, connect as followers. User, who perceive a hospital as none of both, do not connect on Facebook. Most people might combine both motivations or do not adapt their connection settings adequately. However, the ratio of fans to followers on a Facebook fan page provides an aggregate indication of how an organization is perceived by its Facebook community. More concrete, the more the community perceives a hospital only as an attractive brand (including a willingness for open visible word-of-mouth activities) but not as a source of attractive content the higher will be the Fan Follower Ratio (hereinafter referred to as FFR).

In this section, (in regard to RQ1) we developed a concrete understanding of how Facebook communities might differ in their perception of hospitals and how these differences are observable directly on the Facebook fan pages with the FFR.

### 3 Hypotheses Development and Research Model

In a next step, (regarding RQ2) we develop a research model on two different groups of factors (image strength and identity diffusion) influencing the perception of hospitals by their Facebook communities. Figure 1 provides an overview of our research model:

**Fig. 1.** Research model.

### 3.1 Image Strength

In the first set of hypotheses, we follow the overall hypothesis that hospitals with a stronger brand image will show a higher FFR. One important factor influencing the word-of-mouth willingness of hospitals' Facebook communities is the image of a hospital. Image can generally be understood as the sum of attitudes, beliefs, and impressions that is distributed in a relevant population (Barich and Kotler 1991). Therefore, hospitals with a positive image and good reputation are trusted and recommended more by their stakeholders (Wang et al. 2011; Wu 2011). For some Facebook users liking a fan page of a hospital might be an ambiguous decision with regard to privacy risks as well as to potential reputation effects resulting of this connection. In such ambiguous situations people tend to show "herding behavior" by applying a simple social heuristic just following the observable behavior of others (Beier and Wagner 2016b; Duan et al. 2009; Lee and Lee 2012). Facebook supports such tendencies as it displays for fan pages the number of fans (size of a community) as well as an aggregate value for the five-star ratings of the community (reputation within the community). On the one hand, this provides an external indicator for brand image within a Facebook community. On the other hand, this also might increase herding behavior.

In a more general context of word-of-mouth marketing several studies have also observed that users tend more to word-of-mouth activities for an organization or a product, the more others already did it (Moe et al. 2012). Furthermore, people are more motivated to be publicly affiliated to organizations, which have higher reputation (Kovács and Horwitz 2018). This motivation might be enforced by expectations, that word-of-mouth activities for a reputable organization also might increase one's own reputation (Cheung and Lee 2012). Therefore, the visible reputation of a hospital on Facebook is also a relevant driver of word-of-mouth willingness within their community (Wang et al. 2011). In summary, we expect:

**H1a:** *Hospitals with larger Facebook communities (FB Community Size) will show a higher FFR on their Facebook fan page.*

**H1b:** *Hospitals, which are better rated by their Facebook community (FB Review Score), will show a higher FFR on their Facebook fan page.*

### 3.2 Identity Diffusion

In the second group of hypotheses, we follow the overall hypothesis that hospitals with a more diffuse social identity will show a lower FFR. Stakeholders' word-of-mouth willingness for a hospital mainly bases on the personal relatedness and identification with the organization (Medina et al. 2016). Therefore, a more diffuse identity of a hospital might lead to reduced brand attraction and lower word-of-mouth willingness for a hospital.

Generally, the organizational identity defines who an organization is. However, this identity is also related to the boundaries of an organization (Santos and Eisenhardt 2005). We expect two different ways, how hospitals might blur their boundaries and in doing so diffuse their organizational identity. On the one hand, there is a potential effect for hospitals to reduce the clarity of their organizational identity externally if they are

affiliated with a parent group or network (e.g., Hirslanden or Swiss Medical Network in Switzerland). Although, from an image perspective (especially smaller) hospitals might benefit from an affiliation with a bigger brand or parent organization (Vollmers et al. 2010). However, the organizational identity of a hospital might also become more diffuse as with the group affiliation it becomes complemented (or blurred) by its additional identity as part of another organization (Clark et al. 2010; Van Dick et al. 2006). This might decrease the identification of stakeholders with a hospital.

On the other hand, hospitals' organizational identities may also erode internally. This can be the case when internal stakeholders interact intensely in boundary spanning activities with external partners (Bartel 2001). In Switzerland many hospitals use affiliated doctors in addition to physicians directly employed by the hospital. These additional doctors allow to increase service capacities and might foster patient acquisition of hospitals (Kuntz and Scholtes 2013; Liedke et al. 2017). However, they are also external actors in the hospitals, which might diffuse their organizational identity.

Finally, it can be expected that the regional (or national) health service provision of a hospital might be a facilitator for the identification of local stakeholders with it. Furthermore, for many people a local hospital is much more than just a place of health service delivery, providing regional identification (Jones 2015). Accordingly, hospitals' organizational identity might also be connected with the local communities in their region. However, in some cases this connection might be damaged by activities of a hospital, for instance, if it strongly focusses on foreign patients and medical tourism (Blum 2015). Such an orientation on stakeholders from outside the regional community might reduce the perceived connection of the community members (Kaufmann 2017). We thus expect:

- H2a:** *Hospitals, which externally distort their organizational identity (with an affiliation to a group or network), will show a lower FFR on their Facebook fan page.*
- H2b:** *Hospitals, which internally distort their organizational identity (by operating an affiliated doctors system), will show a lower FFR on their Facebook fan page.*
- H2c:** *Hospitals, which weaken their regional/national identity (with a stronger focus on foreign patients), will show a lower FFR on their Facebook fan page.*

## 4 Data and Method

### 4.1 Variables and Measurement

Our research model considers two groups of independent variables with regard to image strength and identity diffusion of hospitals. We measured image strength with respect to the community size and the reputation of hospitals on Facebook. Therefore, we measured the size of hospitals' Facebook communities by the mean of the number of fans and the number of followers of their official Facebook page ("*FB Community Size*"). This aggregate value allowed us to measure the connections of Facebook users with a hospital without distorting the measure towards one specific connection type (fans vs. followers). Furthermore, we measured reputation of hospitals on Facebook with the overall review score displayed on the Facebook fan page of the hospital ("*FB Review Score*").

Regarding different types of identity diffusion of hospitals we defined three variables for our measurement. First, we measured external identity diffusion with a dummy variable (“*Group Affiliation*”), taking the value of one for hospitals where we found any indication on their homepage or in their social media accounts that they are affiliated to a group or network of other hospitals. Second, with regard to internal identity diffusion we computed a dummy variable (“*Affiliated Doctors*”) taking the value of one for hospitals, which have implemented an affiliated doctor system. Third, we measured hospitals’ weakening of their regional/national identity on the basis of their share of foreign patients (“*Foreign Patients %*”).

Additionally, we included several control variables into our analyses. We controlled for the size of hospitals measuring the number of inpatients per year (“*Hospital Size*”). Furthermore, we covered with dummy variables different types of hospitals. We applied hospitals that are established as private companies as reference class. In contrast to this reference class, we measured with two dummy variables public hospitals (“*Public Hospital*”) and hospitals established as foundations or associations (“*Found./Association*”).

We measured our dependent variable regarding hospitals’ perception by their Facebook community by means of a fan/follower ratio of a hospital’s Facebook page. This Fan Follower Ratio (FFR) represents the ratio between Facebook community members wanting to present their relationship with a hospital and their support on the platform against the members who prefer to only follow the fan page invisibly for others. More precisely, we calculated:

$$\text{Fan Follower Ratio (FFR)} = \frac{\text{Number of Fans}}{\text{Number of Followers}} - 1 \quad (1)$$

This specific formula leads to a FFR of 0 if the numbers of fans and followers are equal. The higher the value of the FFR, the more the fans outweigh over the followers in the Facebook community of a hospital. Correspondingly, a FFR of 0 indicates a balanced perception by the Facebook community of a hospital as a provider of valuable content as well as an attractive brand. The more the value of the FFR is greater than 0 the more the Facebook community tends to perceive the hospital more as an attractive brand than as a provider of valuable content. Values below 0 indicate the opposite pattern.

## 4.2 Data Collection and Sample

We tested our research model with data of all hospitals in Switzerland. For this purpose, we collected data from two different sources. First, we gathered relevant hospital data from the official list of the “Federal Office of Public Health” in Switzerland (FOPH) on key figures for all 283 Swiss hospitals (FOPH 2018). Second, we used the FOPH list to collect data on the Facebook pages of all hospitals in Switzerland. In line with previous studies, we collected the Facebook data in a three-step procedure (e.g., Richter et al. 2014; Thaker et al. 2011; Van de Belt et al. 2012; Yang et al. 2018). In the first step, we searched with Google Search for the official homepages of all hospitals. During this step we had to reduce our sample to 279. Two hospitals had closed. Another hospital is listed on the FOPH list as two sub-organizations. This did not allow to merge the FOPH

data with data of a Facebook page. We found for all remaining 279 hospitals an official internet homepage. In 96 cases the homepage provided a link to the official Facebook page of the hospital. In a second step, for hospitals, which did not provide a link to their Facebook page on their homepage, we also searched via Google Search for the name of each hospital in combination with the term “Facebook”. In a third step, we finally searched on Facebook itself for each name of the remaining hospitals.

This procedure resulted in the identification of 163 official Facebook pages (pages, which were not displayed on Facebook as “unofficial”) for hospitals on the official FOPH list. On these Facebook pages we gathered data on FB Review Score, number of page likes as well as page followers. The final collection of Facebook data was performed within three days in September 2018. Information about indications of group affiliations were obtained from the official homepages of the hospitals as well as from their official social media presences on Facebook, Twitter, Google+, Instagram, LinkedIn, XING, and YouTube. Not all official Facebook pages displayed numbers of followers and review scores. Excluding data sets with such missing values our final sample resulted in 130 complete data sets of Swiss hospitals with an official Facebook page.

### 4.3 Analytical Approach

In this study we aim to analyze the influence of several independent variables on one dependent variable (Fan Follower Ratio, FFR). Furthermore, the dependent variable is a metric variable. Therefore, we test our hypotheses by means of a linear regression analysis. We also considered potential multicollinearity issues. Therefore, we computed correlations (Kendall’s tau) between all explanatory variables in our analyses as well as variance inflation factors (VIF) in our regression analysis. All correlations were less than 0.35. Correspondingly, no multicollinearity issues were indicated for our regression analysis.

## 5 Results

We present the results of our data analyses in three steps. First, we provide descriptive statistics on all measures of our research model. Second, regarding RQ1 we present results on the perception of hospitals by their Facebook communities measured by the FFR. Third, regarding RQ2 we present the results of our hypotheses tests on basis of a linear regression.

### 5.1 Descriptive Statistics

In Table 2 we present descriptive statistics for all non-binary measures of our research model. Hospital size ranges between 37 and 41’010 patients per year with a mean of 6’326. Facebook communities of the hospitals in our sample are between 31 and 9’994 users with a mean of 1’047 and a median of 591. The hospitals have an average review score on Facebook of 4.5 with a median of 4.6. The share of foreign patients ranges between 0 and 67.2% with a mean of 3.8% and a median of 1.1%.

25.4% of the hospitals in our sample are affiliated to a network or parent group and 51.5% apply an affiliate doctors system. Furthermore, 18.5% of the hospitals are public and 21.5 are run by a foundation or association.



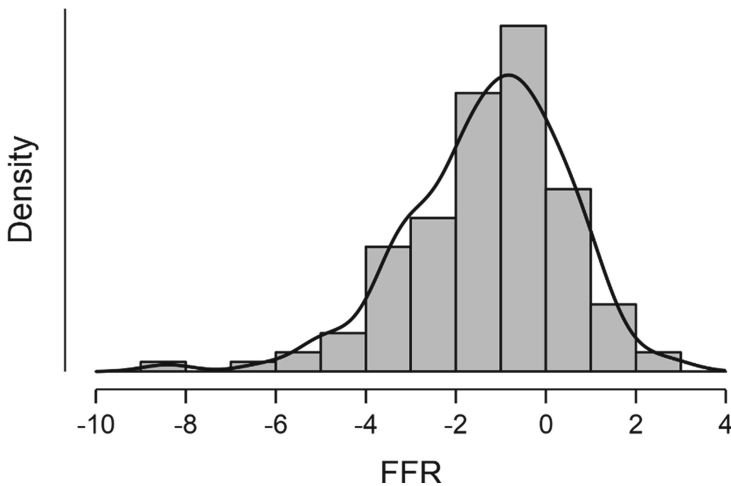
**Table 2.** Descriptive statistics (non-binary variables)

| Variable                     | Mean   | Median | SD     | Min    | Max    |
|------------------------------|--------|--------|--------|--------|--------|
| Fan Follower Ratio (DV) in % | -1.296 | -1.069 | 1.761  | -8.399 | 2.857  |
| Hospital Size (Cont.)        | 6'326  | 2'573  | 8'828  | 37     | 41'010 |
| FB Community Size (H1a)      | 1'047  | 591    | 1'424  | 31     | 9'994  |
| FB Review Score (H1b)        | 4.492  | 4.6    | 0.5338 | 1.0    | 5.0    |
| Foreign Patients % (H2c)     | 3.878  | 1.085  | 9.082  | 0.000  | 67.18  |

Notes: N = 130

### 5.2 Hospitals' Perception by Their Facebook Communities (RQ1)

The descriptive data in Table 2 show that the Fan Follower Ratio (FFR) of hospitals in our sample ranges from a minimum of -8.4 to a maximum of 2.9. Furthermore, the mean (-1.3) and the median (-1.1) are below 0. Correspondingly, hospitals in our sample have on average more followers than fans. Figure 2 displays the distribution and density of the FFR in our sample.



**Fig. 2.** Distribution of the fan follower ratio (FFR) in %; N = 130.

Overall, the hospitals in our sample are by trend perceived more as valuable content providers than as attractive brands by their Facebook communities. However, the propensity to show a connection on Facebook to a hospital might be generally lower than for other brand pages. Healthcare and hospitals are for many people a rather intimate topic, which they tend less to communicate openly online. For instance, in a study on the willingness to like a Facebook fan page on influenza vaccination between five and six percent of the respondents did not want to like the page because others could see it

on Facebook (Mena et al. 2012). So there might be a general tendency by some people not to show a connection with a hospital on Facebook.

Furthermore, the values for FFR have a range of 11.3 in our sample. This seems to be quite a low extent of variation in our data. However, the data only show the cases where Facebook users who engaged with a hospital fan page actively turned of the “fan” or “follower” attribute of their connection. The attitudes or intentions toward such a behavior might be much higher. For instance, under the term “privacy paradox” studies on user behavior on Facebook show that even so a magnitude of users perceive threats to their privacy on the platform, much less apply adequate privacy settings in the options (Jones and Soltren 2005; Kanampiu and Anwar 2018). Often users are not familiar with optional settings on Facebook features. For our study this means, that even the low variation in the FFR values might indicate much higher variations in the underlying perceptions of hospitals. However, at this point this is just an assumption and needs further empirical research to be clarified.

### 5.3 Regression Results (RQ2)

In Table 3 we present the results of our linear regression. Effects are displayed as standardized coefficients. As all variance inflation factors (VIF) are below 2 no multicollinearity issues are indicated (Dormann et al. 2013).

**Table 3.** Linear regression (dependent variable = fan follower ratio, FFR)

| Explanatory variables     | Standardized coefficients |     | VIF   |
|---------------------------|---------------------------|-----|-------|
| Hospital Size (Con.)      | -0.272                    | *** | 1.709 |
| Public Hospital (Con.)    | -0.105                    |     | 1.303 |
| Found./Association (Con.) | -0.208                    | **  | 1.173 |
| FB Community Size (H1a)   | 0.424                     | *** | 1.744 |
| FB Review Score (H1b)     | 0.183                     | **  | 1.124 |
| Group Affiliation (H2a)   | -0.224                    | **  | 1.231 |
| Affiliated Doctors (H2b)  | -0.170                    | **  | 1.190 |
| Foreign Patients % (H2c)  | -0.321                    | *** | 1.369 |
| R <sup>2</sup>            | 0.271                     |     |       |
| Adjusted R <sup>2</sup>   | 0.223                     |     |       |
| Significance              | 0.000                     |     |       |

Notes: N = 130;

\*0.1 > p ≥ 0.05; \*\*0.05 > p ≥ 0.01; \*\*\*p < 0.01.

Our control variables show that smaller hospitals have a lower FFR. Similarly, hospitals operated by a foundation or association show lower FFRs compared to private hospitals. Both seems plausible and can be interpreted as that smaller hospitals as well

as hospitals operated by a foundation or association are less perceived as attractive brands by their Facebook communities. In contrast, large and private hospitals tend more to be perceived as attractive brands.

Regarding effects of image strength of hospitals the regression results fully support our hypotheses. Hospitals with larger Facebook communities (H1a) as well as with higher review scores (H1b) on Facebook show higher FFRs. Correspondingly, these hospitals tend to be perceived by their Facebook communities more as attractive brands than as content providers.

Also all effects regarding identity diffusion of hospitals are fully supported by the linear regression analysis. More concrete, hospitals affiliated to a group or network (H2a), operating a system of affiliated doctors (H2b), or with a higher share of foreign patients (H2c) show significantly lower FFRs. Correspondingly, these hospitals tend to be perceived by their Facebook communities more as providers of interesting content than as attractive brands.

## 6 Discussion and Conclusions

Our statistical analysis provides strong support for all of our hypotheses. Our findings show how the community perception of hospitals (as valuable content providers or as attractive (regional) brands) is influenced by the strength of their image and the diffusion of their organizational identity. Our findings help hospitals to assess their position in their Facebook community better and to adapt their intended roles as well as their strategies, content, or behavior in accordance to that. This might support hospitals in a more dialogic evolution of their organizational identity (Theunissen 2014). On the one hand, own Facebook fan pages provide hospitals a valuable channel to express their own organizational identity to their internal and external stakeholders. On the other hand, own Facebook fan pages also support hospitals in learning more about themselves and how they are perceived by their stakeholders in their Facebook community (Devereux et al. Devereux et al. 2017).

The Fan Follower Ratio (FFR) presented in this study, provides an easy applicable metric for such purposes. Research and practice so far mostly applied only fan page likes as an overall metric for community engagement (e.g., Azar et al. 2016; Phelan et al. 2013). Only few studies included follower numbers on Facebook fan pages or analyzed fan and follower numbers together (e.g., Cao and Smith 2018). On the one hand, page likes are the stronger metric for a strategic focus on brand marketing and intended word-of-mouth effects in Facebook activities of organizations. On the other hand, both measures (fan and follower numbers) show highly correlated values, which is why most studies might have neglected follower metrics in statistical analyses. However, in the context of community perception the FFR (as a combination of fan and follower numbers) might be more beneficial than the usual applications of only fan (or follower) numbers.

This study presents first considerations and empirical results on Facebook community perception on basis of the FFR. The study and its results are rather of exploratory character as the study faces some significant limitations. One main limitation is the one-sided focus of the research model on brand image. Basically, the FFR is conceptualized to measure the balance between users, who are attracted to the fan page by the brand

image, and users, who are just interested in the content of the page. In this regard our research model is quite unbalanced, as all hypotheses only address patterns showing effects via brand attractiveness. In all hypotheses of the current model, the perception of fan page content is only covered as “*ceteris paribus*”. Future research should enhance the model with hypotheses on content of hospitals’ fan pages and its perception by their communities.

Another main limitation is that the FFR is a rather abstract measure with several unobserved effects, which can only be assumed at the current state of the model. For instance, the empirical analysis only applies aggregate numbers of fans and followers, but the overlap of users being both is unknown. Furthermore, the resulting numbers of adapted settings of users’ fan page connections might be biased, for instance, if the ability or motivation to adapt the connection settings differ between only fans and only followers of a fan page. More research is needed on how in detail users connect and relate to hospitals via Facebook. Applications of the Theory of Planned Behavior (Ajzen 1991) could explain in more detail how users develop intentions to adapt the settings for their connection with a Facebook fan page of a hospital and what behavior they finally show (Saeri et al. 2014).

Overall, this study provides some initial ideas and first empirical results on how hospitals are perceived by their Facebook communities and how this perception is influenced by the image strength and identity diffusion of the hospitals. However, more detailed (especially qualitative) research is needed to further develop knowledge and understanding of the topic.

## References

- Ajzen, I.: The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* **50**(2), 179–211 (1991)
- Apenteng B., Ekpo, I.B., Mutiso, F.M., Akowuah, E.A., Opoku, S.T.: Examining the relationship between social media engagement and hospital revenue. *Health Market. Q.* **37**(1), 10–21 (2020). <https://doi.org/10.1080/07359683.2020.1713575>
- Azar, S.L., Machado, J.C., Vacas-de-Carvalho, L., Mendes, A.: Motivations to interact with brands on Facebook. Towards a typology of consumer-brand interactions. *J. Brand Manag.* **23**(2), 153–178 (2016)
- Barich, H., Kotler, P.: A framework for marketing image management. *MIT Sloan Manag. Rev.* **32**(2), 94–104 (1991)
- Bartel, C.A.: Social comparisons in boundary-spanning work: effects of community outreach on members’ organizational identity and identification. *Adm. Sci. Q.* **46**(3), 379–413 (2001)
- Beier, M., Früh, S.: Technological, organizational, and environmental factors influencing social media adoption by hospitals in Switzerland: cross-sectional study. *J. Med. Internet Res.* **22**(3), e16995 (2020). <https://doi.org/10.2196/16995>
- Beier, M., Früh, S.: Social media teams of hospitals as mediators in digital health ecosystems. In: Song, Y., Grippa, F., Gloor, P., Leitão, J. (eds.) *Collaborative Innovation Networks*, pp. 115–124. Springer, Cham (2019a). [https://doi.org/10.1007/978-3-030-17238-1\\_6](https://doi.org/10.1007/978-3-030-17238-1_6)
- Beier, M., Früh, S.: Swiss hospitals on social media platforms: own accounts, communication frequencies and reach. *Soc. Sci. Res. Network Electron. J.*, 3367780 (2019b)
- Beier, M., Wagner, K.: Social media adoption: barriers to the strategic use of social media in SMEs. In: *Proceedings of the European Conference on Information Systems (ECIS)*, AIS, Istanbul, Turkey (2016a)



- Beier, M., Wagner, K.: User behavior in crowdfunding platforms - exploratory evidence from Switzerland. In: Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS), pp. 3584–3593. IEEE, Kauai (2016b)
- Beukeboom, C.J., Kerkhof, P., de Vries, M.: Does a virtual like cause actual liking? How following a brand's Facebook updates enhances brand evaluations and purchase intention. *J. Interact. Market.* **32**, 26–36 (2015)
- Blum, K.: Krankenhauspatienten aus dem Ausland. *Das Krankenhaus* (05) (2015)
- Brandtzaeg, P.B., Haugstveit, I.M.: Facebook likes: a study of liking practices for humanitarian causes. *Int. J. Web Based Communities* **10**(3), 258–279 (2014)
- Bruhn, M., Schoenmueller, V., Schäfer, D.B.: Are social media replacing traditional media in terms of brand equity creation? *Manag. Res. Rev.* **35**(9), 770–790 (2012)
- Campbell, D.A., Lambright, K.T., Wells, C.J.: Looking for friends, fans, and followers? Social media use in public and nonprofit human services. *Public Adm. Rev.* **74**(5), 655–663 (2014)
- Cao, S., Smith, G.P.: Bringing big data from social media reviews to quality improvement. *J. Am. Acad. Dermatol.* **79**(5), 951–952 (2018)
- Carpentier, M., Van Hoye, G., Stockman, S., Schollaert, E., Van Theemsche, B., Jacobs, G.: Recruiting nurses through social media: effects on employer brand and attractiveness. *J. Adv. Nurs.* **73**(11), 2696–2708 (2017)
- Cheung, C.M., Lee, M.K.: What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. *Decis. Support Syst.* **53**(1), 218–225 (2012)
- Clark, S.M., Gioia, D.A., Ketchen Jr., D.J., Thomas, J.B.: Transitional identity as a facilitator of organizational identity change during a merger. *Adm. Sci. Q.* **55**(3), 397–438 (2010)
- De Bruyn, A., Lilien, G.L.: A multi-stage model of word-of-mouth influence through viral marketing. *Int. J. Res. Mark.* **25**(3), 151–163 (2008)
- Devereux, L., Melewar, T.C., Foroudi, P.: Corporate identity and social media: existence and extension of the organization. *Int. Stud. Manag. Organ.* **47**(2), 110–134 (2017)
- Diddi, P., Lundy, L.K.: Organizational Twitter use: content analysis of Tweets during breast cancer awareness month. *J. Health Commun.* **22**(3), 243–253 (2017)
- DiStaso, M.W., Vafeiadis, M., Amaral, C.: Managing a health crisis on Facebook: how the response strategies of apology, sympathy, and information influence public relations. *Public Relat. Rev.* **41**(2), 222–231 (2015)
- Dormann, C.F., et al.: Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**(1), 27–46 (2013)
- Duan, W., Gu, B., Whinston, A.B.: Informational cascades and software adoption on the internet: an empirical investigation. *MIS Q.* **33**(1), 23–48 (2009)
- Duymus, T.M., et al.: Internet and social media usage of orthopaedic patients: a questionnaire-based survey. *World J. Orthop.* **8**(2), 178–186 (2017)
- Facebook (2019). [https://s21.q4cdn.com/399680738/files/doc\\_financials/2018/Q3/Q3-2018-Earnings-Presentation.pdf](https://s21.q4cdn.com/399680738/files/doc_financials/2018/Q3/Q3-2018-Earnings-Presentation.pdf). Accessed 31 Jan 2020
- Facebook Help, How do I unfollow a person, Page or group? (2019). <https://www.facebook.com/help/190078864497547>. Accessed 31 Jan 2020
- FOPH (Federal Office of Public Health), Kennzahlen der Schweizer Spitäler (2018). <https://www.bag.admin.ch/bag/de/home/zahlen-und-statistiken/zahlen-fakten-zu-spitaelern/kennzahlen-der-schweizer-spitaeler.html>. Accessed 31 Jan 2020
- Glazer, H.: “Likes” are lovely, but do they lead to more logins? Developing metrics for academic libraries’ Facebook pages. *College Res. Libr. News* **73**(1), 18–21 (2012)
- Griffis, H.M., et al.: Use of social media across US hospitals: descriptive analysis of adoption and utilization. *J. Med. Internet Res.* **16**(11) (2014)
- Hagg, E., Dahinten, V.S., Currie, L.M.: The emerging use of social media for health-related purposes in low and middle-income countries: a scoping review. *Int. J. Med. Informatics* **115**, 92–105 (2018)

- Halaszovich, T., Nel, J.: Customer–brand engagement and Facebook fan-page “Like”-intention. *J. Prod. Brand Manag.* **26**(2), 120–134 (2017)
- Hamburger, E.: Facebook snubs ‘subscribe’ button in favor of Twitter-esque ‘follow’ on all profile pages (2012). <https://www.theverge.com/2012/12/5/3731986/facebook-like-follow>. Accessed 31 Jan 2020
- Heimbach, I., Schiller, B., Strufe, T., Hinz, O.: Content virality on online social networks: empirical evidence from Twitter, Facebook, and Google+ on German news websites. In: *Proceedings of the 26th ACM Conference on Hypertext and Social Media*, pp. 39–47. ACM (2015)
- Huang, E., Chang, C.C.A.: Patient-oriented interactive e-health tools on US hospital Web sites. *Health Mark. Q.* **29**(4), 329–345 (2012)
- Huang, E., Dunbar, C.L.: Connecting to patients via social media: a hype or a reality? *J. Med. Market.* **13**(1), 14–23 (2013)
- IGEM (Interessengemeinschaft elektronische Medien), IGEM-digiMONITOR (2017). [https://www.igem.ch/download/IGEM-digiMONITOR\\_Factsheet-2017.pdf](https://www.igem.ch/download/IGEM-digiMONITOR_Factsheet-2017.pdf). Accessed 31 Jan 2020
- Jindal, G., Liao, Y.: Living with HIV/AIDS: exploring vloggers’ narratives on YouTube. In: *Proceedings of the 9th International Conference on Social Media and Society*, pp. 320–324. ACM (2018)
- Jones, L.: What does a hospital mean? *J. Health Serv. Res. Policy* **20**(4), 254–256 (2015)
- Jones, H., Soltren, J.H.: Facebook: threats to privacy. Project MAC. MIT Project Math. *Comput.* **1**(01) (2005)
- Kanampiu, M., Anwar, M.: Privacy preferences vs. privacy settings: an exploratory Facebook study. In: Ahram, T.Z., Nicholson, D. (eds.) *AHFE 2018. AISC*, vol. 782, pp. 116–126. Springer, Cham (2019). [https://doi.org/10.1007/978-3-319-94782-2\\_12](https://doi.org/10.1007/978-3-319-94782-2_12)
- Kaufmann, M.: Medizintourismus: Das Geschäft mit den «Selbstzahlern». *Die Schweiz wird zum Spital für reiche Ausländer*. *Blick*, 15 October 2017
- Kordzadeh, N., Young, D.K.: Exploring hospitals’ use of Facebook: thematic analysis. *J. Med. Internet Res.* **20**(5) (2018)
- Kovács, B., Horwitz, S.: Conspicuous reviewing: affiliation with high-status organizations as a motivation for writing online reviews. *Socius* **4**, 1–14 (2018)
- Kuntz, L., Scholtes, S.: Physicians in leadership: the association between medical director involvement and staff-to-patient ratios. *Health Care Manag. Sci.* **16**(2), 129–138 (2013)
- Lee, E., Lee, B.: Herding behavior in online P2P lending: an empirical investigation. *Electron. Commer. Res. Appl.* **11**(5), 495–503 (2012)
- Liedtke, D., Amgwerd, N., Wiesinger, O., Mauer, D., Westerhoff, C., Pahls, S.: The integrated-physician-model: business model innovation in hospital management. In: Pfannstiel, M.A., Rasche, C. (eds.) *Service Business Model Innovation in Healthcare and Hospital Management*, pp. 31–55. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-46412-1\\_3](https://doi.org/10.1007/978-3-319-46412-1_3)
- Medina, P., Buil, P., Heath, R.L.: Establishing and demonstrating US hospital brands through Facebook. *Observatorio* **10**(3), 20–40 (2016)
- Mena, G., Llupià, A., García-Basteiro, A.L., Aldea, M., Sequera, V.G., Trilla, A.: The willingness of medical students to use Facebook as a training channel for professional habits: the case of influenza vaccination. *Cyberpsychol. Behav. Soc. Network.* **15**(6), 328–331 (2012)
- Mochon, D., Johnson, K., Schwartz, J., Ariely, D.: What are likes worth? A Facebook page field experiment. *J. Mark. Res.* **54**(2), 306–317 (2017)
- Moe, W.W., Schweidel, D.A.: Online product opinions: incidence, evaluation, and evolution. *Market. Sci.* **31**(3), 372–386 (2012)
- Moorhead, S.A., Hazlett, D.E., Harrison, L., Carroll, J.K., Irwin, A., Hoving, C.: A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *J. Med. Internet Res.* **15**(4) (2013)
- Nisar, T.M., Whitehead, C.: Brand interactions and social media: enhancing user loyalty through social networking sites. *Comput. Hum. Behav.* **62**, 743–753 (2016)

- O'Connor, M.I., Brennan, K., Kazmerchak, S., Pratt, J.: YouTube videos to create a "virtual hospital experience" for hip and knee replacement patients to decrease preoperative anxiety: a randomized trial. *Interact. J. Med. Res.* **5**(2) (2016)
- Peters, M.: Facebook subscribe button: what it means for each type of user (2011). <https://mashable.com/2011/09/15/facebook-subscribe-users>. Accessed 31 Jan 2020. Visited on 31 Jan 2020
- Phelan, K.V., Chen, H.T., Haney, M.: "Like" and "Check-in": how hotels utilize Facebook as an effective marketing tool. *J. Hospitality Tourism Technol.* **4**(2), 134–154 (2013)
- Richter, J.P., Muhlestein, D.B., Wilks, C.E.: Social media: how hospitals use it, and opportunities for future use. *J. Healthc. Manag.* **59**(6), 447–461 (2014)
- Saeri, A.K., Ogilvie, C., La Macchia, S.T., Smith, J.R., Louis, W.R.: Predicting Facebook users' online privacy protection: risk, trust, norm focus theory, and the theory of planned behavior. *J. Soc. Psychol.* **154**(4), 352–369 (2014)
- Santos, F.M., Eisenhardt, K.M.: Organizational boundaries and theories of organization. *Organ. Sci.* **16**(5), 491–508 (2005)
- Smith, K.T.: Hospital marketing and communications via social media. *Serv. Market. Q.* **38**(3), 187–201 (2017)
- Thaker, S.I., Nowacki, A.S., Mehta, N.B., Edwards, A.R.: How U.S. hospitals use social media. *Ann. Internal Med.* **154**(10), 707–708 (2011)
- Theunissen, P.: Co-creating corporate identity through dialogue: a pilot study. *Public Relat. Rev.* **40**(3), 612–614 (2014)
- Van de Belt, T.H., Berben, S.A., Samsom, M., Engelen, L.J., Schoonhoven, L.: Use of social media by western European hospitals: longitudinal study. *J. Med. Internet Res.* **14**(3) (2012)
- Van Dick, R., Ullrich, J., Tissington, P.A.: Working under a black cloud: how to sustain organizational identification after a merger. *Br. J. Manag.* **17**(S1), S69–S79 (2006)
- Vanzetta, M., Vellone, E., Dal Molin, A., Rocco, G., De Marinis, M.G., Rosaria, A.: Communication with the public in the healthcare system: a descriptive study of the use of social media in local health authorities and public hospitals in Italy. *Annali dell'Istituto Superiore di Sanità* **50**, 163–170 (2014)
- Ventola, C.L.: Social media and health care professionals: benefits, risks, and best practices. *Pharmacy Therapeutics* **39**(7), 491–499 + 520 (2014)
- Vollmers, S.M., Miller, D.W., Kilic, O.: An analysis of hospital brand mark clusters. *J. Hosp. Market. Public Relat.* **20**(2), 87–99 (2010)
- Wang, Y.C., Hsu, K.C., Hsu, S.H., Hsieh, P.A.J.: Constructing an index for brand equity: a hospital example. *Serv. Ind. J.* **31**(2), 311–322 (2011)
- Wong, C.A., Ostapovich, G., Kramer-Golinkoff, E., Griffis, H., Asch, D.A., Merchant, R.M.: How US children's hospitals use social media: a mixed methods study. *Healthcare* **4**(1), 15–21 (2016)
- Wu, C.C.: The impact of hospital brand image on service quality, patient satisfaction and loyalty. *Afr. J. Bus. Manage.* **5**(12), 4873–4882 (2011)
- Yang, P.C., et al.: Use of Facebook by hospitals in Taiwan: a nationwide survey. *Int. J. Environ. Res. Public Health* **15**(6), 1188 (2018)
- Zaglia, M.E.: Brand communities embedded in social networks. *J. Bus. Res.* **66**(2), 216–223 (2013)
- Zhou, L., Zhang, D., Yang, C.C., Wang, Y.: Harnessing social media for health information management. *Electron. Commer. Res. Appl.* **27**, 139–151 (2018)



# The Ideal Topic: Interdependence of Topic Interpretability and Other Quality Features in Topic Modelling for Short Texts

Ivan S. Blekanov<sup>(✉)</sup> , Svetlana S. Bodrunova , Nina Zhuravleva,  
Anna Smoliarova, and Nikita Tarasov

St. Petersburg State University, St. Petersburg 199004, Russia  
i.blekanov@spbu.ru

**Abstract.** *Background.* Topic modelling is a method of automated probabilistic detection of topics in a text collection. Use of topic modelling for short texts, e.g. tweets or search engine queries, is complicated due to their short length and grammatical flaws, including broken word order, abbreviations, and contamination of different languages. At the same time, as our research shows, human coding cannot be perceived as a baseline for topic quality assessment. *Objectives.* We use biterm topic model (BTM) to test the relations between two topic quality metrics independent from topic coherence with the human topic interpretability. Topic modelling is applied to three cases of conflictual Twitter discussions in three different languages, namely the *Charlie Hebdo* shooting (France), the Ferguson unrest (the USA), and the anti-immigrant bashing in Biryulevo (Russia), which represent, respectively, a global multilingual, a large monolingual, and a mid-range monolingual type of discussions. *Method.* First, we evaluate the human baseline coding by providing evidence for the Russian case on the coding by two pairs of coders who have varying levels of knowledge of the case. We then measure the quality of modelling on the level of topics by looking at topic interpretability (by experienced coders), topic robustness, and topic saliency. *Results.* The results of the experiment show that: 1) the idea of human coding as baseline needs to be rejected; 2) topic interpretability, robustness, and saliency can be inter-related; 3) the multilingual discussion performs better than the monolingual ones in terms of interdependence of the metrics. *Conclusion.* We formulate the idea of an 'ideal topic' that rethinks the goal of topic modelling towards finding a smaller number of good topics rather instead of maximization of the number of interpretable topics.

**Keywords:** Twitter · Inter-ethnic discussions · Topic modelling · Ideal topic · Topic quality · Human coding

## 1 Introduction

The studies of large text corpora via automated detection of topicality are a growing area of social research. Topic modelling is a probabilistic tool for discovering non-evident topics in a collection of texts. After the potential of topic modelling for Twitter



has been proven [1], it has become a separate and growing area within topic detection research. Several works have shown that topic modelling of Twitter data is suitable for both detection of hidden discussion topics and data pre-processing aimed at reducing the topical dimensionality of the studied corpora [2, 3], as well as for a range of comparative tasks. In both capacities, it has already been applied to practical tasks like drug use analysis [4] or assessment of popular behavior during natural disasters [5, 6].

But, despite the promise, topic modelling is highly problematic when applied to short texts [7], especially of the oral-written nature [8] often met online, of which Twitter is exemplary. Short texts, including Twitter posts (tweets), are today a popular form of public speech online, but their nature creates substantial obstacles to modelling topicality in Twitter corpora of virtually any size.

Topic modelling algorithms for the English language on Twitter have been explored widely, even if not extensively. Several models beyond LDA with Gibbs sampling, such as vector models or pLSA, as well as extensions for LDA, have been proposed. In most cases, classic unsupervised LDA performed worse than extended LDA or other models like biterm topic modelling (BTM). But, in some cases, LDA is used unquestioned, with results interpretable enough; this is why we will compare LDA which is in wide use with two other models described below.

Till today, comparative works on Twitter topic models for various languages beyond English remain very rare, mostly focusing on German and Spanish. Comparative studies of topic modelling for the francophone segment of Twitter are truly scarce, and, for the Russian-language Twitter, except for the earlier pilot works by our working group [1, 2, 9], are almost non-existent. For a rare comparative work on five languages including French and Russian. This work, similar to ours in its goals, tests three models of topic detection, including the authors' own development, and tests them by human coders and objective measures. The paper reports the comparative results on topic coherence and the so-called 'utility' of topic clusters (not topics!) without explaining how utility relates to human interpretability (whether this actually is interpretability or not). The authors also claim that their own model performs better than both LDA and BTM, but they do not reveal its full details for that would allow for comparison of the algorithms. Moreover, the results are not presented in full but, e.g. for all the languages beyond English, only on LDA and the authors' model. Due to these reasons, we will look to French and Russian to once again compare the quality of topic detection.

Another gap in the current research on Twitter topics detection is that, except for e-health and natural disasters, issue- or event-oriented discussions are practically not studied. The methodological papers, including [10], do a lot to avoid limitations to overall Twitter topicality and work with large data that contain all possible topics. But, within a case-oriented discussion, detection of topics must be seriously complicated by the fact that the inner sub-topics share very similar lexicons.

To partly cover these gaps, we have taken under scrutiny three conflictual cases in three different countries, as they have provoked the respective Twitter discussion of a large scale but at the same time were focused enough for our experiment. In our earlier research, we have applied unsupervised LDA, as well as WNTM and BTM, to these three cases (see below), with BTM showing the best results by automated quality assessment metrics, namely Umass and NPMI [2]. But here we need to re-discuss our conclusions

and juxtapose them to other, more subjective quality metrics such as topic saliency and human interpretation. Also, we try to see whether topic saliency and topic interpretability are linked to each other; to our best knowledge, this idea has not yet been explored by the academic community.

The remainder of the paper is organized as follows. In Sect. 2, we discuss the topic modelling algorithms that we use and the results we have received by applying the automated quality metrics and human coding, with the conclusion to reject both baselines. In Sect. 3, we present our ideas on the two other measures of topic detection quality, namely topic saliency and topic robustness. Section 4 serves for describing the cases and formulating the hypotheses. Section 5 provides the description of methods. In Sect. 6, we present the results and discuss them. To conclude, we formulate our idea of the ‘ideal topic’.

## 2 Human and Automated Evaluation of Quality: Rejection of Both Baselines

### 2.1 Automated Quality Assessment

For automated quality assessment, we have used the coherence metrics that have gained popularity in the recent academic research, namely Umass and normalized PMI (NPMI). As we have shown in our earlier pre-tests, by both metrics, we have seen BTM to be performing better than the two other algorithms [2].

But, after preliminary evaluation by human coders, we have seen that there are problems in topic interpretation that are not captured by the topic coherence metrics. Thus, the problems that we have discovered may be summarized as follows: 1) seemingly low number of interpretable topics; 2) a relatively big number of topics too similar to each other; 3) one-retweet-based topics gathered around one popular tweet rather than around a real topic in the discussion; 4) chained topics that unite several micro-themes but are non-clear to the human coders.

### 2.2 Human Coding and Its Artifacts

Interpretability is understood here as an ability of a human coder to say why the topic descriptors (the most relevant words that describe the topic) stay together in a given topic. This is a traditional and ultimate metric of topic quality, as human interpretability is the ultimate goal of the whole modelling procedure. But this metric is non-feasible if the number of topics in multiple runs of the algorithms is overwhelming.

Interpretability is really hard to measure in an objective way. Moreover, below we show that human coding is critically dependent on three types of knowledge: that on the case, that on method, and that on the dataset. The difference in preparation to coding critically alters the coding results.

In Table 1, we show the coding results for three pairs of coders: unexperienced coders who have no detailed knowledge on the case, unexperienced coders who have received detailed instructions on the case and a lot of background knowledge, and members of our working group who have knowledge on the case, the dataset, and the algorithms.

**Table 1.** Human coding of the Russian dataset, 100 topics: the artifacts of coder training

|      | Unexperienced |         | Trained |         | Experienced |         |
|------|---------------|---------|---------|---------|-------------|---------|
|      | Coder 1       | Coder 2 | Coder 1 | Coder 2 | Coder 1     | Coder 2 |
| LDA  | 33%           | 11%     | 65%     | 94%     | 53%         | 63%     |
| WNTM | 12%           | 11%     | 75%     | 92%     | 76%         | 73%     |
| BTM  | 22%           | 8%      | 88%     | 86%     | 76%         | 83%     |

As we see from Table 1, the coding result is, indeed, highly skewed by the coder background knowledge. Also, experienced coders from the working group show the tendency to give out a more balanced coding output, as they are more strict on the method performance and do not imply the inner grammar relations to the topics, filtering interpretability by the knowledge on the internal discourses in the dataset.

Thus, we see both automated dataset-level and human-based topic-level metrics to be unreliable in terms of finding ‘real topics’. This is why we suggest two other metrics and test their interconnectedness with human interpretability, to be able in future to substitute human coding by measuring these metrics.

### 3 Topic-Level Quality Assessment: In Search of Independence Form Human Eye

Thus, we have decided to introduce topic-level metrics of quality of topic modelling and test whether human interpretability will be related to them; we do it in an attempt to find topic-level metrics to substitute human coding without the necessity to assess the quality of the algorithm. At the same time, we still consider the well-known *tf-idf* metric relevant for topic quality detection and will test our metrics against it in future.

#### 3.1 Topic Saliency

The idea of topic saliency lies in the fact that, for various time slots within a discussion, different sub-themes may emerge as the leading ones. Saliency is calculated in the following way. Each document in the dataset is assigned a topic distribution; then, based on the time of publication, the topic distributions in the tweets are summarized, and, for each topic, the saliency is calculated as the sum of the tweets belonging (with a certain probability) to the particular topic. Since the intensity of the discussion may vary in time, the overall topic saliency for all topics may vary (from 0 to 1, in our case). We have used a 24-h step and measured the topic saliency for each day in aggregate. This strategy constitutes one of the limitations for our study, as the French dataset encompasses four days only, due to its overwhelming volume (see below); in future, for the 4-day French dataset, we plan to change the step and conduct an hourly analysis.

To our best knowledge, topic saliency has not yet been discussed in the literature as a possible quality metric for topic modelling. But, by commonsense logic, topic saliency should be related to topic interpretability: the more salient a topic is, the bigger it is

within the dataset and must be easier to interpret. And, if so, topic saliency may in future be seen a quality metric – a proxy for understanding topics by human coders.

### 3.2 Topic Robustness

Topic robustness shows whether the topic is composed of a relatively large number of relevant texts/words. We will look at the relevance levels of topic descriptors (top words). The relevance is, basically, the extent to which a particular word belongs to a given topic, depending on how many times it is met in the topic-relevant tweets; the relevance is measured 0 to 1. There is no clear agreement in today's research what relevance is to be considered high; this metric is dataset-dependent and thus needs to be assessed for each topic depending on the highest levels for a given run of topic modelling. Usual values for word relevance range from 0.01 to 0.1 for short-text small datasets to 0.001 to 0.05 for longer-text larger datasets.

Topic robustness, thus, combines two internal parameters: the word relevance to a topic and the number of relevant words in a topic. Robust topics are those who have many highly relevant words; it means they attract a lot of similar texts. We will provide the exact measurements for topic robustness in Sect. 5. But topics with lower robustness must not be viewed as 'bad': robustness highly depends on the nature of the discussion.

We will check the relationships between these metrics, to see how one can describe an ideal topic: a robust, understandable, salient one – and assess which datasets (bigger/smaller, monolingual/multilingual) provide for a bigger quantity of ideal topics.

## 4 The Cases Under Scrutiny and the Research Hypotheses

Here, we describe the cases we work with, which will help formulate the hypotheses. For this research, we work with the datasets from Twitter on the three cases:

- Anti-immigrant riots in the Moscow district of Biryulevo, Russia, 2013: number of users who published tweets – 3574, total number of user messages – 10215;
- The Ferguson unrest, USA, 2014: number of users who published tweets – 70018, total number of messages – 193812;
- The Charlie Hebdo shooting, France, 2015: number of users who published tweets – 238491, total number of messages – 505069.

The datasets are highly uneven in volume, and thus we had conducted pre-tests of the number of topics that would provide the most interpretable results and, at the same time, remain comparable. We ran the topic modelling for all the three cases with the BTM algorithm for the following number of topics set arbitrarily: for Biryulevo, 10, 50, and 100; for Ferguson, 50, 100, and 200; for Charlie Hebdo, 100, 200, and 400 topics. We have pre-tested their interpretability for small numbers of topics for each run (10 to 20), and we have seen that, for 100 topics, the results may be comparable. This is why we will further on use the 100-topic runs for comparing the quality assessment results.

Having in mind the aforementioned metrics for quality assessment and the number of topics in each run, we have set the following hypotheses:

**H1.** Human-based interpretability will correspond to the volume of the dataset: the bigger the dataset, the more interpretable the topics are.

**H2.** Human-based interpretability will be higher for monolingual discussions (Russia, the USA) than for multilingual discussions (France).

**H3.** More interpretable topics will have higher saliency in all the datasets.

**H4.** More interpretable topics will show higher robustness in all the datasets.

**H5.** More robust topics will have higher saliency in all the datasets.

We note that H1 and H2 are mutually exclusive.

## 5 The Research Methods

Here, we will in short describe the methods we use to test the hypotheses. As our data collection has been well-described in our previous works, we will only provide the description for the methods directly used in this paper.

For H1 and H2, we have worked with native speakers as human coders. Unlike in other research, we do not test the inter-rater reliability for the coders and then let them code separate segments of the task. Instead, two coders were asked to code the topic descriptors independently, and then, their inter-coder reliability was checked. We used this to see how divergent the opinions of the coders could be and whether the number of interpretable topics corresponds to earlier studies of longer texts, e.g. Russian blogs [3]. Thus, six coders were instructed to code the topics as interpretable or non-interpretable based on the abovementioned assumption of comprehensibility of why the topic descriptors stay together in a particular topic. Then, the number of interpretable topics for each coder and the meta-coding results were calculated. Each topic was assigned an interpretability index of 0 (non-interpretable for both coders), 1 (interpretable for one coder), or 2 (interpretable for both coders).

In case of France, multi-lingual coders had to work (those who could recognize English, French, Spanish, Italian, German, and, with additional help, Russian). Finding such coders constitutes a separate problem in assessing the quality of multilingual global-scale discussions. We consider the procedures of assignment of meaning comparable for Russia, the USA, and France, as interpreting the topic descriptors requires simple word recognition procedures and basic knowledge of language(s), which makes recognition of a top word the unit of interpretation, disregarding the language belonging of a word.

For H3, topic saliency was automatically calculated based on the modelling results, and the saliency thresholds were defined based on the overall saliency picture for a given case. But, as the saliency measurements have shown, topics could be considered salient if they reached circa 30% of the overall saliency of topics in a given day. Then each topic was assigned a saliency index of 0 (the topic has never reached over 30% of the overall saliency for a particular time slot), 1 (the topic has at least once reached over 50% of saliency for a particular time slot), or 2 (the topic stably reached over 50% of saliency in at least 30% of the overall time span for the case). Then, Spearman's rho was used to see the dependencies between topic interpretability and topic saliency.

For H4 and H5, we have calculated the topic robustness for each topic. First, we have calculated the word relevance and have established 0.02 as the high relevance threshold. Then, we have introduced the robustness score: if a topic has no top words with the

relevance reaching 0.02, it is non-robust (0); if there are up to 5 words of the relevance of 0.02 or higher, the topic is acceptably robust (1); if there are more than 5 words with the relevance of 0.02 or higher, the topic is fully robust (2).

## 6 Results and Discussion

**H1.** We have presupposed that the datasets with a bigger amount of tweets will provide for better topic extraction, if the number of topics is the same. But H1 has to be rejected, as we have found that the number of highly interpretable topics was almost the same for all the datasets: 45 of 99, for Russia; 40 of 96, for the USA; and 41 of 100, for France (topics on non-Russian and non-English were eliminated from the first two cases). Thus, we do not observe any growth of interpretability for bigger tweet collections.

**H2.** The same goes for H2: we do not observe higher interpretability for monolingual cases, given that the coders understand the languages of the multilingual discussion.

Rejection of H1 and H2 is telling, as it provides input for understanding the nature of topic interpretability on Twitter using BTM. Only circa 40–45% of the topics get surely interpreted, which needs to be addressed in the future research. In the French case, most non-interpretable topics were composed by tweets in different languages, first and foremost French, English, and Spanish. The algorithm that puts the tweets in different languages into one topic is still to be analysed, to prevent the mixing of languages in future. Another problem discovered both in the ‘international’ topics and in the francophone one is the over-abundance of pronouns, service words and particles as top words. In the biterm-based approaches, eliminating them from the pre-processed dataset would significantly change the clustering results; but maybe a decision here is simple – they should not be allowed to the lists of topic descriptors, thus leaving more space to the meaningful top words.

**H3–H5.** For the results for H3 to H5, see Table 2.

**Table 2.** Topic quality metrics and their interdependence for the three datasets

|                  | Russia   |          |         | The USA  |          |        | France   |          |          |
|------------------|----------|----------|---------|----------|----------|--------|----------|----------|----------|
|                  | Interpr. | Saliency | Rob.    | Interpr. | Saliency | Rob.   | Interpr. | Saliency | Rob.     |
| Interpretability | –        | –0,080   | 0,064   | –        | 0,226*   | 0,192  | –        | 0,337*** | 0,396*** |
| Saliency         | –0,080   | –        | 0,261** | 0,226*   | –        | –0,086 | 0,337*** | –        | 0,345*** |
| Robustness       | 0,064    | 0,261**  | –       | 0,192    | –0,086   | –      | 0,396*** | 0,345*** | –        |

*Note.* \* -  $p \leq 0, 05$ ; \*\* -  $p \leq 0,01$ ; \*\*\* -  $p \leq 0,001$ .

Table 2 does not provide for any systematic picture of the interdependence of the three topic quality metrics. Our hypotheses were formulated the way that they demanded a clear picture, which is not always the case in practice. In the way they are formulated, they have to be rejected; but important conclusions can be drawn.

Thus, more interpretable topics are more salient in the US and French cases. This might have to do with the dataset volume. And the French case, despite its multilingualism, shows that the three aspects of the topic detection quality may be interdependent.

The correlations for the French case are stronger than for the monolingual cases, and this definitely demands future research on how multilingual discussions are constructed in terms of topicality and why the multilingual discussions perform better than monolingual ones, which is quite counter-intuitive.

To conclude, we underline the following. Putting our results against previous research, we need to state that interpretability for Twitter is significantly lower than for longer-text datasets (cf. [3]). Also, looking at topic/dataset characteristics like robustness and saliency may provide for better understanding of the nature of an ideal topic. In all 300 topics we have assessed, only 7 were all robust, salient, and well-interpretable.


**Acknowledgements.** This research has been supported in full by Russian Science Foundation, grant 16-18-10125-P (2016–2018, prolonged to 2019–2020).

## References

1. Ramage, D., Dumais, S., Liebling, D.: Characterizing microblogs with topic models. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 130–137. AAAI (2010)
2. Blekanov, I., Tarasov, N., Maksimov, A.: Topic modeling of conflict Ad Hoc discussions in social networks. In: Proceedings of the 3rd International Conference on Applications in Information Technology, pp. 122–126. ACM (2018)
3. Koltsova, O., Koltcov, S.: Mapping the public agenda with topic modeling: the case of the Russian live journal. *Policy Internet* **5**(2), 207–227 (2013)
4. Jonnagaddala, J., Jue, T.R., Dai, H.J.: Binary classification of Twitter posts for adverse drug reactions. In: Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, pp. 4–8 (2016)
5. Ligutom, C., Orio, J.V., Ramacho, D.A.M., Montenegro, C., Roxas, R.E., Oco, N.: Using Topic Modelling to make sense of typhoon-related tweets. In: Proceedings of the 2016 International Conference on Asian Language Processing (IALP), pp. 362–365. IEEE (2016)
6. Maceda, L.L., Llovido, J.L., Palaoag, T.D.: Corpus analysis of earthquake related Tweets through topic modelling. *Int. J. Mach. Learn. Comput.* **7**(6), 194–197 (2017)
7. Mazarura, J.R., De Waal, A., Kanfer, F., Millard, S.M.: Topic modelling for short text (2015). [researchgate.net/publication/279195527\\_Topic\\_Modelling\\_for\\_Short\\_Text](https://researchgate.net/publication/279195527_Topic_Modelling_for_Short_Text)
8. Lutovinova, O.V.: Internet as a new ‘oral-written’ system of communication. *Bull. Russ. State Pedagogical Univ.* **71**, 58–65 (2008). <https://cyberleninka.ru/article/n/internet-kak-novaya-ustno-pismennaya-sistema-kommunikatsii>
9. Smoliarova, A.S., Bodrunova, S.S., Yakunin, A.V., Blekanov, I., Maksimov, A.: Detecting pivotal points in social conflicts via topic modeling of Twitter content. In: Bodrunova, S.S., et al. (eds.) INSCI 2018. LNCS, vol. 11551, pp. 61–71. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-17705-8\\_6](https://doi.org/10.1007/978-3-030-17705-8_6)
10. Sridhar, V.K.R.: Unsupervised topic modeling for short texts using distributed representations of words. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pp. 192–200 (2015)



# Making Reproducible Research Simple Using RMarkdown and the OSF

André Calero Valdez<sup>(✉)</sup> 

Human-Computer Interaction Center, RWTH Aachen University,  
Campus Boulevard 57, 52076 Aachen, Germany  
`calero-valdez@comm.rwth-aachen.de`

**Abstract.** The replication crisis has further eroded the public’s trust in science. Many famous studies, even published in renowned journals, fail to produce the same results when replicated by other researchers. While this is the outcome of several problems in research, one aspect has gotten critical attention—reproducibility. The term reproducible research refers to studies that contain all materials necessary to reproduce the scientific results by other researchers. This allows others to identify flaws in calculations and improve scientific rigor. In this paper, we show a workflow for reproducible research using the R language and a set of additional packages and tools that simplify a reproducible research procedure.

**Keywords:** Reproducible research · Replication crisis · Literate programming

## 1 Introduction

The scientific database Scopus lists over 73,000 entries for the search term “reproducible research” at the time of writing this document. The importance of making research reproducible was recognized in the early 1950s in multiple research subjects. And with the reproducibility project, the Open Science Foundation [26] found that merely half of all studies conducted in psychological research can be replicated by other researchers. Several factors have contributed to this problem. From a high-level perspective, the pressure to publish and the increase in scientific output has led to a plethora of findings that will not replicate. Both bad research design and (possibly unintentional) bad research practices have increased the number of papers that hold little to no value. More than half of researchers agree that there is a severe reproducibility crisis in science according to Baker [3] and her article in *Nature*. The study also found that problems for reproducibility include: a lack of analysis code availability, a lack of raw data availability, and problems with reproduction efforts.

## 2 Problematic Research Practices

One problem that is often mentioned is HARKing [16] or “hypothesizing after results are known”. When multiple statistical tests are conducted with a normal



alpha-error rate (e.g.,  $\alpha = .05$ ), it is expected that some tests will reject the null-hypothesis on mere randomness alone. Hence, the error-rate. If researchers now claim that these findings were their initial hypotheses, results will be indiscernible from randomness. However, this is unknown to the reviewer or reader who only hears about the new hypotheses. HARKing produces findings were there are none. It is thus crucial to determine the research hypothesis before collecting (or analyzing) the data.

Another strategy applied (often without ill intent) is p-hacking [13]. This technique is widespread in scientific publications and probably already is shifting consensus in science. p-hacking refers to techniques that alter the data until the desired  $p$ -value is reached. Omitting individual *outliers*, creating different grouping variables, adding or removing control variables—all these techniques can be considered p-hacking. This process also leads to results that will not hold under replication. It is crucial to show what modifications have been performed on data to evaluate the interpretability of  $p$ -values.

When researchers already “massage” the data to attain *better*  $p$ -values, it is additionally bad that many researchers do not understand the meaning of  $p$ -values. As Colquhoun [9] found, many researchers misinterpret  $p$ -values and thus frame their findings much stronger than they really are. Adequate reporting of  $p$ -values is thus important to the interpretability of results as well.

Lastly, scientific journals have the problem that they are mostly interested in publishing significant results. Thus contradictory “non-findings” seldom get published in renowned journals. There is little “value” for a researcher to publish non-significant findings, as the additional work to write a manuscript for something like *arXiv* does often not reap the same reward as a journal publication. This so-called *publication bias* [29] worsens the crisis. As now only significant findings are available. It is thus necessary to simplify the process of publishing non-significant results.

### 3 Reproducible Research Workflows

Many different solutions to this process have been proposed to address these challenges (e.g., [22,37]). However, no uniform process exists that allows the creating of documents and alternative reproducibility materials in one workflow.

In this paper, we demonstrate a research workflow based on the R-language and the R Markdown format. This paper was written using this workflow and the sources are freely available online (<https://www.osf.io/kcbj5>). Our workflow directly addresses the challenge of writing LNCS papers and a companion paper website (<https://sumidu.github.io/reproducibleR/>) that includes additional material and downloadable data.

In this paper, we will focus on the following aspects:

- Creating a reproducible research compendium using RMarkdown
- Using GitHub and the OSF to make research accessible
- Packages that simplify research in RStudio

We assume that the reader is somewhat familiar with the R Programming language and knows that scientific analyses can be run using computational tools such as R, Python, Julia or others. The guidance in this paper addresses the R user.

### 3.1 What Is Reproducibility?

The Open Science Foundation (OSF) speaks of three different kinds of reproducibility [24]. *Computational reproducibility* refers to the quality of research that when other researchers get access to your code and data that they will be able to reproduce your results. *Empirical reproducibility* means that your research has sufficient information that allows other researchers to recreate your experiments and copy your study. *Replicability* refers to the quality of an outcome and a study, meaning that given that you were to reproduce the experiment, you would also reach the same outcome. In this article, we provide tools for the first type of reproducibility only, as the latter are both dependent on your research content not exclusively on your procedure. It is important to note that creating computationally reproducible research is important, but it is also worthless when basic concepts of methods and research processes are ignored. If you measure incorrectly, your result may reproduce, but the finding may be wrong anyways. Hopefully, when you are using the suggested workflow here, others will be able to point out mistakes to you more easily.

## 4 Writing a Research Compendium

The central aim of a research compendium is to provide all data and information necessary to allow others to reproduce your findings from your data [12]. There are several different ways of achieving this but a central theme of a research compendium is to organize data in a meaningful fashion. Since we are addressing R users, it makes sense to consider possible computing environments for R first.

You can find detailed information on how to create a research compendium online here <https://research-compendium.science/>.

### 4.1 Why R and RMarkdown?

R is the de-facto standard when it comes to statistical analysis tools that are open source and free to use. In economics and the social sciences, similar tools that provide a GUI like SPSS are used with one immediate downside for reproducibility. If your analysis toolkit is proprietary, other users will not be able to reproduce your work without a significant investment.

Moreover, using a GUI makes it untraceable—even to yourself—what analyses you have conducted later. You might have manually deleted a row with broken data, or might have recoded a typing error in your data manually. If this is not documented, this information is lost. Using a language like R, where every

change of the data corresponds to a line of code, no accidental “quick fixes” will get lost over time. R also provides a rich set of tools for reproducible research on CRAN<sup>1</sup>.

## 4.2 Literate Programming

*RMarkdown* is a tool that is extremely helpful for researchers, as it allows us to combine analysis code with regular text. This document was written using RMarkdown and integrating some analysis code in between. RMarkdown is a *literate programming* approach. The documentation of code is equally necessary for understanding the code, as the code itself. By interleaving code and text, the intentions of the developer are implicitly communicated. Python and Julia have similar approaches by using Jupyter notebooks.

RMarkdown allows not only for the integration of text and figures directly from code, but it also allows writing in an abstract format. A single document (such as this) can be rendered to various output formats. In this case, it is rendered to the LNCS styled Latex output format, as well as to a website using bootstrap. The benefit is that text and code are reusable, so when papers get rejected no excessive reformatting has to be made. Formatting is done using Markdown (see [here](#)<sup>2</sup> for a tutorial). Code and analyses are interleaved in text in so-called “code chunks”. Code chunks can contain R code, but also code from other languages (e.g., Python).

## 4.3 Project Workflows

The most popular integrated development environment (IDE) for R is RStudio. RStudio comes with a license that allows researchers to freely use it for scientific purposes and it integrates many of the tools described in this paper. The first strong tool for reproducible research using R is using RStudio projects.

RStudio projects contain information about where your code, your data, and your output should reside on your computer. The benefit of RStudio projects is that they contain relative path information, so when another user installs your project on their computer, it should work without a problem. Since you need to refer to files in some cases, even relative paths work well. The `here` package provides a helpful tool to access data relative to the project main directory. This works on Linux, Windows, and Mac computers.

## 4.4 Package Management

Another key requirement for computational reproducibility is that the software versions on different computers actually produce the same analysis. This is most safely achieved by keeping all libraries in the same version as in the original analysis. Sometimes libraries change their features and this can render old projects

---

<sup>1</sup> <https://cran.r-project.org/web/views/ReproducibleResearch.html>.

<sup>2</sup> <https://www.markdowntutorial.com/>.

unusable. Using package management is not only a necessity for computational reproducibility, but it is also helpful for yourself when you get back at a project. There are several tools in the R universe that address this challenge. All of them have different efforts involved and provide different benefits.

The **packrat** package [32] comes integrated into RStudio and allows you to create a localized copy of the used libraries in your analysis. Packrat even downloads sources of these packages and allows using libraries from different sources (CRAN, GitHub, etc.). Packrat provides a function (`packrat::bundle()`) to pack everything into a shareable file. However, packrat sometimes has problems with multiple RMarkdown files in the project, causing it to re-render all documents to infer the used packages. In these cases packrat is not a viable solution. When you find your project becoming very slow, it might make sense to remove packrat.

This is where the **renv** package [31] comes into play. It is a simplified version of package management and runs relatively reliable even with multiple RMarkdown files in the project. By calling `renv::init()` a lock file is created that contains information on all packages used in the project. It does, however, not download sources, so it will only work if the packages you use are expected to be available in the future as well.

Neither of these options though addresses the challenge of using the same R version or the same operating system. Differences between Windows and Linux could yield different results in the future. This is where **docker** comes into play. Docker is a light-weight virtualization software that allows you to run a virtual machine based on other users' machine images. It also provides a sharing platform for these images. Rocker provides a set of default virtual machine images<sup>3</sup> that contain both a fixed R version and a set of libraries usable in research. By adding a `dockerfile` to a project, you can create a definition of your project that will build a matching machine image. Docker does require the user to install the docker software on their machine.

There are options for sharing your run-time environment without asking other researchers to install any software. RStudio comes with a cloud version that can (currently) be used for free if the projects are either private or completely public. By running your project in `rstudio.cloud` and sharing the public link to a project, others can create copies of your run-time environment in their **rstudio.cloud** account.

An even simpler version is the use of **binder**. Binder can be set up to automatically build your project on a virtual machine and provide an RStudio instance on the virtual machine that has access to your project. However, to make this work, you need to use a version control system, more specifically you need to use GitHub with your project. The auto-generated binder link from the [www.mybinder.org](https://www.mybinder.org) website can be extended to use RStudio by adding `?urlpath=rstudio` to the URL. You can have a look at the readme of this project on GitHub to see how it is done.

---

<sup>3</sup> <https://github.com/rocker-org/rocker>.

## 4.5 Writing Articles Using `rmdtemplates`

The package `rmdtemplates` [7] provides templates for writing RMarkdown files that adhere to the Lecture Notes in Computer Sciences series. It also contains a template for an open data website. This website allows creating a reader-friendly version of your paper to send around. Moreover, you can add additional analyses in the open data website which can then be added as supplementary materials to your paper. You must ensure though that copyrights are respected when sharing the written content of your paper.

Key benefits of the `rmdtemplates` package are that it supports to automatically generate citations for the R packages that you use. RMarkdown uses `bibfiles` to store your references. To make referencing easier, install the `citr` [2] package in RStudio to enable a GUI to use references from your library. `Citr` allows connecting your Zotero database and using those references as well.

## 5 Open Data and Open Code

One key idea of reproducibility is making data and analysis code openly available. Sharing your code allows other researchers to inspect it and verify that your results are valid conclusions from your analyses. Research and statistical analyses are complex processes and mistakes are bound to happen sometime. Mistakes are less severe when they can be retraced and results adapted. Sharing your data allows other researchers to see what other information might have been undiscovered by your analyses. Studying open data can be used to explore new theories, used in meta-analyses and be used in teaching settings. Typically data is released under the CC0 license, making data part of the public domain.

### 5.1 Data Sharing and Anonymization

Sharing data is not just uploading your data to a website. First, several considerations must be made before data can be shared. The most important question is: “Does my data contain personal information?” Any data that was collected on human subjects potentially contains personal information. This has several implications.

First, you must ask whether the participants agreed with data sharing. Typically, participants sign data waivers allowing researchers to use data for scientific purposes. It is important to inform participants about the possibilities of sharing.

Second, information on people can be damaging to these people upon release. By allowing others to utilize your data, you must consider possible threats to your participants before deciding what data to release. It is crucial to inform yourself about data and anonymization before carelessly releasing information. For example, releasing information on your participants when they were students from a certain semester might leave individuals identifiable in your data set. Thus it may be necessary to either anonymize your data or to limit additional

information on data gathering procedures (or both). It makes sense to speak to an expert on anonymization about this topic and to ask for permission from your organization's ethics board.

**K-Anonymity.** The simplest form of anonymity can be generated if all quasi-identifiers of a person appear multiple times in the database from several other persons. Each person is then represented by the same data attributes so that each person can no longer be distinguished from other persons with the same attributes. This concept is called  $k$  anonymity [1]. If at least  $k$  persons exist in a given data set, who are identically represented in terms of their quasi-identifiers, the data is  $k$ -anonymized. Each person is now in an *equivalence class* of at least  $k - 1$  other people who share the same *quasi-identifiers*.

This method provides an intuitive version of privacy that is both algorithmically simple to implement and easy to explain to the participant. Technically, we can simply add noise to the data to enhance privacy. For example, we can remove the last digits of postal codes (data deletion) until at least  $k$  equal entries exist for each postal code. We can also store age groups instead of birth dates (data aggregation).

The advantage of this method is that our data is only slightly changed, as only the quality of the data is reduced. One problem is not solved with this method. It could be that the combination of quasi-identifiers and sensitive data could still be too informative for an external attacker. An insurance company might want to know that all persons from a region aged 65 and older suffer from heart disease. Even if no person is de-anonymized in this scenario, all persons in the data set may suffer from the consequences of possible secondary use of the data.

The most important finding of  $k$ -anonymity is that the most important problem in anonymization is not user identification, but data sensitivity and the possibilities of secondary use. For this purpose  $l$ -diversity [20] or  $t$ -closeness [19] may be considered. A package for R that provides an interactive tool for applying anonymization techniques to a data set is the `sdcMicro` [30] package. Another option is the `anonymizer` [14] package which provides methods for detecting potentially identifying information and replacing it with hashes.

**Differential Privacy.** Often other scientists are not interested in individual data. If only the statistical properties of a data set are interesting, it should be easy to ensure the privacy of individuals. However, it is still possible to obtain sensitive information about individual users by repeatedly querying a database.

Attackers can combine multiple queries to narrow down sensitive information about individuals. The idea behind Differential Privacy is to establish a privacy budget [11]. Whenever statistics are calculated on the data, the amount of information in these statistics is deducted from this privacy budget. This is achieved by replacing data with noisy data. This means that two identical database queries will most likely yield different results. The more queries are received, the more different the results become until the database returns only noise. The database must

now either be discarded or new data must be collected to increase the budget for data protection.

The advantage of Differential Privacy is that there is a mathematically guaranteed privacy for each user. It is therefore impossible to gain knowledge about individuals from the retrieved information [18].

## 5.2 GitHub and Git

Sharing of code is a procedure that is natural to computers scientists, as almost all larger software products are team efforts. GitHub has crystallized as the default standard of sharing code for open-source software. What is GitHub and how do you use it?

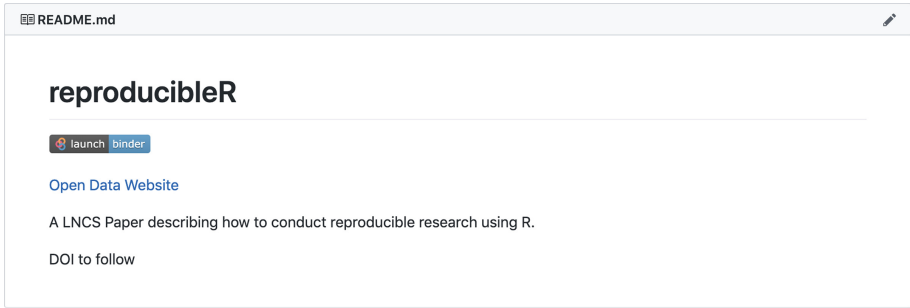
First, we must understand **Git**. Git is a version control software (VCS). Git allows you to keep track of changes in your files. It allows you to store individual changes as so-called *commits*. Each individual commit can always be restored from the git *repository* on your computer. This gets read of the challenge of keeping multiple version files of a document. Git works completely locally, so you can move project folders that are tracked by git around on your computer or someone elses computer, without losing tracking information.

RStudio is completely integrated with Git, so committing new versions of your project is as simple as a click. Git has proven to be the most valuable tool in literate programming for science [6].

**GitHub** is a website that provides free repositories for open source software or open-source research. GitHub allows you and your collaborators to work on the same project asynchronously. By uploading (called *pushing*) your local git repository to the public GitHub repository your collaborators or other researchers get access to this project. These people can now download the repository (called *pulling*) to their computer and work on the project or reproduce your analysis. Git has extensive mechanisms for merging your progress and your collaborator's project progress. Changes can be integrated on a line-by-line basis. Thus it is best to break lines in your code frequently.

It is important to note that GitHub is not the best place to store your data. Individual files are limited to 100 MB and projects are limited to 2 GB.

**GitHub Readme.** Beyond providing a publicly available place to store your analysis code. GitHub serves as a publicly accessible website for your research project. It is recommended to upload a `README.md` file that contains basic information about your research project. It could contain a DOI of the published article, it could contain links to other parts of the project such as data stores on the web. The benefit of GitHub readme files is that they will automatically render a pretty HTML output on the website (see Fig. 1).



**Fig. 1.** Rendered Readme.MD file on GitHub.

**GitHub Pages.** If you have generated your analysis using RMarkdown you can render your output to a website as well. This provides the benefit of adding additional figures and making your document more accessible. By using libraries such as `plotly` other researchers can even explore your data using interactive visualizations. The template from the `rmtemplates` [7] package provides a nice pre-structured interactive website that allows you to include tabular downloadable data in the website.

When you store your projects on GitHub, you can make your website publicly available easily. By copying the output format to a sub-folder called `docs` and enabling GitHub Pages in your GitHub settings your page is exposed to the public without requiring a hosting service (except for GitHub).

### 5.3 OSF

While GitHub is an excellent provider for storing the code of your analysis, it is not very well suited for sharing data and for reviewing purposes. The Open Science Foundation (OSF) provides a service where researchers can create projects that have Wikis, file storage, and transparent referencing. You may even choose the server where your data is stored when data protection laws require your data to be in your country.

A key benefit of the OSF is that you can create sub-modules in your project and share the whole project or the sub-modules individually with others. Each “node” in your project gets an easy-to-recognize and short unique URL which can be added to a paper or a website. More importantly, it allows the sharing of parts of your project anonymously, during the reviewing process. The reviewers can see the available data, without seeing the authors’ names. But they also can verify that data has not been changed since the project has gone public.



To make things even better, there is a package called `osfr` [38] that lets you download (or upload) your data directly from your R or RMarkdown documents. This can be leveraged in the setup procedures of a document, as in: “If the data is not available, try downloading it automatically.”

**Preregistration.** Another benefit of the OSF is that you may preregister your study before collecting data. By setting up preregistration at the OSF and setting up the rest of your project before collecting data, you prevent yourself from HARKing in your research. Preregistered trials are part of the gold standard of high-quality social science research. Some journals have decided to accept studies after preregistration to prevent a publication bias towards significant findings. This does not mean that you can no longer conduct exploratory research on your data, it simply ensures that confirmatory and exploratory findings are clearly separated.

## 6 Helpful Tools for Everyday Tasks

In this section, we will introduce some tools that make life as a researcher easier, when relying on computational analysis using R and RStudio.

### 6.1 Automizing Builds Using Drake

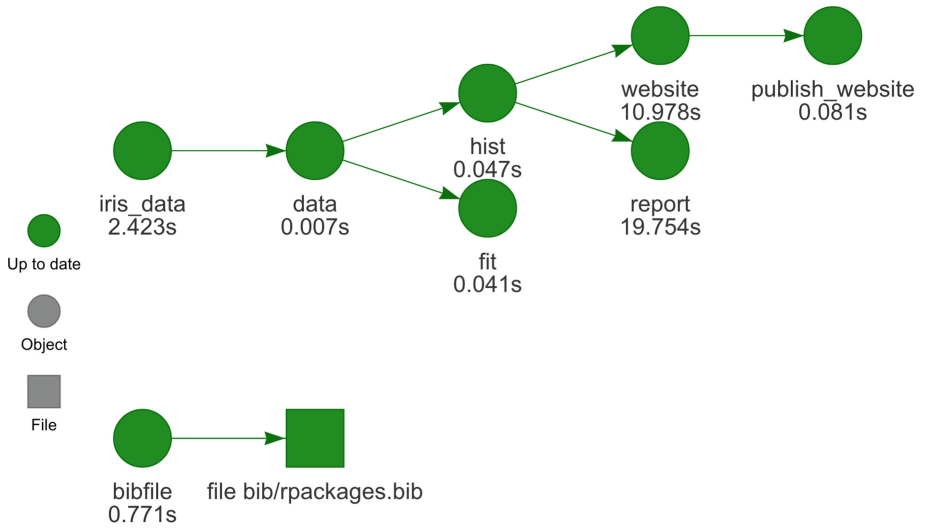
One large challenge for literate programming is that for every small change in a document you need to re-run your complete analysis. This is not a problem with modern computers in many cases, but when your data is large enough, it can become a hassle. While RMarkdown does provide a caching mechanism, it is limited to the individual computer and may not be shared among researchers. And when individual code chunks depend on external data that has changed, RMarkdown caching no longer registers these changes.

The package `drake` [17] addresses this challenge. By creating a plan using the `drake_plan` function we first determine what steps are necessary for our analysis. Drake then analyzes our code and files for implicit dependencies. It derives a dependency tree (see Fig. 2) that visually shows how the project should run.

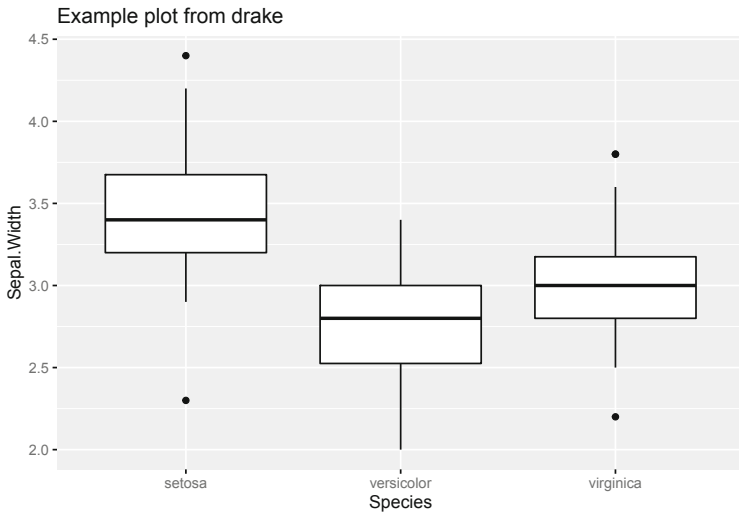
Each build target (e.g., `report`) is the call to an R function. The result of which can easily be reloaded anywhere inside the project using the `load` or `read` function. In the simplistic build path for this document, the `hist` target creates a histogram of the `iris.data`, which is downloaded from the OSF. The figure (see Fig. 3) is then included in the document. Whenever a single previous dependency becomes outdated, the rest of the dependency graph is executed.

When all dependencies are properly modeled, drake allows running individual targets on multiple CPUs or a compute cluster without much overhead. The interested reader can take a look at the `make.R` file in the project root of this document.

### Dependency graph



**Fig. 2.** The drake plan to generate this document.



**Fig. 3.** This figure was created outside of the document in the hist target.

## 6.2 Ensuring Relative Paths Using `here`

The `here` [25] package provides a useful tool to find files on all operating systems. Windows and Unix are famous for using different slashes as directory separators, which can make addressing file paths complicated for reproducible research. By encapsulating all file operations in the `here` function, the relative root is set to the destination of the R project file and directory separators are automatically inserted in accordance with the local operating system.

## 6.3 Automating Project Setup Using `usethis`

When setting up a project, many tasks have to be conducted over and over again. To simplify this the `usethis` [35] package provides a set of tools to start working on a project. A typical workflow for setting up a project using `usethis` could look like this.

1. Create a project using `create_tidy_project`
2. Setup the license using, e.g., `use_mit_license`
3. Setup using git `use_git`
4. Setup using GitHub `use_github`
5. Setup a readme using `use_readme_rmd`
6. Setup a citation for your project `use_citation`

Running `git_vaccinate` once will add all files that typically contain credentials or other personal information to the global git-ignore file, preventing them from being accidentally shared.

Another library that helps with setting up a project is the `rrtools` [21] package.

## 6.4 Onboarding New Users to RStudio with `Addins`

Learning how to write code can be hard for someone changing over from UI-based approaches such as SPSS. However, there are several tools that help you create reproducible R code from the UI of RStudio. Such tools can be found in the “Addins” menu and we will highlight some that make research easier.

Working with factors is not always easy in R. They will appear in alphabetic order in plots, they are hard to rename and hard to reorganize. The `forcats` [33] package simplifies the use of factors, by unifying the interface to them. An addin from the `questionr` [5] package allows for easy recoding and relabeling of factors.

Creating custom plots using the `ggplot2` package creates usable figures for scientific papers. Yet, it may take a while to get accustomed to `ggplot`. The `esquisse` [23] package has an interactive addin (see Fig. 4) that lets you drag and drop variables to axes, adjust layout and color, filter the data, and export the script that would generate the matching plot.

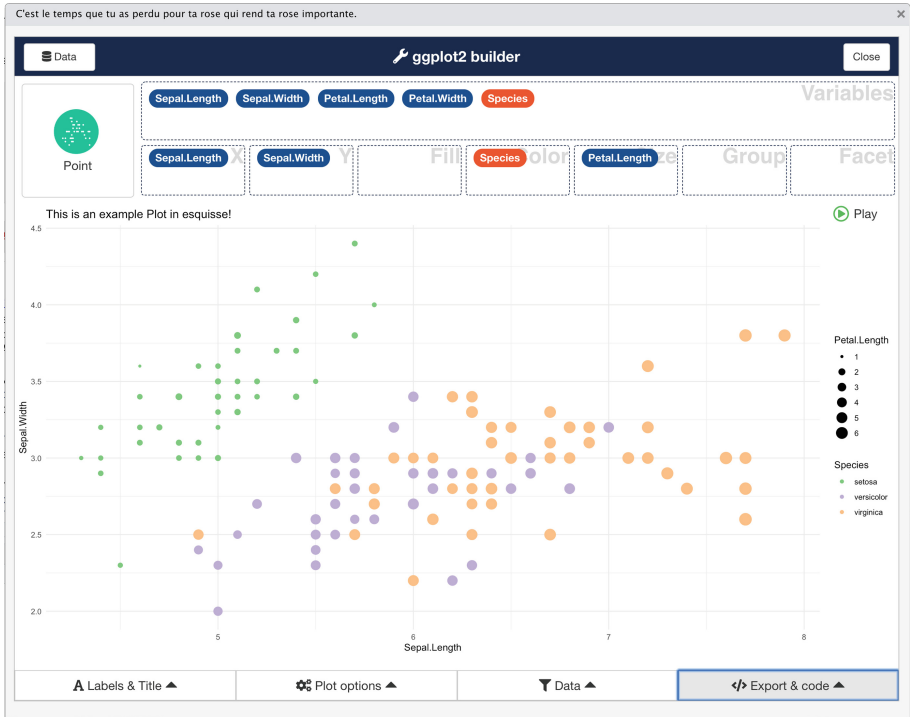


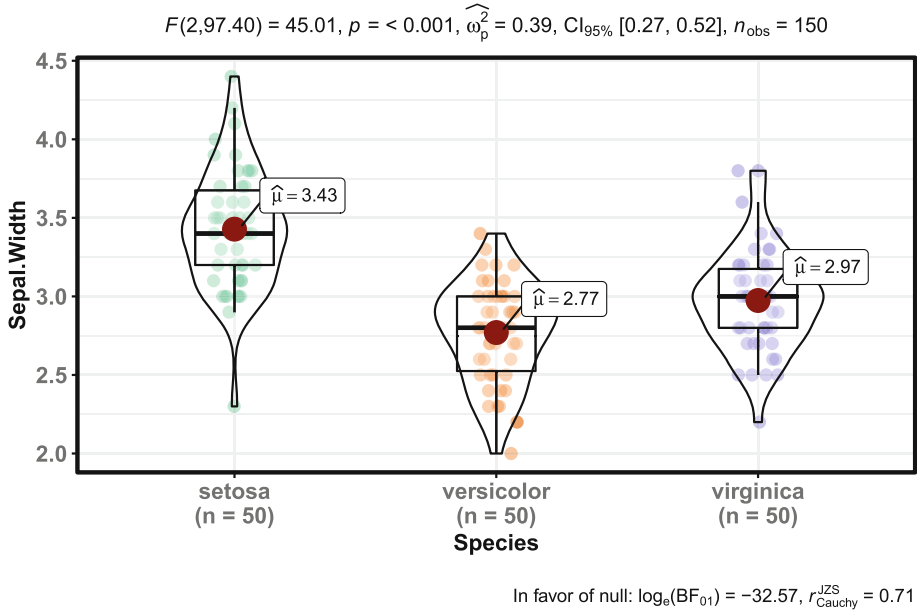
Fig. 4. This is the interface to the ggplot builder from esquisse.

### 6.5 Using ggstatsplot to Generate Meaningful Statistical Analyses

Knowing how to report a statistical finding often requires a deep understanding of what test to use and how to report the results adequately. The `ggstatsplot` [27] package provides a set of plotting functions that use very sensible defaults derived from the data input. If, for example, you want to compare means between multiple groups the `ggbetweenstats` function will produce a nice plot with all relevant statistical information—including effects sizes, confidence intervals, and Bayes factors.

The upcoming plot (Fig. 5) was created from a single line of code.

```
ggbetweenstats(iris, Species, Sepal.Width, messages = F)
```



**Fig. 5.** A comparison of means using the iris data.

## 6.6 Create Research Plans Using DiagrammeR

Even process diagrams as in Fig. 6 can easily be created using the DiagrammeR [15] package. It requires writing a process description in the dot language which is relatively easy to learn.

```
library(DiagrammeR)

grViz(diagram = "
  digraph boxes_and_circles {

  graph [rankdir = TB]

  node [shape = box
        fontname = Helvetica
        ]
    'Setup OSF Project Site'
    'Setup R Project'
    'Setup GitHub Repo'
    'Ensure reproducibility using renv'
    'Write analysis'
    'Preregister Study'
    'Collect Data'
```

```
node [shape = circle]

Start
'Submit Paper'

edge []

Start->'Setup OSF Project Site';
'Setup OSF Project Site'->'Setup R Project';
'Setup R Project'->'Setup GitHub Repo';
'Setup GitHub Repo'->'Ensure reproducibility using renv';
'Ensure reproducibility using renv'->'Write analysis';
'Write analysis'->'Preregister Study';
'Preregister Study'->'Collect Data';
'Collect Data'->'Submit Paper'
}
")
```

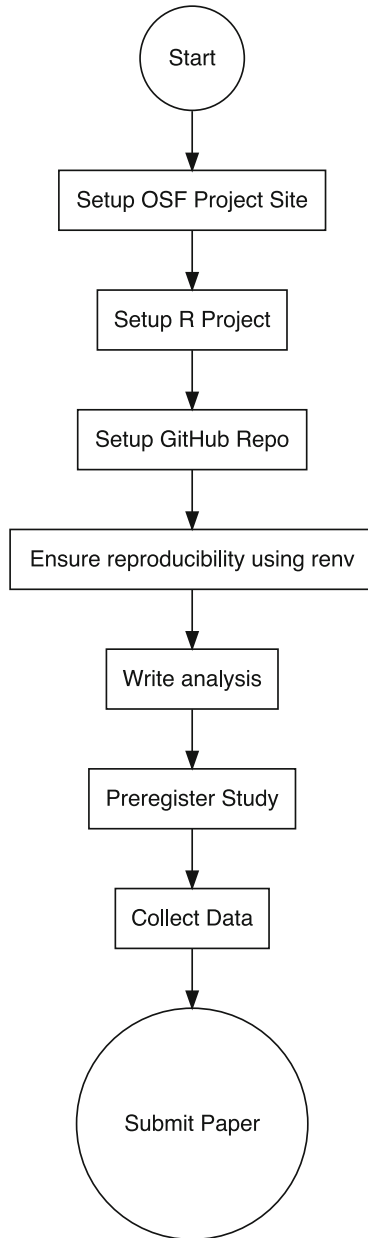
## 7 Discussion

In this paper, we have very shortly introduced a large number of tools that can be utilized to make research more computationally reproducible. By integrating the different tools into a complete workflow, emergent effects from the interactions of the individual steps can be reaped. However, this workflow can easily become overwhelming as it includes multiple tools, each of which has large documents attached to them explaining how to use them. You as a reader may decide on which tools to use from this set and which to ignore. Be warned though, that skipping some of the steps will reduce the benefits for other researchers and yourself.

We did not extensively elaborate on how some of these tools are used most effectively. We leave it up to you to deepen the knowledge of some of these tools. In the future, we want to create additional tutorial materials made available on the website for this paper<sup>4</sup>.

---

<sup>4</sup> (<https://sumidu.github.io/reproducibleR/>).



**Fig. 6.** Reproducible workflow using the tools from this paper.

**Acknowledgements.** This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia. We would further like to thank the authors of the

packages we have used. We used the following packages to create this document: `knitr` [39], `tidyverse` [34], `rmdformats` [4], `kableExtra` [40], `scales` [36], `psych` [28], `rmdtemplates` [7], `sdcMicro` [30], `webshot` [8], `here` [25], `DiagrammeR` [15], `citr` [2], `drake` [17], `esquisse` [23], `usethis` [35], `gramr` [10], `questionr` [5], `ggstatsplot` [27].

## References

1. Aggarwal, C.C., Philip, S.Y.: A general survey of privacy-preserving data mining models and algorithms. In: Aggarwal, C.C., Yu, P.S. (eds.) *Privacy-preserving data mining*, vol. 34, pp. 11–52. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-0-387-70992-5\\_2](https://doi.org/10.1007/978-0-387-70992-5_2)
2. Aust, F.: `citr`: RStudio Add-in to Insert Markdown Citations. R package version 0.3.2. (2019). <https://CRAN.R-project.org/package=citr>
3. Baker, M.: Reproducibility crisis. *Nature* **533**(26), 353–66 (2016)
4. Barnier, J.: `rmdformats`: HTML Output Formats and Templates for ‘rmarkdown’ Documents. R package version 0.3.6. (2019). <https://CRAN.R-project.org/package=rmdformats>
5. Barnier, J., Briatte, F., Larmarange, J.: `questionr`: Functions to Make Surveys Processing Easier. R package version 0.7.0. (2018). <https://CRAN.R-project.org/package=questionr>
6. Bryan, J.: Excuse me, do you have a moment to talk about version control? *Am. Stat.* **72**(1), 20–27 (2018)
7. Valdez, A.C.: `rmdtemplates`: `rmdtemplates` - an opinionated collection of rmarkdown templates. R package version 0.4.0.0000. (2020). [https://github.com/statisticsforsocialscience/rmd\\_templates](https://github.com/statisticsforsocialscience/rmd_templates)
8. Chang, W.: `webshot`: Take Screenshots of Web Pages. R package version 0.5.2. (2019). <https://CRAN.R-project.org/package=webshot>
9. Colquhoun, D.: The reproducibility of research and the misinterpretation of p-values. *Roy. Soc. Open Sci.* **4**(12), 171085 (2017)
10. Dumas, J., Marwick, B., Shotwell, G.: `gramr`: The Grammar of Grammar. R package version 0.0.0.9000. (2020). <https://github.com/ropenscilabs/gramr>
11. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) *TCC 2006*. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). [https://doi.org/10.1007/11681878\\_14](https://doi.org/10.1007/11681878_14)
12. Gentleman, R., Lang, D.T.: Statistical analyses and reproducible research. *J. Comput. Graph. Stat.* **16**(1), 1–23 (2007)
13. Head, M.L., et al.: The extent and consequences of p-hacking in science. *PLoS Biol.* **13**(3), e1002106 (2015)
14. Hendricks, P.: `anonymizer`: Anonymize data containing personally identifiable information. R package version 0.2.2. (2020). <https://github.com/paulhendricks/anonymizer>
15. Iannone, R.: `DiagrammeR`: Graph/Network Visualization. R package version 1.1.0. (2020). <https://github.com/rich-iannone/DiagrammeR>
16. Kerr, N.L.: HARKing: hypothesizing after the results are known. *Pers. Soc. Psychol. Rev.* **2**(3), 196–217 (1998)
17. Landau, W.M.: `drake`: A Pipeline Toolkit for Reproducible Computation at Scale. R package version 7.10.0. (2020). <https://CRAN.Rproject.org/package=drake>
18. Lee, J., Clifton, C.: How much is enough? choosing  $\epsilon$  for differential privacy. *Inf. Secur.* **7001**, 325–340 (2011)



19. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering, pp. 106–115. IEEE (2007)
20. Machanavajjhala, A., et al.: l-diversity: privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data (TKDD)* **1**(1), 3 (2007)
21. Marwick, B.: rrtools: Creates a Reproducible Research Compendium. R package version 0.1.0. (2019). <https://github.com/benmarwick/rrtools>
22. Marwick, B., Boettiger, C., Mullen, L.: Packaging data analytical work reproducibly using R (and friends). *Am. Stat.* **72**(1), 80–88 (2018)
23. Meyerm, F., Perrier, V.: esquisse: Explore and Visualize Your Data Interactively. R package version 0.3.0. (2020). <https://CRAN.Rproject.org/package=esquisse>
24. Meyers, N.K.: Reproducible Research and the Open Science Framework (2017). <https://osf.io/458u9/>
25. Müller, K.: here: A Simpler Way to Find Your Files. R package version 0.1. (2017). <https://CRAN.R-project.org/package=here>
26. Open Science Collaboration et al.: Estimating the reproducibility of psychological science. *Science* **349**(6251), aac4716 (2015)
27. Patil, I.: ggstatsplot: “ggplot2” Based Plots with Statistical Details. R package version 0.2.0. (2020). <https://CRAN.R-project.org/package=ggstatsplot>
28. Revelle, W.: psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 1.9.12.31. (2020). <https://CRAN.R-project.org/package=psych>
29. Simonsohn, U., Nelson, L.D., Simmons, J.P.: p-curve and effect size: correcting for publication bias using only significant results. *Perspect. Psychol. Sci.* **9**(6), 666–681 (2014)
30. Templ, M., Meindl, B., Kowarik, A.: sdcMicro: Statistical Disclosure Control Methods for Anonymization of Data and Risk Estimation. R package version 5.5.1. (2020). <https://CRAN.Rproject.org/package=sdcMicro>
31. Ushey, K.: renv: Project Environments. R package version 0.9.3-30. (2020). <https://rstudio.github.io/renv>
32. Ushey, K., et al.: packrat: A Dependency Management System for Projects and their R Package Dependencies. R package version 0.5.0. (2018). <https://CRAN.R-project.org/package=packrat>
33. Wickham, H.: forcats: Tools for Working with Categorical Variables (Factors) (2020). <http://forcats.tidyverse.org>, <https://github.com/tidyverse/forcats>
34. Wickham, H.: tidyverse: Easily Install and Load the ‘Tidyverse’. R package version 1.3.0. (2019). <https://CRAN.R-project.org/package=tidyverse>
35. Wickham, H., Bryan, J.: usethis: Automate Package and Project Setup. R package version 1.5.1. (2019). <https://CRAN.Rproject.org/package=usethis>
36. Wickham, H., Seidel, D.: scales: Scale Functions for Visualization. R package version 1.1.0. (2019). <https://CRAN.R-project.org/package=scales>
37. Wilson, G., et al.: Good enough practices in scientific computing. *PLoS Comput. Biol.* **13**(6), e1005510 (2017)
38. Wolen, A., Hartgerink, C.: osfr: Interface to the ‘Open Science Framework’ (‘OSF’). R package version 0.2.8. (2020). <https://CRAN.Rproject.org/package=osfr>
39. Xie, Y.: knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.28. (2020). <https://CRAN.Rproject.org/package=knitr>
40. Zhu, H.: kableExtra: Construct Complex Table with ‘kable’ and Pipe Syntax. R package version 1.1.0. (2019). <https://CRAN.R-project.org/package=kableExtra>



# Visual Saliency: How Text Influences

Ying Fang<sup>1</sup>, Liyu Zhu<sup>1</sup>, Xueni Cao<sup>1</sup>, Liqun Zhang<sup>1</sup>(✉), and Xiaodong Li<sup>2</sup>(✉)

<sup>1</sup> School of Design, Shanghai Jiao Tong University, Shanghai, China  
zhangl1liqun@gmail.com

<sup>2</sup> China National Gold Group Gold Jewellery Co., Ltd., Beijing, China  
lixiaodong@chnau99999.com

**Abstract.** Posters are widely used a powerful tool for communication. They are very informative but are normally viewed for only 3 s, which calls for efficient and effective information delivery. It is thus important to know where people would look for posters. Saliency models could be of great help where expensive and time-consuming eye-tracking experiment isn't an option. However, current datasets for saliency model training mainly deal with natural scenes, which makes research on saliency models for posters difficult. To address this problem, we collected 1700 high-quality posters as well as their eye-tracking data where each image is viewed by 15 participants. This could be the groundwork for future research in the field of saliency prediction for posters. It is noticeable that posters are rich in texts (e.g. title, slogan, description paragraph). The various types of texts serve respective functions, making some relatively more important than others. Nevertheless, the difference is largely neglected in current studies where researchers put same emphasis on all text regions, and the problem is especially crucial when it comes to saliency model for posters. Our further analysis of the eye-tracking results with focus on text offers some insights into the issue.

**Keywords:** Visual saliency prediction · Eye-tracking dataset of posters · Relative importance of different text regions

## 1 Introduction

Poster is a popular and rather unique means of information delivery. Unlike photos or artworks which are more likely to be carefully looked at, they are normally viewed for around 3 s [1]. Plus, they are often rich in context, which requires higher level of semantic understanding. All these call for efficient and high-quality information delivery, and it is thus important to know where people would pay more attention to when they look at posters.

Along with the progress of machine learning, studies on visual saliency models using neural networks have gain strong momentums these years. Visual saliency models can make predictions for fixations of viewers, offering help where expensive, tedious and time-consuming eye-tracking experiment isn't an option [2]. Most saliency models deal with prediction of fixation points of natural scenes, and few other studies focus on other image types such as webpages and infographics [3]. This is partly due to the fact that

current datasets used for saliency model training basically consist of natural scenes. MIT300 which contains 300 natural images is one of the most frequently used datasets, and the same researchers later established MIT1003 which consists of 1003 landscapes and portraits [4, 5]. Other datasets include UNCEF (758 images, many of them emotion-evoking) [6], AVA (over 250,000 photography work) [7], CAT2000 (4000 images from 20 categories) [8] etc. Shen and Zhao [9] provided one of the very few datasets for webpages which contains 149 webpages and eye-tracking from 11 participants. To our knowledge, there isn't yet a dataset of posters. It is therefore important to introduce a dataset of posters as the groundwork for saliency models of posters.

Compared to natural scenes, posters are rich in semantics which requires higher level of cognition. Bottom-up and top-down mechanism are the two cognitive mechanisms in human's visual system: bottom-up mechanism refers to the instinctive and automatic deployment of attention, while top-down mechanism is driven by subjective factors such as pre-knowledge and personal interest [10]. Many studies have looked into how to incorporate top-down semantic features into saliency models. High-level semantic features such as face, person, object, text etc. have been employed by researchers. Among them, face is found the most important semantic feature, followed by text and other pop-out elements [11].

It's important to note that posters are especially rich in texts (e.g. title, slogan, description paragraph), making it necessary to put special emphasis on this semantic feature. The various types of texts serve respective functions, making some relatively more important than others. However, current saliency models incorporate texts as feature for model training by using a text detector, which puts same weight on every piece of text and thus results in significant inaccuracy for saliency prediction of text regions [12]. In terms of why different texts attract differently, there are two levels of reasons: the lower-level reason is about typography, which leads to bottom-up saliency difference and is largely decided the type of text; the higher-level one concerns top-down understanding of words. In this study we focused on the former and tried to alleviate the latter. Analysis of eye-tracking results of our dataset with text as the focus aims to offer some insights into how text influences saliency.

## 2 Large-Scale Eye-Tracking

### 2.1 Dataset Gathering

Our dataset consists of 1700 posters from a wide range of fields such food, cosmetics, electronics, jewelry, auto etc. This was done to cover more design patterns and reduce the impact of personal preference or pre-knowledge of our participants. The posters were then divided into three categories based on the ratio of graph and text: The first contains almost no readable text (group A), the second only a piece of title/slogan (group B), and the rest are the third (group C). The number of images for each category is almost equal. In order to ease top-down semantic influence, we only included posters in Chinese for group B and group C, and posters which featured copywriting weren't chosen at all. After clarity examination, watermark editing and resizing, the three categories of pictures were mixed and evenly distributed into four groups.

## 2.2 Eye-Tracking Experiment

60 participants (29 males, 31 females) were students between the age of 18 and 25. They were divided into four groups, making each image viewed by 15 participants. All viewers sat at a distance of about 50 cm from a screen with resolution of 1280 \* 1024, and they were asked to stabilize their heads. Tobii T60 eye tracker was used to record eye movement and calibration was examined before each run. During the free-viewing task, each image was shown for three seconds.

## 3 Data Analysis

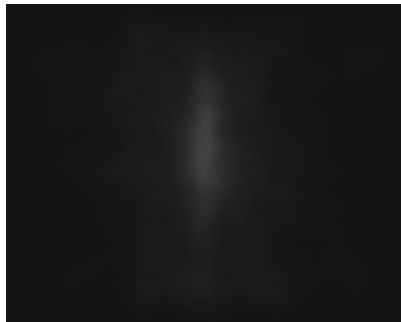
In order to statistically study the saliency pattern of our dataset, we chose one group of images from our dataset and used tools provided by Tobii Studio such as heatmap and area of interest to analyze the eye-tracking data. Further investigation into the differences in statistic results of the three categories of images reveals how various text elements interact with each other and how they impact other elements.

### 3.1 Heatmap

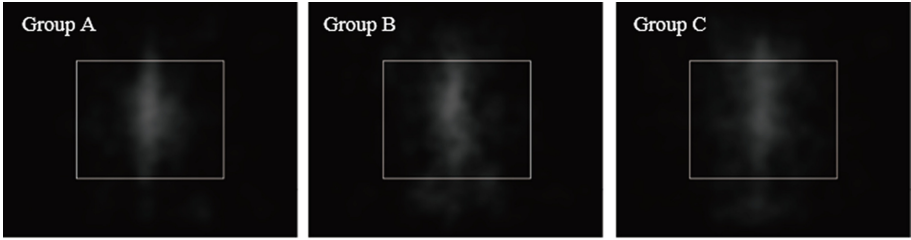
#### Center-Bias

The term center-bias refers to the phenomenon that most fixation points tend to be located in the center of an image during a free-viewing task. Center-bias has been widely proved in eye-tracking studies using datasets of natural scenes [5, 13]. It is often attributed to the preference of photographers, who are likely to put objects of interest in the center of the image. In turn, viewers would form the habit of searching for important content near the central area of an image [14]. Also, due to experiment settings the fact that viewers directly face the center of the screen could contribute [15].

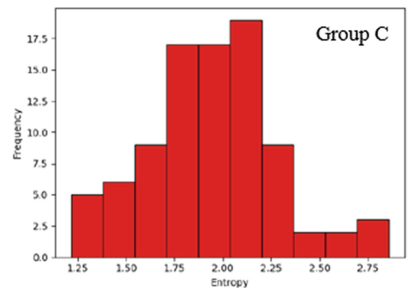
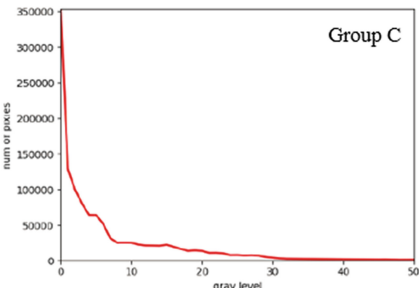
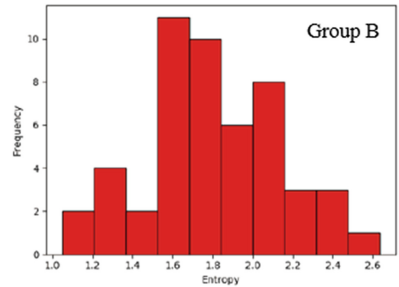
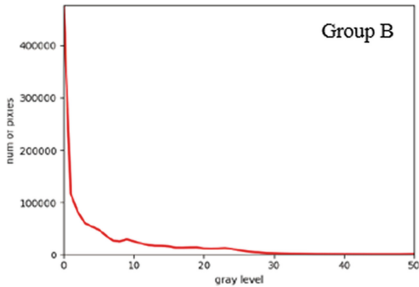
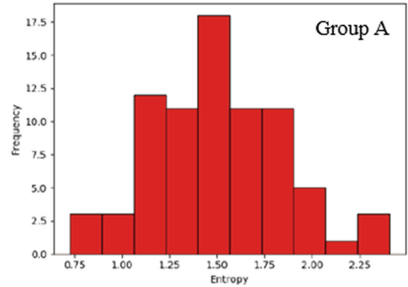
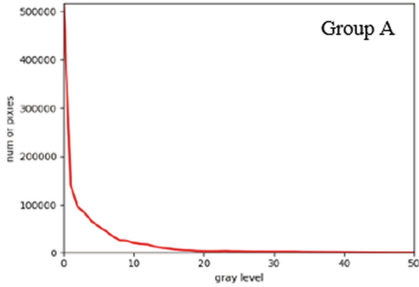
It is interesting to note that center-bias still applies for our dataset of posters. Figure 1 shows the mean heatmap of all posters within the group. The average heatmap has spindle-shaped clustering of fixation points in the central region, proving the center-bias hypothesis. However, the tendency towards the center isn't very strong compared with previous work on dataset of natural scenes [5]. This may be due to the fact that posters are naturally more informative and efficient compared to natural scenes. This means that more scattering elements, which serve specific functions, are intentionally designed to attract viewers' attention.



**Fig. 1.** Average heatmap of all posters within the group



(A)



(B)

(C)

**Fig. 2.** (A) Average heatmap of group A, group B and group C; (B) Gray level plot of group A, group B and group C; (C) Entropy map of group A, group B and group C

We then looked at the saliency difference between our three categories of stimuli. We found that fixation region is more concentrated for group A, and group C is the

least concentrated. Figure 2(A) shows the average saliency map for the three different categories. From our observation, Group A enjoys more clustered fixation areas which form a clear spindle-shaped fixation region near the center of the image. However, for group C the fixation area is rather vague. Statistically, for group A 41.1% of salient areas fall in the 25% central region, and for group C the number is 34.2%. According to the gray level plot, for group A there is a larger number of black pixels (more than 500 K) and smaller number of gray pixels, and the decrease slope is rather smooth (see Fig. 2(B)). In comparison, group C owns fewer black pixels (around 350 K) as well as more gray pixels, and its decrease slope fluctuates. The reason for the difference in concentration of salient area may lie in the fact that designers tend to place the most important and eye-catching element in the center when designing posters with only graph and logo (group A). Without other distracting elements, viewers are naturally drawn to the center. With the increase in dispersed text elements, viewers' fixation is more scattered.

In order to analyze whether human fixation consists for different people, we measured the entropy of saliency maps averaged across all viewers. Our Entropy histograms present gaussian-like distribution (see Fig. 2(C)). Group A exhibits lower level of entropy compared with group B, and group C has the highest entropy. For images with various text elements (group C), viewers' choice of where to look tends to differ based on subjective top-down factors such as personal interest and pre-knowledge. However, for images with fewer places to look at, viewers' decisions tend to be driven by bottom-up factors which basically act the same on everyone.

### Specific Salient Element

From observation of the saliency maps, we noticed that viewers tend to focus on specific elements such as face, object and text.

#### A. Face

Our saliency maps show a strong bias for people to focus on face, and the same applies for faces of animals, personified objects and even sculpture (see Fig. 3). Within the face, fixations are more likely to fall on eyes, nose and lips. When several faces appear, people tend to focus on faces in the middle and probably the more good-looking face.



Fig. 3. Examples of heatmap for images with face

B. Object

Objects can also catch a lot of attention (see Fig. 4). Objects placed in the center are more likely to be looked at than those which scatter around. It’s interesting to note that there seems tendency for fixations to fall on objects presented by a spokesperson.



Fig. 4. Examples of heatmap for images with object

C. Text

Texts are very common and very carefully designed in posters, as they could convey information both effectively and efficiently. Through observation of our saliency maps, it is clear that text areas always rank high on the fixation list (see Fig. 5). This could be attributed to the fact that we are almost instinctively driven to look at text and to try to understand it. However, not every piece of text enjoys the same degree of saliency because of the differences in text type, and this would be further investigated in the next part. Additionally, there are two interesting points to notice: text-on-object can often draw lot of attention, and both text saliency and object saliency may have contributed; name of the spokesperson can always attract attention. This could be explained by the location of the text which is normally near face of the person, making movement of attention easy. Also, the drive to know the name of the spokesperson is a possible reason.



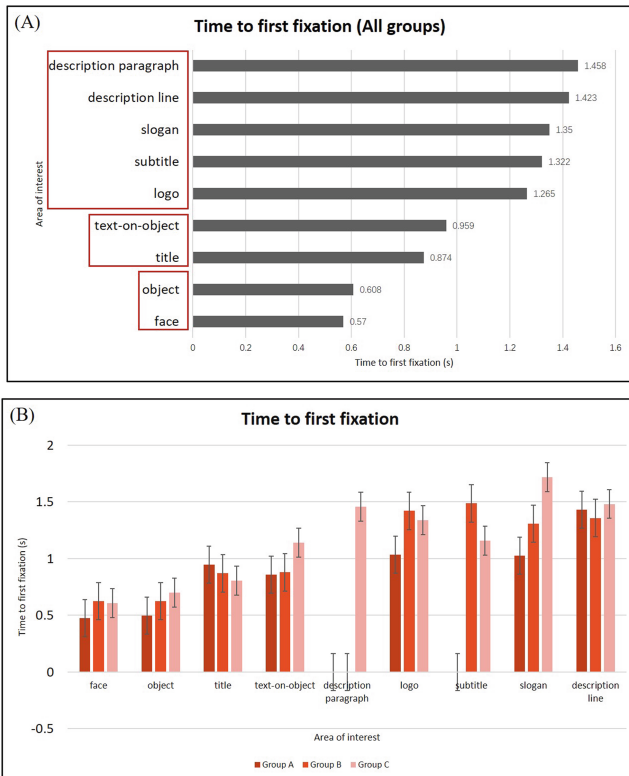
Fig. 5. Examples of heatmap for images with text-on-object

### 3.2 Area of Interest

In order to quantitatively study how people fixate on specific elements we hand-labeled our dataset. The elements chosen are face, object, logo and text. Text is further categorized into title, subtitle, description line, description paragraph, slogan and text-on-object so that we could study at length how different types of text influence overall visual saliency.

#### Time to First Fixation

Time to first fixation demonstrates the order in which viewers fixate on an area of interest. As in shown in Fig. 6(A), all the areas in concern could be grouped into three tiers: object and face first catch attention at around 0.6 s, then around 0.9 s title and text-on-object take the focus, after that comes all the text elements such as paragraph, logo, subtitle and slogan during 1.2 s and 1.4 s. This shows that object and face could catch viewers' attention straight away. Title and text-on-object come the next. Fixation of title could be explained by typography of the title: as the usually most important piece of text in a poster, a title is normally big in size, placed in the center of the image and with carefully chosen font. As for text-on-object, perhaps object saliency itself and the instinct drive to know what the object is by reading text-on-object could explain. Within the remaining



**Fig. 6.** (A) Average time to first fixation; (B) Time to first fixation of group A, group B, and group C

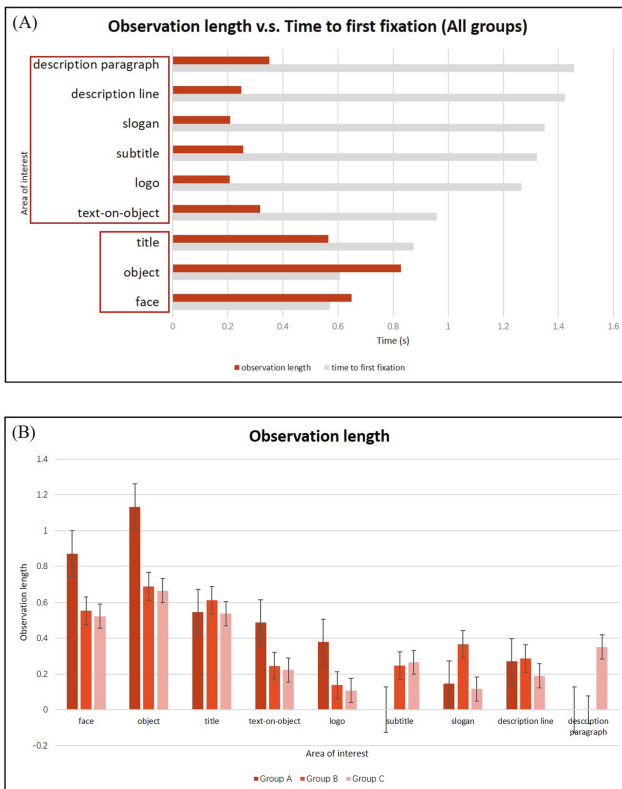


text elements, description paragraph and description line seem the least attractive for viewers, as they often are not very interesting. On the contrary, a logo, though not big in size and very often placed in the corner of the poster, is always emphasized through color, font etc.

Further analysis of average time to first fixation of three different groups reveals the differences between the groups (see Fig. 6(B)). With the increase in text elements, there is tendency of delayed time to first fixation for object, face, text-on-object, logo and slogan. The reason may lie in the fact that as the amount of text elements increases people spend longer time deciding where to look at. For title and subtitle, however, time to first fixation seems to be brought forward, as title and subtitle might be more intentionally designed to catch attention among all the elements in the poster.

### Observation Length

Observation length measures the degree to which an area attracts. It stands for the total time a person looks within an area of interest, starting from a fixation within the area and ending with a fixation out of the area. Figure 7(A) shows the average observation length of posters. Object, face and title are among the top three areas which could hold attention for a fairly long period of time. Observation length of the remaining text elements are



**Fig. 7.** (A) Average observation length; (B) Average observation length of group A, group B, and group C

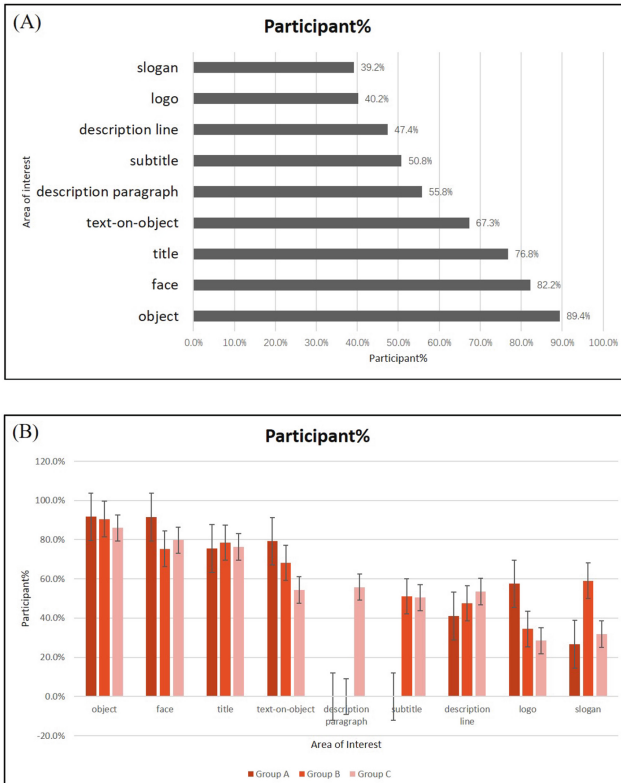
mostly the same, with description paragraph a bit longer than others. It is noticeable that the order of fixation and the length of fixation go hand in hand, which means that the area which attracts attention first is more likely to hold attention longer. Object, text-on-object and description paragraph, however, are the exceptions. For object, the rather long observation length could be because that viewers are actually spending time on text-on-object, which is hard to separate; For text-on-object, although it catches attention very fast it might not hold it very long because the words may be easily understood; For description paragraph, the larger amount of message contained requires more time of understanding.

Figure 7(B) demonstrates the comparison between observation length of posters of different groups. It is clear that observation length of face, object, text-on-object and logo is much longer for group A than that for group B and group C. This is because that with more distracting text elements in Group B and Group C, viewers tend to search for more information instead of dwelling on the first few interest areas.

**Participant%**

Participant% refers to the ratio of participants who fixated on a specific area of interest. It has strong impacts on the intensity of an area on a saliency map.

As is shown in Fig. 8(A), among all the text-related areas, title ranks first (89.4%), next comes text-on-object (67.3%), followed by description paragraph (55.8%), subtitle



**Fig. 8.** (A) Average participant%; (B) Average participant% of group A, group B, and group C

(50.8%) and description line (47.4%), and finally logo (40.2%) and slogan (39.2%). This reveals the relative importance of different text elements, and could help determine weights of those elements in studies of saliency prediction neural networks. Comparison of participant% between three groups (see Fig. 8(B)) shows that less people fixate on text-on-object and logo when the number of text elements increases.

## 4 Discussion

Our eye-tracking dataset of 1700 posters with category annotations based on the ratio of graph and text allowed us to quantitatively inspect the overall saliency pattern of posters with focus on text. Subsequent between-group studies helped us gain knowledge of how texts interact with each other and other elements in posters.

We learned that center-bias still holds true for posters, though not so strong and rather spindle-shaped. With the increase in text elements, fixations are more dispersed and people's choices of where to look at tend to differ. Observation of heatmap shows that face, object and text are the specific elements people focus on. Also, different types of text are found to attract differently. All the elements mentioned above were labeled as area of interest, with text further categorized into title, subtitle, description line, description paragraph, slogan and text-on-object. Analysis of time to fixation shows that object and face are the first elements that draw attention, followed by title and text-on-object. However, as the amount of text elements increases people tend to take longer time to fixate on object and face, as they need more time to decide where to look at. Observation length reveals the degree to which an area attracts, and the result mostly seems to go hand in hand with the order of fixation, with face, object and title remaining the top 3. There is a strong tendency for people to spend less time on face and object when more text elements exist. The ratio of participants who fixated on a specific region (participant%) distributes almost evenly from object (89.4%) to slogan (39.2%). Less people focus on text-on-object and logo when there are more text elements. The relative importance of different types of text could offer some advice on the weights for different text features in saliency model studies.

## 5 Conclusion

In this paper we made the following contributions: We provided a dataset containing 1700 high-quality posters falling into a wide range of areas, with eye-tracking data collected from 15 people for each image. To our knowledge there aren't yet such datasets for posters. Our dataset could thus be the foundation for research on visual saliency model for posters. Statistical analysis of heatmap and area of interest reveals the overall saliency pattern of posters. Specifically, the relative importance of different text regions and how they influence saliency of other elements. This could provide some advice on saliency prediction of posters.

For future work it is advisable to work on saliency prediction model for posters, as it is a relatively novel field and has strong practical value. To take a step further, research could be done on automatic evaluation system for posters.

## References

1. Hutton, S.B., Nolte, S.: The effect of gaze cues on attention to print advertisements. *Appl. Cogn. Psychol.* **25**(6), 887–892 (2011)
2. Xu, P., Ehinger, K.A., Zhang, Y.: *TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking*. Computer Science (2015)
3. Kim, N.W., et al.: *BubbleView: an interface for crowdsourcing image importance maps and tracking visual attention*. *ACM Trans. Comput. Hum. Interact.* **24**(5), 1–40 (2017)
4. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations (2012)
5. Judd, T., Ehinger, K., Durand, F.: Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2106–2113. IEEE (2009)
6. Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., Chua, T.-S.: An eye fixation database for saliency detection in images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part IV*. LNCS, vol. 6314, pp. 30–43. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-15561-1\\_3](https://doi.org/10.1007/978-3-642-15561-1_3)
7. Murray, N., Marchesotti, L., Perronnin, F.: AVA: a large-scale database for aesthetic visual analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2408–2415. IEEE (2012)
8. Borji, A., Itti, L.: Cat2000: A large scale fixation dataset for boosting saliency research. arXiv preprint [arXiv:1505.03581](https://arxiv.org/abs/1505.03581) (2015)
9. Shen, C., Zhao, Q.: Webpage saliency. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *ECCV 2014, Part VII*. LNCS, vol. 8695, pp. 33–46. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-10584-0\\_3](https://doi.org/10.1007/978-3-319-10584-0_3)
10. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundations: a survey. *ACM Trans. Appl. Percept. (TAP)* **7**(1), 1–39 (2010)
11. Kümmerer, M., Theis, L., Bethge, M.: Deep gaze I: boosting saliency prediction with feature maps trained on imagenet. arXiv preprint [arXiv:1411.1045](https://arxiv.org/abs/1411.1045) (2014)
12. Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., Durand, F.: Where Should Saliency Models Look Next? In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016, Part V*. LNCS, vol. 9909, pp. 809–824. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46454-1\\_49](https://doi.org/10.1007/978-3-319-46454-1_49)
13. Kruthiventi, S.S., Ayush, K., Babu, R.V.: Deepfix: a fully convolutional neural network for predicting human eye fixations. *IEEE Trans. Image Process.* **26**(9), 4446–4456 (2017)
14. Tseng, P.H., Carmi, R., Cameron, I.G., Munoz, D.P., Itti, L.: Quantifying center bias of observers in free viewing of dynamic natural scenes. *J. Vis.* **9**(7), 4–4 (2009)
15. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.: WSUN: a Bayesian framework for saliency using natural statistics. *J. Vis.* **8**(7), 32–32 (2008)



# Improving the Web Accessibility of a University Library for People with Visual Disabilities Through a Mixed Evaluation Approach

Milda Galkute<sup>1</sup>, Luis A. Rojas P.<sup>2</sup>(✉), and Victor A. Sagal M.<sup>3</sup>

<sup>1</sup> Pontificia Universidad Católica de Chile, Santiago, Chile  
mgalkute@uc.cl

<sup>2</sup> Universidad Central de Chile, Santiago, Chile  
lrojas.larp@gmail.com

<sup>3</sup> Universidad Andres Bello, Santiago, Chile

**Abstract.** Web accessibility is an essential aspect in nowadays technologically ever-advancing societies as it promotes equality and inclusiveness of those people with different kind of disabilities, such as visual impairment. This article examined the accessibility of one Chilean University's Virtual Library platform in order to generate significant improvements and to make it more accessible for visually impaired students. To do so, the present research used a mixed method approach, which included heuristic usability evaluation, user experience evaluation through the usability test involving real users, and both automatic and manual evaluations of accessibility. Results revealed that the most remarkable accessibility issues were related to the general structure of the website and the user control of the moving elements. Based on the information gathered through the different evaluations, it was possible to outline a proposal for changes to the University's Virtual Library platform. The proposed changes were subsequently implemented, resulting in the platform's compliance with the A and AA levels.

**Keywords:** Accessibility · Heuristic usability evaluation · Methodology · Visual disabilities

## 1 Introduction

The importance of Web accessibility lies in the fact that it allows the social inclusion of people with disabilities, favoring participation and interaction in access to knowledge. Currently, there are several studies that describe how to promote Web accessibility, especially in different university libraries. Although this research topic is not new, it is worth drawing attention to the evaluation techniques used in different researches. Some scholars have relied exclusively on heuristic evaluation [1–3], while usability testing through experiments have been implemented in others [4, 5]. Likewise, an automated evaluation instrument has also been used by several researchers as the main technique to check library websites and to identify their accessibility issues [6, 7].

What has been missing in this research area so far is an all-inclusive methodology that would combine heuristic evaluation, usability testing involving real users, as well as automated and manual evaluation, giving a more complete and holistic view of Web accessibility-related issues.

Therefore, this paper aims to redesign the current digital platform of one Chilean University library, by applying all the aforementioned methods, in order to make it more accessible for students with visual impairment, so that they can access electronic material of the digital library more easily.

The following evaluations have been carried out: heuristic usability evaluation involving experts in the area; evaluation of user experience through the usability test involving real users; and evaluation of Web accessibility, both automatic and manual.

Based on the results of different evaluations performed at the University's Virtual Library platform, a set of improvements were proposed, implemented and validated. The evaluations have provided comprehensive and relevant information to analyze and to decide the breadth and depth of the set of improvement proposals. Likewise, these proposals present their traceability with regard to the evaluations they were identified. As a result, it was possible to determine the efficiency of different evaluations and how they complement each other in the analysis of the needs and acceptance criteria of the end users. Additionally, formal tests were carried out with visually impaired students in order to validate the implementation of the elements proposed in the improvements. The User Experience Questionnaire (UEQ) [8] was applied in both formal tests, which were compared using the UEQ Compare Products tool [9]. These results show a general improvement in all evaluation scales, validating the success of the proposal.

## 2 Related Work

Indeed, heuristic evaluations which focus on the new interface assessment conducted by experts (see [10]) are likely to overcome a number of disadvantages faced by other methods. For instance, heuristic evaluations tend to be less time-consuming and less costly, providing efficient guidance to assess and to enhance systems [1, 11]. Among the recent studies that applied heuristic evaluation, it is worth mentioning Inostroza et al. [12] who identified a total of 53 usability-related issues in an experiment with touchscreen mobile devices using heuristic evaluation.

Likewise, Rusu and others [13] argue that “there is no evidence that a formal process or methodology had been used in order to establish sets of heuristics” (p. 51), proposing their own stages to develop usability heuristics. Similarly, a recent research conducted by Fung et al. [1] who evaluated the University of Hong Kong Library mobile website using the ten usability heuristic proposed by Nielsen [10], resulted in only 5 usability heuristics with adequate performance. Thus, a great number of international studies in heuristic evaluation have shown that the findings in this research area are still few and scattered, which was also confirmed by a systematic literature review on usability heuristics for mobile phones conducted by Salazar et al. [11], calling in this way for the need for more investigation.

As for the experiments-based usability testing, a recent qualitative study of Mulliken and Falloon [4] discovered several limitations in accessibility of blind screen readers to

academic library in the USA, including locating articles through links or accessing printed library books. Another recent study conducted by Arzola [14] that involved interviews with students with visual impairment revealed a serious concern related to documents' accessibility due to the lack of implementation of accessibility requirements in higher education in the USA. Indeed, this kind of studies that involve a considerable number of in-depth interviews are likely to contribute to the research area by showing first-hand experiences and challenges encountered by the population under study.

When it comes to the automated evaluation as the main tool to identify the accessibility issues of library websites, Nir and Rimmerman [6] used the WAVE instrument to check different Israeli university web pages and to identify accessibility issues, revealing very low accessibility levels. These findings go in line with the findings of previous studies that used automated evaluation as the main instrument, including [15] and [16]. Nevertheless, as mentioned by Wijayaratne and Singh [16], the automated evaluation tool is not able to examine every single issue in the guidelines.

Therefore, there have been several previous international studies that assessed both, usability and accessibility of different Web pages particularly for visually impaired people using diverse methodologies, such as heuristic evaluation, usability testing through experiments and automated evaluation instruments. Nonetheless, to our knowledge, there are no previous studies in this research field that would combine all three methodologies in order to overcome the drawbacks of each methodology separately and to obtain a more comprehensive view of Web accessibility-related issues.

### 3 Experimentation

In the present research, several usability and web accessibility evaluations were carried out with the Virtual Library platform of the University of Andrés Bello (UoAB) in Chile, with the aim of evaluating the user experience as a whole (Fig. 1).



Fig. 1. Virtual Library Platform of the UoAB.

#### 3.1 Usability Test

The experimentation with users was performed with 5 higher education students with visual impairment, belonging to either UoAB or University of the Americas (UoA). The

experimental sessions were carried out in the premises of both universities using the personal equipment of each user.

**Participants.** All participants were visually impaired (totally blind). They performed all tests with the help of assistive technologies (screen reader software). It is worth mentioning that none of them had previously used the UoAB Virtual Library platform.

Table 1 shows the characteristics of each of the collaborators, including gender, age, screen reader and the degree of web expertise.

**Table 1.** Data of the usability test participants.

| Participants | Gender | Age | Screen reader | Web expertise |
|--------------|--------|-----|---------------|---------------|
| P#1          | Male   | 26  | Jaws          | High          |
| P#2          | Male   | 30  | Jaws y NVDA   | High          |
| P#3          | Female | 26  | Jaws          | Medium        |
| P#4          | Female | 24  | Voice over    | High          |
| P#5          | Female | 25  | Jaws          | Medium        |

This data was obtained through the interviews prior to the usability test as well as during the test.

The following techniques were selected to carry out the usability test: Retrospective test, Thinking Aloud, and the application of the UEQ user experience questionnaire. The aim of the above-mentioned techniques was to gather more information about the interaction with the platform and users' experience.

The tasks that the users had to perform were composed of two cases. The first case corresponded to a free search. The aim of this case was to search for a particular e-book on the website, employing the resources that the students used frequently, and to interact with the e-book. The second case consisted of a library catalogue search. In this opportunity the same resource had to be found, however, using the UoAB Virtual Library platform.

**Case 1: Free Search.** The tasks proposed for the first free search case are described hereafter:

1. Open the web browser used frequently.
2. Search for a specific book. Book title: Organic Chemistry. Author: John McMurry.
3. Open the resource and go to chapter 3: Organic compounds: alkanes and its stereochemistry.

**Results of the Case 1.** Table 2 and Table 3 present the results of the tasks performed by the participants in case 1.



**Table 2.** Results of the completed tasks of case 1.

| Participants | Completed tasks |     |     |     |                 |            |
|--------------|-----------------|-----|-----|-----|-----------------|------------|
|              | T#1             | T#2 | T#3 | T#4 | Number of tasks | % of tasks |
| P#1          | ✓               | ✓   | ✓   | ×   | 2               | 66.66%     |
| P#2          | ✓               | ✓   | ✓   | ✓   | 3               | 100%       |
| P#3          | ✓               | ×   | ✓   | ×   | 1               | 33.33%     |
| P#4          | ✓               | ✓   | ✓   | ✓   | 3               | 100%       |
| P#5          | ✓               | ✓   | ✓   | ×   | 2               | 66.66%     |

**Table 3.** Results of the times used in the tasks of case 1.

| Participants                   | Time spent on tasks (Seconds) |     |      | Total time (Seconds) |
|--------------------------------|-------------------------------|-----|------|----------------------|
|                                | T#1                           | T#2 | T#3  |                      |
| P#1                            | 6                             | 130 | 1000 | 1136                 |
| P#2                            | 2                             | 309 | 2292 | 2603                 |
| P#3                            | 4                             | 892 | 0    | 896                  |
| P#4                            | 5                             | 260 | 1243 | 1508                 |
| P#5                            | 8                             | 511 | 765  | 1284                 |
| Average time to complete tasks | 5                             | 420 | 1060 |                      |

**Case 2: Library Catalogue Search.** The tasks proposed for the second case of free search in library catalogue are described below:

1. Open the web browser used frequently.
2. Enter the Virtual Library page <http://biblioteca.unab.cl/>
3. Search for a specific book. Book title: Organic Chemistry. Author: John McMurry.
4. Open the resource and go to chapter 3: Organic compounds: alkanes and its stereochemistry.

**Results of the Case 2.** Table 4 and Table 5 present the results of the tasks performed by the participants in case 2.

Moreover, during the interaction carried out, valuable information could be collected by the researchers. This information was grouped into two categories: on the one hand, strategies used by the participants and, on the other hand, problems identified during the development of the tasks. Both categories are specified below.

**Table 4.** Results of the completed tasks of case 2.

| Participants | Completed tasks |     |     |     |                 |            |
|--------------|-----------------|-----|-----|-----|-----------------|------------|
|              | T#1             | T#2 | T#3 | T#4 | Number of tasks | % of tasks |
| P#1          | ✓               | ✓   | ✓   | ×   | 3               | 75%        |
| P#2          | ✓               | ✓   | ✓   | ×   | 3               | 75%        |
| P#3          | ✓               | ✓   | ✓   | ×   | 3               | 75%        |
| P#4          | ✓               | ✓   | ✓   | ×   | 3               | 75%        |
| P#5          | ✓               | ✓   | ✓   | ×   | 3               | 75%        |

**Table 5.** Results of the times used in the tasks of case 2.

| Participants                   | Time spent on tasks (Seconds) |     |     |     | Total time (Seconds) |
|--------------------------------|-------------------------------|-----|-----|-----|----------------------|
|                                | T#1                           | T#2 | T#3 | T#4 |                      |
| P#1                            | 4                             | 102 | 709 | 240 | 1055                 |
| P#2                            | 2                             | 132 | 423 | 612 | 1169                 |
| P#3                            | 7                             | 23  | 494 | 282 | 806                  |
| P#4                            | 6                             | 183 | 619 | 480 | 1288                 |
| P#5                            | 5                             | 254 | 388 | 219 | 866                  |
| Average time to complete tasks | 5                             | 139 | 527 | 367 |                      |

**Strategies:**

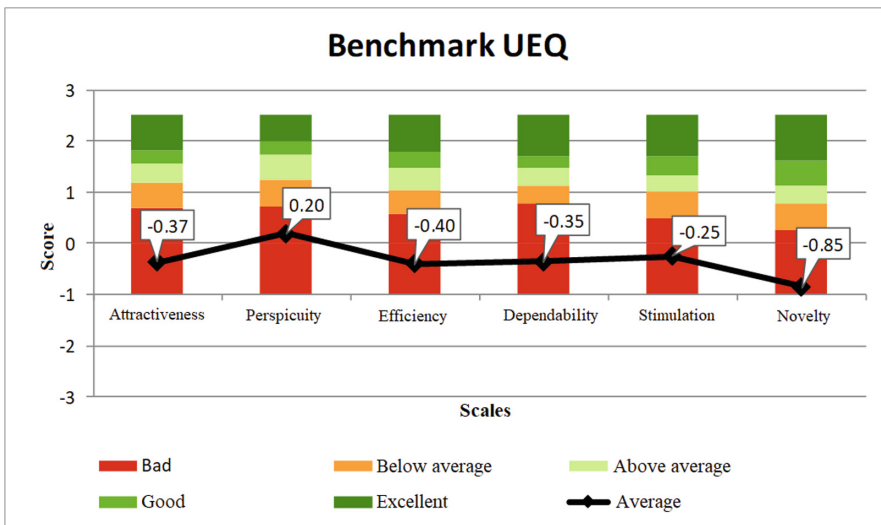
- Links and files were sent to email as a method of downloading the resources or opening them via email.
- Navigation was performed using keyboard commands in all system options (web browser, PDF viewer). No touchpad or mouse were used.
- The most frequently used browser on Windows was Internet Explorer, followed by Mozilla Firefox and Google Chrome. Students indicated that Internet Explorer had presented more accessibility options for them.

**Problems:**

- There was a large amount of PDF in non-accessible formats, which were similar to an image where the text was not recognizable, making it difficult for screen readers to read it.
- The images and links in the UoAB Virtual Library lack a description that could be interpreted by screen readers, making it difficult to access these resources. In these cases, the help of a third party was needed.

- The format of the books, downloaded from the library platform, was not a standard one, such as PDF. Therefore, a special program called Adobe Digital Editions was needed in order to view it.
- The advertising on the pages was not correctly interpreted by the screen readers, making navigation more challenging and problematic.

**UEQ Questionnaire.** Once the experimentation with the users was completed, they were asked to fill in the UEQ questionnaire [8]. The results of the questionnaire are presented in Fig. 2. As it can be seen, the scores were poor in all scales with regard to other products evaluated with the same questionnaire. It is important to notice that a negative evaluation corresponded to a value lower than  $-0.8$  [17]. The novelty scale had a value of  $-0.85$  and thus it was the only one to obtain such a negative value.



**Fig. 2.** Benchmark results of the UEQ questionnaire.

The benchmark or the UEQ questionnaire reference index classifies a software within 5 categories or scales [17]:

- Excellent: In the range of the top 10% of best results.
- Good: 10% of the results in the benchmark data set are better while 75% are worse.
- Above average or average: the 25% of the results in the benchmark data set are better than the results for the evaluated product, while the other 50% of the results are worse.
- Below average or average: the 50% of the results in the benchmark data set are better than the results for the evaluated product, while the other 25% of the results are worse.
- Bad: In the range of 25% of worst results.

The measurement scales were established in relation to the existing values of a reference data set. This allowed researchers to draw conclusions on the relative quality of the evaluated product, compared to other products [9].

The negative evaluation obtained through the UEQ user experience questionnaire is in line with the accessibility problems and barriers identified above. For example, there was a lack or absence of descriptions of the different elements that composed the Virtual Library platform, making it difficult for screen readers to interpret them; the navigation flow to be followed was not intuitive or explicit; in most cases, access to resources required authentication by means of a user and password.

### 3.2 Heuristic Evaluation

Heuristic evaluation is a screening method that helps identify usability problems based on usability principles. An expert evaluator judges the application interface by listing the usability problems that were found. These problems are categorized and evaluated, and a report is then generated with the analysis of the results and possible suggestions for the system developers [18].

In order to carry out the usability evaluation, in the present study researchers applied specific heuristics for library web services proposed by [19]. On the other hand, to measure the severity of the identified usability problems, the following factors were considered [20]:

- Frequency: The frequency of the problem and whether it is common or not.
- Impact: Whether users tend to have many difficulties due to the occurrence of the problem.
- Persistence: Whether the problem appears repeatedly or not.

Finally, a score was established for each of the findings or problems as follows [20]:

- 0 = Not a usability problem.
- 1 = A minor problem: Does not need to be corrected unless there is enough time to do so.
- 2 = A small problem: Fixing the problem is not relevant.
- 3 = A serious problem: It is important to fix it.
- 4 = A disastrous problem: It is mandatory or essential to fix it.

**Results.** The usability problems that have been identified are presented below, according to their degree of severity and impact on the system. It should be noted that the only results included were those from levels 3 and 4, since these impact on making changes to the system:

As presented in Table 6, the most serious problems that needed to be addressed were related to the following aspects: the contrast of background and text colors; the absence of <alt> labels for graphics and other elements so that they could be read by the screen readers; the total absence of prevention and error messages in the platform.

**Table 6.** Usability problems detected.

| Problems   | Heuristic (#) | Severity |
|--|---------------|----------|
| Use of frames  | No.2          | 3        |
| Background color does not contrast clearly with the text colors          | No.4          | 4        |
| Use of embedded links  | No.5          | 4        |
| Links without description  | No.5          | 3        |
| Graphics as links  | No.5          | 3        |
| No information about links leading to files other than HTML is attached  | No.5          | 3        |
| Text is presented as a graphic   | No.6          | 3        |
| Graphics without text on the label < alt>                                | No.6          | 4        |
| No additional text is provided for media files                           | No.6          | 3        |
| Graphics with text without label < alt>                                  | No.6          | 4        |
| Not all buttons are recognizable   | No.6          | 3        |
| Text colors are different from black, very low contrast with backgrounds | No.6          | 4        |
| Strong colors used in excess   | No.6          | 4        |
| There are no error messages  | No.7          | 4        |
| The accessibility of the site cannot be checked                          | No. 8         | 3        |

### 3.3 Automatic Evaluation of Accessibility

The automatic evaluation was performed using the W3C Markup Validation Service and the Wave browser extension in Google Chrome.

The results of both tools indicated that there were certain elements that did not comply with the WCAG web accessibility guidelines, such as alternative text for images, color contrast, empty HTML headers, and the presence of frames.

### 3.4 Manual Evaluation of Accessibility

For a more detailed analysis, a manual evaluation of accessibility was performed. This evaluation included three stages: the initial accessibility check, the execution of the WCAG-EM methodology, and the generation of the report using the W3C template.

**Initial Accessibility Check.** In the initial check, the search for accessibility problems in the page code was done manually and was then complemented with the use of automatic evaluation tools, which yielded the following findings:

- Absence of alternative text for all images.
- Headings: Empty headings were found.
- Contrast ratio: Some did not follow a contrast ratio of at least 4.5:1 for normal-sized text.

**Execution of the WCAG-EM Methodology.** Once the initial check was completed, the WCAG-EM methodology was applied, with the following steps:

5. Scope of the assessment: The web accessibility of the UoAB Virtual Library page was assessed. The objective was to identify its current status of compliance with the WCAG guidelines at the AA compliance level.
6. Exploring the website: Main page and sub-pages, where the most important functionality to be evaluated was the search for electronic resources and the requirements to be able to carry it out.
7. Selection of the representative sample: Main page and sub-pages related to the search for electronic and physical library resources.
8. Assessment of the selected sample: Both successes and failures in the WCAG guidelines were determined.
9. Reporting the Assessment Results: The accessibility report was prepared using the report template recommended by W3C (See Table 7).

**Table 7.** Results of the accessibility report.

| Principle      | Guideline                            | Compliance                | Level |
|----------------|--------------------------------------|---------------------------|-------|
| 1: Perceivable | 1.1 Text alternatives                | 1.1.1 Non-textual content | A     |
|                | 1.4 Distinguishable                  | 1.4.1 Use of color        | A     |
|                |                                      | 1.4.3 Contrast (minimum)  | AA    |
|                |                                      | 1.4.5 Text images         | AA    |
| 2: Operative   | Guideline 2.1 Accessible by keyboard | 2.4.3 Order of the focus  | A     |
|                |                                      | 2.4.6 Headings and labels | AA    |
|                |                                      | 2.4.7 Visible focus       | AA    |

Finally, and by way of summary, the following identified elements can be mentioned:

- The UoAB Virtual Library site did not meet the conditions for the AA compliance level, nor did it meet the conditions for the A compliance level.
- The most remarkable accessibility features were the general structure of the website and the user control of the moving elements.
- The most important measures to be implemented were the following: alternative text labels for all images; contrast correction of site elements; removal of empty headings; removal or correction of frames.

## 4 Proposal

Based on all the information gathered through the different evaluations that were carried out in this study, it was possible to outline a proposal for changes to the UoAB Virtual Library platform. Table 8 indicates the traceability between improvement proposals and the type of evaluation that was identified.

**Table 8.** Traceability between improvement proposals and the type of evaluation that was identified.

| Proposal   | Type of evaluation |                      |                                    |                                       |
|--|--------------------|----------------------|------------------------------------|---------------------------------------|
|  | Usability test     | Heuristic evaluation | Manual evaluation of accessibility | Automatic evaluation of accessibility |
| Adding the alt attribute with a comprehensible and logical description for all the images on the Virtual Library website.                        | ✓                  | ✓                    | ✓                                  | ✓                                     |
| Changing the color contrast for all elements that did not comply with the 4.5:1 ratio recommended by the WCAG web accessibility guidelines [21]. |                    | ✓                    | ✓                                  | ✓                                     |
| Removing empty headings in the website structure.  |                    |                      | ✓                                  | ✓                                     |
| Correcting the frames by adding a title attribute so that the screen readers could identify it as an element.                                    |                    | ✓                    | ✓                                  | ✓                                     |

As shown in Table 8, both accessibility evaluations (manual and automatic) allowed to identify all improvement proposals. Manual evaluation requires more time for its implementation, but it provides valuable insights to get a more comprehensive view. On the other hand, even though most improvement proposals were not identified in the usability test, this evaluation allowed to discover participants' interaction styles with the platform. The latter was valuable information to identify improvement opportunities.

## 5 Experimentation of the Proposal Implementations

The proposals identified above were implemented on the UoAB Virtual Library platform. Subsequently, several evaluations have been conducted to assess whether user-perceived improvements had been achieved.

**Evaluation Method.** In order to evaluate the new version of the UoAB Virtual Library platform, a usability test was carried out. In this evaluation, users had to perform a number of tasks so as to interact with the platform. Afterwards, the UEQ user experience questionnaire was applied to them, with the aim of comparing the results obtained before and after the implementation of the changes.

### 5.1 Usability Test

**Participants.** Five higher education students (belonging to UoAB and UoA) with visual impairment (total blindness) participated in the user test. In this way, the same profile of the first evaluation described above has been complied with.

**Tasks.** The tasks associated with this evaluation correspond to those applied in the previous evaluation (see Sect. 3.1).

**Procedure.** The XAMPP software was installed on the participants’ personal computers. Afterwards, the Virtual Library platform was executed. The tasks described in the previous point were then carried out. Finally, the UEQ user experience questionnaire was applied to every participant.

**Results.** As can be seen in Fig. 3, unlike the evaluation conducted prior to the implementation of changes (see Sect. 3.1), none of the scales had a negative evaluation, meaning that no lower value than  $-0.8$  was obtained.

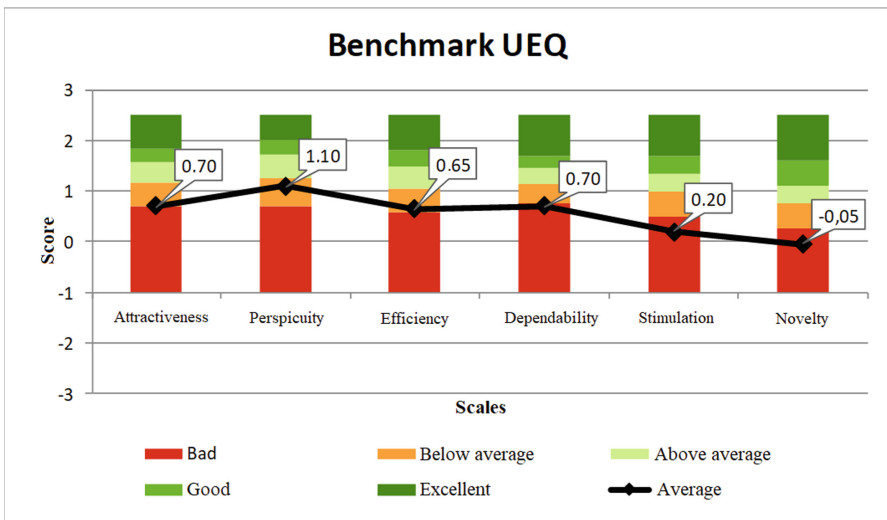


Fig. 3. Benchmark results of the UEQ questionnaire.



While there were four scales that were still in the poor category, two other scales have changed significantly. For instance, the perspicuity was above average, and the efficiency was below average. This shows that the modifications made to the site had an impact on the user experience of the participants.

On the other hand, using the User Experience Questionnaire tool [9], the comparison between both user experience evaluations (pre and post changes in the Virtual Library) was obtained. The results are presented in Fig. 4. In this way, Data Set 1 is the evaluation made before the changes while Data Set 2 is the evaluation made after the changes. As can be observed, there was a general improvement in all scales from one evaluation to another on the UoAB Virtual Library site, which also validates the success of the proposal.

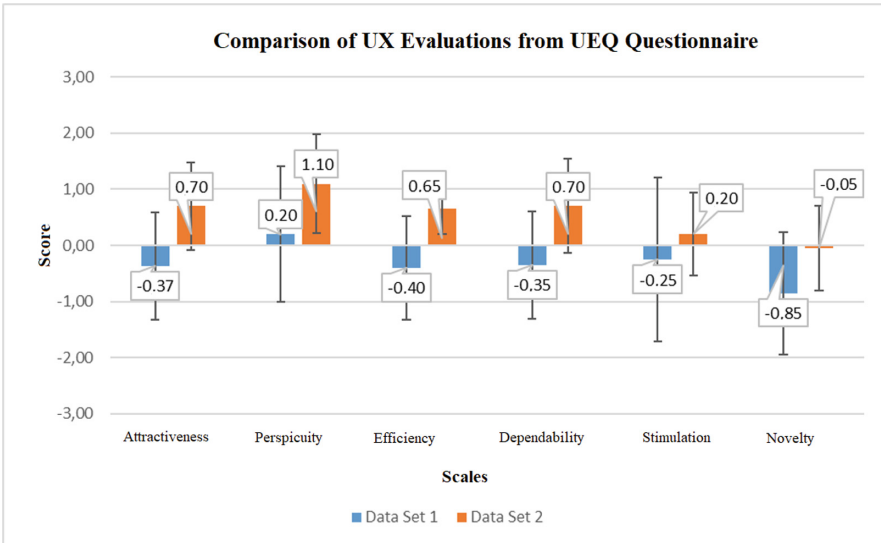


Fig. 4. Comparison of UX evaluations from UEQ questionnaire

### 5.2 Automatic Evaluation of Accessibility

The automatic accessibility evaluation could not be carried out via the W3C’s Markup Validation Service because the changes were not uploaded to the Library System server. Instead, it was attempted to upload the .PHP files. However, the latter were not supported by the evaluation tool as the validator exposed in it could only analyse HMTL or XHTML files. Therefore, the manual evaluation of the web accessibility of the Virtual Library platform was carried out.

### 5.3 Manual Evaluation of Accessibility

In this second evaluation, the WCAG-EM methodology was applied, using exactly the same steps and objectives as described in Sect. 3.4. In this way, only those WCAG

guidelines that had not met the compliance criteria in the previous assessment were verified.

**Report Results.** The results of the evaluation indicate that the Virtual Library platform met the A and AA compliance levels.

The aspects that stand out in accessibility were the following: the correct overall structure of the website; user control over moving elements; the correct contrast ratio; all images had the alt attribute so that they could be interpreted by the screen readers.

On the other hand, inaccessible features of the site that still persisted despite the level of compliance achieved were the following: text images, such as computer graphics. Although a description was provided by means of alt attributes in the HTML code, more information would be specified if they were replaced by some more accessible formats such as PDF.

Both the user experience evaluation and the web accessibility evaluation show that the system has made significant progress, validating the proposal of this study. This is based on the results obtained in the UEQ questionnaire where the scores of all the scales have increased, as well as on the accessibility of the website which has reached the AA compliance level with the WCAG 2.0 guidelines.

## 6 Conclusions

Although a great number of international studies have been conducted so far on Web accessibility, they all tend to focus on one methodological perspective only. By contrast, the present study offered a novel and all-encompassing approach to Web accessibility by combining three different methods: heuristic evaluation, experiments-based usability testing and both automatic and manual evaluation.

Overall the evaluation results showed that the University's Virtual Library platform selected for this study did not meet the conditions for either the A or the AA compliance level at first instance. Likewise, the most remarkable accessibility features detected in this study were related to the general structure of the website and the user control of the moving elements. Moreover, it was revealed that alternative text labels for images, contrast correction of site elements, removal of empty headings and removal or correction of frames were the most important measures to be implemented in the platform.

Based on the results obtained through mixed method evaluations, improvement suggestions were made and implemented, resulting in a considerable enhancement of University's Virtual Library platform accessibility. As a result, the Virtual Library platform finally met the A and AA compliance levels.

Indeed, the present research has meaningful contributions to the academics and practitioners. Regarding the former, this is one of the first studies to combine three different methodologies in order to examine and to get an in-depth understanding of Web accessibility for visually impaired students. In this sense, this study provides an insight on how each evaluation method can overcome the drawbacks of other methods, complementing each other. As for the practitioners, this study provides valuable guidelines for website designers for identifying the main obstacles that visually impaired users face on a daily basis and thus creating more accessible and inclusive sites.

Nevertheless, this research also presents some limitations. First, it is important to note that this study was based on one particular University's website. A comparison with other Virtual Library platforms would be advantageous. Second, there were several obstacles to access to visually impaired students during the research, so the number of participants has not been as expected initially. However, these limitations have been considered for future research work. Therefore, it is expected to continue the research by expanding the number and type of websites, carrying out comparative evaluations. The researchers also hope to contact associations with people who suffer from different disabilities, in order to recruit a greater variety and number of participants.


## References

1. Fung, R.H.Y., Chiu, D.K., Ko, E.H., Ho, K.K., Lo, Y.P.: Heuristic usability evaluation of university of Hong Kong libraries' mobile website. *J. Acad. Librariansh.* **42**(5), 581–594 (2016)
2. Bertot, J.C., Snead, J.T., Jaeger, P.T., McClure, Y.C.R.: Functionality, usability, and accessibility. *Perform. Meas. Metr.* (2006)
3. Comeaux, D., Schmetzke, Y.A.: Web accessibility trends in university libraries and library schools. *Libr. Hi Tech* **25**(4), 457–477 (2007)
4. Mulliken, A., Falloon, Y.K.: Blind academic library users' experiences with obtaining full text and accessible full text of books and articles in the USA. *Libr. Hi Tech* (2018)
5. Allen, M.: A case study of the usability testing of the University of South Florida's virtual library interface design. *Online Inf. Rev.* (2002)
6. Nir, H.L., Rimmerman, Y.A.: Evaluation of Web content accessibility in an Israeli institution of higher education. *Univ. Access Inf. Soc.* **17**(3), 663–673 (2018)
7. Idrees, H., Mudassir, Y.K.: Library Web sites for people with disability: accessibility evaluation of library websites in Pakistan. *Libr. Hi Tech News* (2015)
8. Schrepp, M., Hinderks, A., Thomaschewski, Y.J.: Construction of a Benchmark for the User Experience Questionnaire (UEQ). *IJIMAI* **4**(4), 40–44 (2017)
9. Schrepp, M.: Data Analysis Tool (2018). <https://www.ueq-online.org>
10. Nielsen, J.: Finding usability problems through heuristic evaluation. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 373–380 (1992)
11. Salazar, L.H.A., Lacerda, T., Nunes, J.V., von Wangenheim, Y.C.G.: A systematic literature review on usability heuristics for mobile phones. *Int. J. Mob. Hum. Comput. Interact. IJMHCI* **5**(2), 50–61 (2013)
12. Inostroza, R., Rusu, C., Roncagliolo, S., Jimenez, C., Rusu, Y.V.: Usability heuristics for touchscreen-based mobile devices. In: *2012 Ninth International Conference on Information Technology-New Generations*, pp. 662–667 (2012)
13. Rusu, C., Muñoz, R., Roncagliolo, S., Rudloff, S., Rusu, V., Figueroa, Y.A.: Usability heuristics for virtual worlds. In: *Proceedings of the Third International Conference on Advances in Future Internet*, ser. pp. 16–19 (2011)
14. Arzola, R.: Collaboration between the library and office of student disability services. *Digit. Libr. Perspect.* (2016)
15. Ringlaben, R., Bray, M., Packard, Y.A.: Accessibility of American university special education departments' web sites. *Univ. Access Inf. Soc.* **13**(2), 249–254 (2014)
16. Wijayaratne, A., Singh, Y.D.: Is there space in cyberspace for distance learners with special needs in Asia? A review of the level of Web accessibility of institutional and library homepages of AAOU members. *Int. Inf. Libr. Rev.* **42**(1), 40–49 (2010)

17. Schrepp, M.: User Experience Questionnaire Handbook. All you need to know to apply the UEQ successfully in your projects. ResearchGate GmbH, Berlin (2015)
18. Quiñones, D., Rusu, Y.C.: How to develop usability heuristics: a systematic literature review. *Comput. Stand. Interf.* **53**, 89–122 (2017)
19. Aitta, M.-R., Kaleva, S., Kortelainen, Y.T.: Heuristic evaluation applied to library web services. *New Libr. World* (2008)
20. González, M.P., Pascual, A., Lorés, Y.J.: Evaluación heurística. In: 2001 Introd. Interacción Pers.-Ordenad. AIPO Asoc. Interacción Pers.-Ordenad. (2001)
21. Caldwell, B., Cooper, M., Reid, L.G., Vanderheiden, Y.G.: Web content accessibility guidelines (WCAG) 2.0. In: WWW Consort. W3C (2008)



# Utilization of Vanity to Promote Energy Saving Activities

Kyoko Ito<sup>1</sup>, Yasutaka Kishi<sup>2</sup>, and Shogo Nishida<sup>2</sup>

<sup>1</sup> Office of Management and Planning, Osaka University,  
1-1, Yamadaoka, Suita, Osaka 5650871, Japan  
[ito.kyoco@gmail.com](mailto:ito.kyoco@gmail.com)

<sup>2</sup> Graduate School of Engineering Sciences, Osaka University,  
1-3 Machikaneyama-cho, Toyonaka, Osaka 5608531, Japan

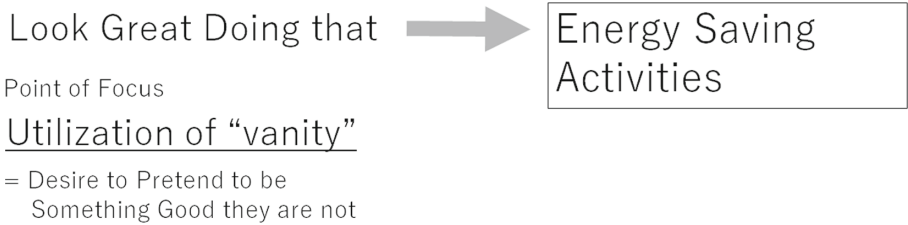
**Abstract.** Energy saving activities are human activities to reduce energy consumption. These activities can be an effective measure to reduce energy consumption and humans can start them immediately, so it is expected as a measure against resource depletion and increased carbon dioxide emissions. In this context, there have been attempts made from various viewpoints based on the knowledge that the improvement of understanding it would lead to putting it into practice. However, the results of a questionnaire survey on energy saving activities have indicated that there is a gap between understanding and in-practice regarding energy saving, and improvement of understanding does not lead to in-practice. From this point, it is expected not to improving the understanding, but to propose a method that gives direct motivation for energy saving activities. In this study, we focus on “vanity”, which is considered to be one of human needs for promoting energy saving activities. Here, “vanity” means a desire to pretend to be something good they are not. The purpose of this study is to consider experimentally whether energy saving activities could be applied to the target of vanity. At first, we will consider a method of applying vanity for energy saving activities and construct an experimental system to experimentally verify whether or not vanity could act as a motivation to promote energy saving activities. The experiment will be conducted using the experimental system to check if there are any persons who activate vanity regarding energy saving activities.

**Keywords:** Energy saving activities · Vanity · Action

## 1 Introduction

Energy saving activities are human activities to reduce energy consumption. These activities can be an effective measure to reduce energy consumption and humans can start them immediately, so it is expected as a measure against resource depletion and increased carbon dioxide emissions [1, 2]. In this context, there have been attempts made from various viewpoints based on the knowledge

that the improvement of understanding it would lead to putting it into practice [3,4]. However, the results of a questionnaire survey on energy saving activities have indicated that there is a gap between understanding and in-practice regarding energy saving, and improvement of understanding does not lead to in-practice. From this point, it is expected not to improving the understanding, but to propose a method that gives direct motivation for energy saving activities. In this study, we focus on “vanity”, which is considered to be one of human needs for promoting energy saving activities, as shown in Fig. 1. Here, “vanity” means a desire to pretend to be something good they are not.



**Fig. 1.** Our approach for promoting energy saving activities.

In recent years, there have been many studies that encourage people to act to support energy saving activities [5,6]. On the other hand, on social networking services (SNS) such as Facebook and Instagram, there are many posts including photos. Among them, there is a term “Instagrammable”, which is a photo that shows well on Instagram. The term means assuming the value by others, and aims at getting others to feel “good”. In addition, it shows the desire to pretend to be something good they are not. The desire to be recognized well by others is approval desire and a mindset of “vanity”.

The purpose of this study is to consider experimentally whether energy saving activities could be applied to the target of vanity. At first, we will consider a method of applying vanity for energy saving activities and construct an experimental system to experimentally verify whether or not vanity could act as a motivation to promote energy saving activities. The experiment will be conducted using the experimental system to check if there are any persons who activate vanity regarding energy saving activities.

Vanity originates from a human desire to be recognized by others and is an internal motive. If vanity could be used to promote energy saving activities, it is expected that it would lead to voluntary and sustainable energy saving activities. Figure 2 shows the concept for utilization of vanity on energy saving activities. If there is a person who utilize vanity for energy saving activities, it can be expected to promote energy saving activities by creating a place to activate vanity for energy saving activities. However, energy saving activities generally have different characteristics from the target of vanity. In other words, the targets of vanity are assumed to be what many people want, such as possessing goods that many people want to get, traveling to popular places, etc.

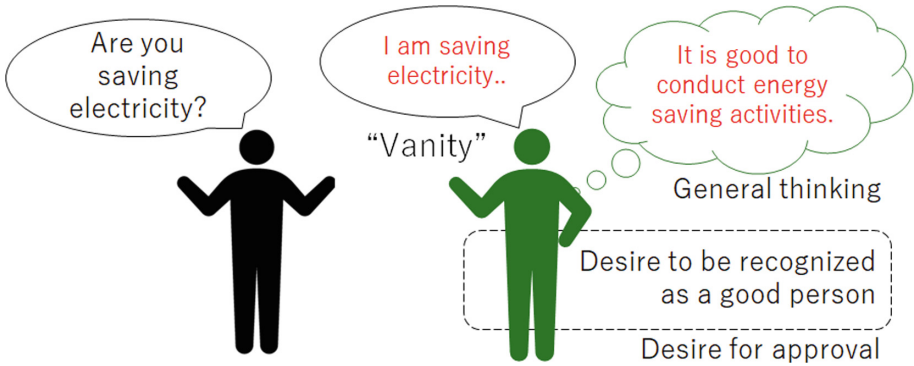


Fig. 2. Out concept for utilization of vanity on energy saving activities.

## 2 Related Studies

In recent years, there have been many studies on interfaces that encourage people to promote energy saving activities [5,6]. Gustafsson *et al.* have developed an extension cord that shines when power is used to give people awareness of power consumption [7]. Yagida *et al.* have investigated the impact of household electricity consumption by visualization on energy consumption in order to promote human understanding of energy [8]. Ueno *et al.* have focused on the fact that people have multiple ways of thinking about energy saving, classified people into multiple patterns, and proposed a method of supporting energy saving activities that matches the patterns [9]. In this way, there are some studies that promote understanding of energy and expect energy saving activities.

However, in order to put them into practice, it is difficult to motivate human energy saving behavior only because of the understanding of energy problem, that is, “to reduce energy consumption”. Therefore, studies has been conducted to motivate energy saving activities by using social influences such as the effects of the existence, attitude, and behavior of others. Petersen *et al.* have focused on competition among groups, and proposed a method of competing for low power consumption among groups and motivating energy saving activities [10]. Aoyagiet *al.* focused on the sympathy of groups and proposed a method of supporting energy saving activities, which reporting each other on energy saving activities would lead to synchronizing with other people, that is, “If everyone is doing energy saving activities, let’s do it [11]”.

In these methods of supporting energy saving activities using various social influences, Abrahamse *et al.* have proposed that a method of supporting energy saving activities by using the influence of the group would be more effective than supporting individuals [12]. In particular, they have pointed out that face-to-face interaction would have a high energy-saving effect [12]. However, face-to-face interaction is characterized by a low diffusion rate of influence and high human cost. Therefore, we use the situation that arises from one-to-many interactions by using the Web. In this study, we would like to promote the energy-saving

effect by using the social influence, that is, the mindset of “vanity” of a person in a situation where some people see it.

### 3 Method to Utilize “vanity” for Energy Saving Activities

When the targets of vanity are energy saving activities, the targets are classified into intentions and actions. As for the intentions, the vanity for the intension would lead to trying to show more intentions than the actual one. As for the actions, the vanity for the action would lead to trying to do more actions than the real quantity of actions.

In order to form a place where vanity would be shown on the Web, we examine a method of presenting intentions and actions on the Web. The targets of the vanity should be limited to energy saving activities when presenting the intentions and actions. In addition, in order to confirm the presentation of oneself and to expect the competition of vanity, it is necessary to make the presentation of intentions and actions by oneself and others visible. Figure 3 shows the outline of the utilization of vanity.

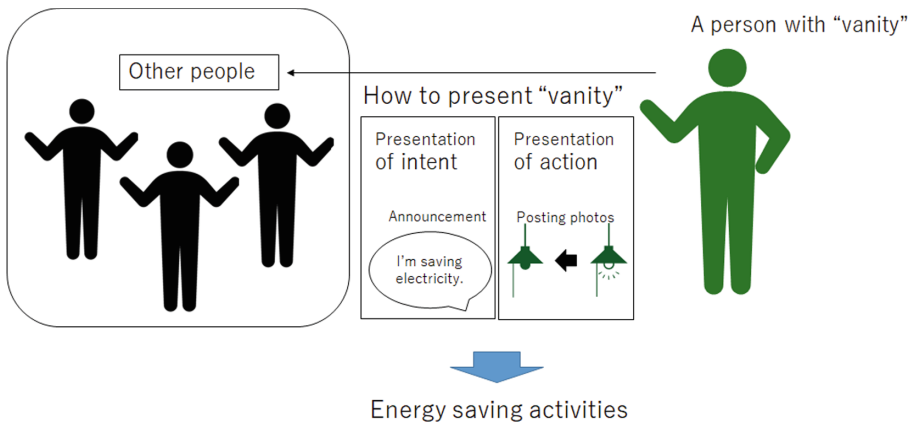


Fig. 3. Outline of utilization of vanity.

As for the intentions, we request users to reduce the power consumption. Here, we use “public commitment” as a mechanism for energy saving activities. Public commitment is to publicly declare the positions on an issue [13]. For example, Pallak and Cummings have shown that when homeowners make a public commitment to energy conservation, they are more likely to comply, compared to homeowners who make the declaration in a less public manner or those who do not make a public commitment at all [14].

When we use the mechanism of public commitment, at first we ask a user for energy saving activities. Then, the user announce the intention. Concretely, the



user could select the intension on how much effort s/he wants to reduce power consumption. When presenting the intention, s/he is required to input in the range of 0% to 100% in increments of 10%. The output of the intension is a bar graph and it shows the intension (%) and the others' intentions (%). Figure 4 shows an example of the output.

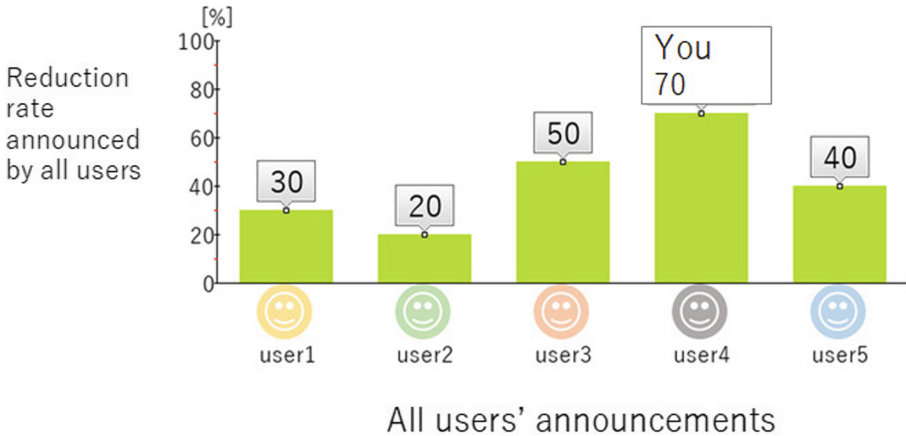


Fig. 4. An example of an output for a user's intention.

As for the actions, we request users to show an action for energy saving. Specifically, in order to confirm the presented action, a photograph is used as a recording medium of the action. And, we request users to show two photos before and after the action for energy saving. The photos would clarify the action the user did by comparing the photo before with the photo after. In addition, in order to show the focused point on the photos, we request the users to add a description about the action. Figure 5 shows an example of a presented action.

## 4 Experiment

### 4.1 Purpose

The purpose of this experiment is to confirm whether anyone is going to utilize vanity with regard to energy saving activities.

### 4.2 Method

The number of the participants is thirty, and the participants were divided into six groups (g1–g6; one group of five participants). The period of the experiment was two weeks. We have prepared an experimental environments on the Web for their presentation of energy saving activities. We gave them the instructions on



**Fig. 5.** An example of a presented action (photos before and after the action).

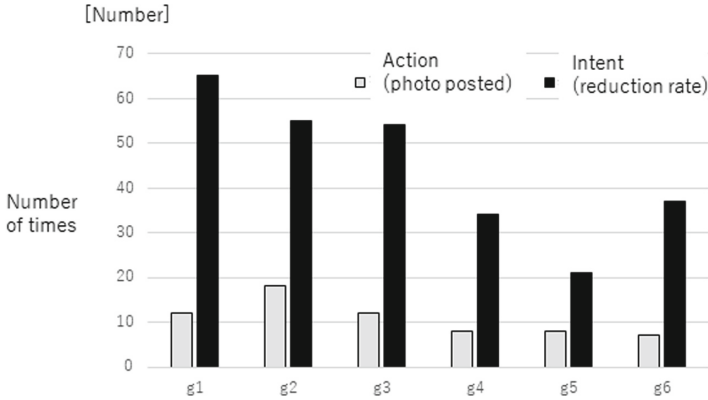
how to use the environment and showed how to input each presentation and how to see the output for the intent and action on energy saving activities. And, there was no explanation on vanity for energy saving activities. After the experiment, we asked them to answer the following questions regarding the intension and action on energy saving activities at the time of their presentation and browsing.

- I want to show it off to the others.
- I want to show the power to someone.
- I want to be praised by someone.
- I want to show it off to the others.
- I want to be praised by someone.

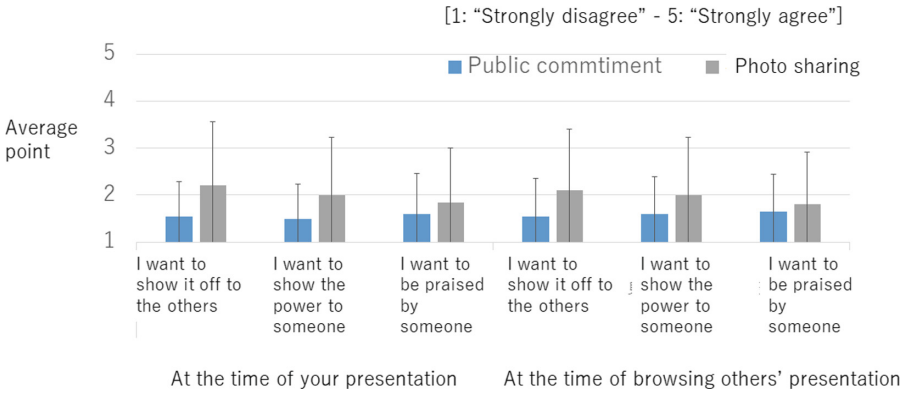
The answers were selected between 1 (“strongly disagree”) and 5 (“Strongly agree”) (five-degree scale).

### 4.3 Results

Figure 6 shows the total number of posts on action and intent for each group. Figure 7 shows the average of the results of the questionnaire asked about vanity.



**Fig. 6.** The number of posts on action and intent for each group (g1–g6).



**Fig. 7.** Results of the questionnaire on the vanity for energy saving activities .

From Fig. 7, the average value of all items was less than 3 for all items, however, there were six participants out of 30 who showed vanity (the value of the answer was 4 or more). In addition, “action” has a higher average value for both input and output than “intent”.

Next, there were the differences in the situation among the groups. There were 4 groups who showed vanity and 2 groups who did not. The former groups were g1, g2, g3, and g4, and latter groups were g5 and g6. From Fig. 6, the number of photos posted was larger in the former groups than in the latter groups. From the results, the participants with vanity might influence the other participants.

## 5 Conclusion

In this study, we proposed a method to support energy saving activities. Specifically, we focused on the mindset of vanity generated from the desire for approval.

Then, we have designed and developed an experimental environment for energy saving activities on the Web, and conducted a two-week experiment with 30 participants. The results would suggest the following points.

1. Among thirty participants of the experiment, there were six participants whose mindset of vanity would be utilized for energy saving activities.
2. When a participant utilizing vanity was in a group, the other participants in the group might have been influenced.

The contribution of this study is to suggest that vanity could be utilized for energy saving activities. In the future, we would like to investigate how to create a place to promote vanity for energy saving activities.

## References

1. Consumer Affairs Agency of Japan: White Paper on Consumer Affairs 2019 (Summary) -Table of Contents-. [https://www.caa.go.jp/en/publication/annual\\_report/2019/](https://www.caa.go.jp/en/publication/annual_report/2019/). Accessed 31 Jan 2020
2. Agency for Natural Resources and Energy of Japan: White Paper on Energy 2019. <https://www.enecho.meti.go.jp/about/whitepaper/2019html/> (in Japanese). Accessed 31 Jan 2020
3. Hirose, Y.: Determinants of environment-conscious behavior. *Japan. J. Soc. Psychol.* **10**(1), 44–55 (1994). (in Japanese)
4. Hirose, Y.: Two-phase decision-making model of environmental conscious behavior and its application for the waste reduction behavior. *Saf. Sci. Rev.* **5**, 81–91 (2015). Faculty of Societal Safety Science, Kansai University
5. Kjeldskov, J., Skov, M.B., Paay, J., Pathmanathan, R.: Using mobile phones to support sustainability: a field study of residential electricity consumption. In: Proceedings the SIGCHI Conference on Human Factors in Computing Systems, pp. 2347–2356 (2012)
6. Froehlich, J., et al.: UbiGreen: investigating a mobile tool for tracking and supporting green transportation habits. In: Proceedings the SIGCHI Conference on Human Factors in Computing Systems, pp. 1043–1052 (2009)
7. Gustafsson, A., Gyllensard, M.: The power-aware cord: energy awareness through ambient information display. In: Proceedings the SIGCHI Conference on Human Factors in Computing Systems, pp. 1423–1426 (2005)
8. Yagita, Y., Iwafune, Y.: The role of visualization with home energy audit to promote energy conservation behavior. *J. Japan Soc. Energy Resour.* **32**(4), 25–33 (2011). (in Japanese)
9. Ueno, T., Nakano, Y.: Development of a support tool for ranking energy-saving activities considering residents preferences. *J. Japan Soc. Energy Resour.* **32**(5), 33–41 (2011). (in Japanese)
10. Petersen, J.E., Shunturov, V., Janda, K., Platt, G., Weinberger, K.: Dormitory residents reduce electricity consumption when exposed to real-time visual feedback and incentives. *Int. J. Sustain. High. Educ.* **8**(1), 16–33 (2007)
11. Aoyagi, S., Okamura, T., Ishii, H., Shimoda, H.: Proposal and evaluation of a method for promoting continuous pro-environmental behavior with moderate communication. *Trans. Hum. Interface Soc.* **13**(3), 31–44 (2011). (in Japanese)

12. Abrahamse, W., Steg, L.: Social influence approaches to encourage resource conservation: a meta-analysis. *Glob. Environ. Change* **23**, 1773–1785 (2013)
13. Cialdini, R.B.: *Influence: Science and Practice*. Allyn & Bacon, Boston (2001)
14. Pallak, M.S., Cummings, W.: Commitment and voluntary energy conservation. *Pers. Soc. Psychol. Bull.* **2**, 27–30 (1976)



# Verification of the Effect of Presenting a Virtual Front Vehicle on Controlling Speed

Tetsuma Konishi<sup>1</sup>(✉), Takayoshi Kitamura<sup>2</sup>, Tomoko Izumi<sup>2</sup>, and Yoshio Nakatani<sup>3</sup>

<sup>1</sup> Graduate School of Information Science and Engineering, Ritsumeikan University,  
Kusatsu, Shiga 525-8557, Japan  
is0258sv@ed.ritsumei.ac.jp

<sup>2</sup> College of Information Science and Engineering, Ritsumeikan University,  
Kusatsu, Shiga 525-8557, Japan  
{ktmr, izumi-t}@fc.ritsumei.ac.jp

<sup>3</sup> Ritsumeikan Trust, Kyoto, Kyoto 604-8520, Japan  
nakatani@is.ritsumei.ac.jp

**Abstract.** Whenever there are no vehicles ahead and there is good road visibility, drivers tend to exceed speed limits even when they normally take care to drive safely. However, in some of these cases, overspeeding may cause serious accidents. In this research, to encourage driving at safe speeds, we propose a system that uses a mobile device installed in a car to visualize a virtual front vehicle. Specifically, when the speed of the real car is faster, the size of the visualized vehicle becomes bigger, as though the driver were approaching the virtual front vehicle. That is, the size simulates approaching the front vehicle. On the other hand, when the speed is slower, the size becomes smaller, as though the front vehicle were moving further away from the driver. We expect that a driver will feel a sense of approaching the front vehicle, notice their fast driving speed from the size of the virtual front vehicle, and slow down. To verify this effect, we conducted a driving simulation experiment.

**Keywords:** Driving support system · Speed control · User interface · Animation

## 1 Introduction

Traffic accidents and traffic violations occur frequently in various places in Japan. According to the National Police Agency, the number of traffic accidents in Japan in 2018 was 430,601, and the number of traffic violations was 6,015,297 [1]. As a measure for the effective prevention of accidents and violations, facilities such as traffic enforcement cameras (ORBIS), traffic signs, and speed bumps have been introduced to road environments [2]. However, because these facilities are installed on the roads, their benefits are only partial. For example, drivers decelerate only in the areas where they know the facilities are installed, and accelerate as soon as they pass these areas. Moreover, to realize the automatic operation (Level 3–5) of all driving tasks in a practical manner, improving traffic laws, in addition to dealing with technical problems, is necessary, an

issue that may require a long time to discuss [3]. Thus, providing a supporting system for safe driving that can be installed immediately is necessary.

In this study, we focus on the problem of the tendency of drivers to exceed legal speeds when driving on good clear roads with no vehicle ahead. To solve this problem, we propose a system for visualizing information as though a vehicle were driving in front of the user, using a mobile device installed in the car of this user. Specifically, when the speed of the real car is faster, the size of the visualized vehicle becomes bigger, as though the driver were approaching the virtual front vehicle. That is, the size of the visualized front vehicle simulates approaching this front vehicle. On the other hand, when the speed of the real car is slower, the size of the visualized vehicle becomes smaller, as though the virtual front vehicle were moving away from the driver. We expect that a driver will feel a sense of approaching this vehicle, notice that his/her driving speed is fast, from the size of the virtual front vehicle, and thus slow down. In this research, a demonstration experiment using a driving simulator was conducted to verify this hypothesis. This paper reports that the proposed method had a certain effect on the reduction of driving speed, based on the results of the demonstration tests.

## 2 Related Works

### 2.1 Relationship Between Driving and Vision

Drivers have to keep recognizing their surrounding situations while they are driving. Of the information that drivers generally recognize, 90% is said to be visual information [4]. The carelessness of a driver with confirming the surrounding situation via his/her visual sense may prevent him/her from grasping the situation accurately and may therefore cause a serious accident. One piece of information that drivers confirm frequently is the distance to the vehicle ahead, which is an important factor in preventing vehicular collisions. Generally, the distance to the vehicle ahead changes depending on the driving speed, and the sense of danger about the collision is approximately proportional to the inverse of the inter-vehicle time. Therefore, the distance between the vehicles is determined based on the recognition of a sense of crisis about a possible collision [5]. Therefore, visual changes of the appearance of the preceding vehicle may be an important clue to the awareness of the sense of crisis and driving speed.

### 2.2 Speed Control Method Using Vision

Much research has been performed about systems that support driving at safe speeds, are implemented through the installation of mobile devices in car interior spaces, such as on the dashboards of automobiles, and present useful information on the screens of these mobile devices. Shimizu et al. [6] have proposed a system to promote safe driving by providing a game point to a user from the viewpoint of collision safety with respect to front vehicles, and by showing the rank of the user. Takada et al. [7], on the other hand, have proposed a game that rewards the user with the pleasure of driving safely, to encourage speeding drivers to slow down spontaneously. The system, which is assumed to be used on an expressway, evaluates a compliance situation with respect to speed

limit and driving speed in real time. According to the results of the evaluation, points are acquired by the user. All of these systems have demonstrated the effectiveness of providing game points as reward for good driving behavior, which encourages speed compliance naturally. However, if a driver prioritizes collecting points over actual safe driving, the driver may perform driving behavior unsuitable for the surrounding situation just to obtain points.

### 3 Proposed Method

In this research, to encourage drivers to control their speeds, we propose a method of visualizing a virtual vehicle travelling in front of a driver. We assume a situation where there is no preceding vehicle on roads with good visibility, such as expressways and bypasses. By showing a front vehicle virtually, we expect that a driver recognizes visually that he/she is approaching the vehicle and slows down because he/she feels anxiety due to the close distance to the virtual vehicle. Specifically, similar to what is shown in Fig. 1, an illustrations or photographs of a vehicle seen from behind is shown on the screen of a device, such as a mobile device or a car navigation system, that is installed in the real vehicle. The size of the illustration is changed based on the driving speed of the real vehicle, to express the feeling of approaching the vehicle in front. That is, when the speed is faster, the size of the visualized vehicle becomes bigger, as though the driver were approaching the virtual front vehicle. On the other hand, when the speed is slower, the size of the visualized vehicle becomes smaller, as though the front vehicle were moving away from the driver.

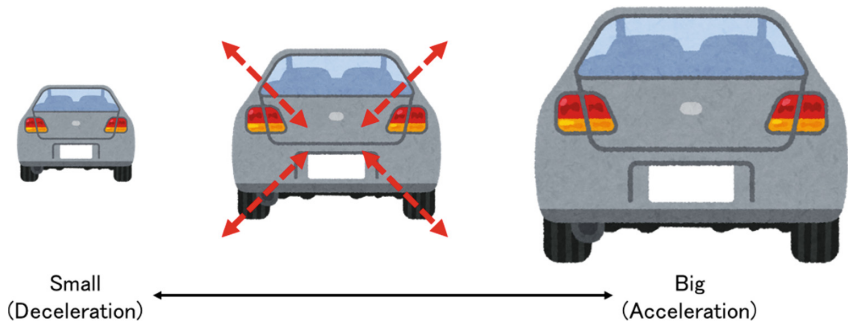


Fig. 1. Example of information visualization for a virtual front vehicle

## 4 Verification Experiment

### 4.1 Outline of Experiment

In this research, we conducted an experiment using a driving simulator to maintain the same environment for all participants. To verify the effectiveness of the proposed system,



we required each of the participants to run three test courses in the driving simulator. In these three driving cases, the participants used a system that shows virtual vehicles under different conditions (listed in Table 1). The conditions of the experiment were different in terms of whether the size of the virtual vehicle displayed on the device changes or not and of whether prior explanation about the system was given or not to the participants. That is, the 1st drive was a normal run, whereas the 2nd drive was a run in which the size of the virtual vehicle changed according to the driving speed, but the size-changing mechanism was not explained before the experiment. This mechanism was then explained before the 3rd drive. We asked the participants to drive under each pattern of conditions and to answer a questionnaire regarding their impressions and feelings about the displayed information.

In this experiment, we evaluate the proposed system via subjective and objective evaluations.

- Subjective evaluation: Do you feel the approximate speed when the size of the virtual vehicle changes to large or small?
- Subjective evaluation: Do you feel proximity and anxiety when the size of a virtual vehicle changes to large or small (become particularly large)?
- Objective evaluation: Is it possible to make the participants follow safe speed using this system? (This question is evaluated from the behavior of the participants and the driving log.)

**Table 1.** Patterns of the experiment

|           | Size of virtual front vehicle changes? | Mechanism of our proposed system explained beforehand? |
|-----------|--|--|
| 1st drive | No                                     | No   |
| 2nd drive | Yes                                    | No   |
| 3rd drive | Yes                                    | Yes  |

To investigate these evaluation points, we required participants to complete the questionnaire shown in Table 2. For each question, the participants answered one of five possible grades (1: Strongly disagree, 5: Strongly agree). Questions No. 1, 2, and 3 were answered after the 1st driving experiment, questions No. 1 to 7 were answered after the 2nd driving experiment, and questions No. 1 to 8 were answered after the 3rd driving experiment.

## 4.2 Experiment Environment

In this experiment, we used the driving simulator shown in Fig. 2, which consisted of a PC, a driving seat, three displays for the driving simulator (for the front, right, and left windows of the simulated car that the participant was driving), a display for operating the

**Table 2.** Questionnaire

| No. | Question items  |
|-----|---|
| 1.  | Did you check the speed while driving?  |
| 2.  | Did you check the system (smartphone) while driving?  |
| 3.  | Did you notice the system (smartphone) while driving?   |
| 4.  | Did it seem as though you were approaching the vehicle, based on the changing information on the display?           |
| 5.  | Did the vehicle seem as though it was driving away from your car, based on the changing information on the display? |
| 6.  | Did you have a sense of approaching the vehicle, based on the changing information on the display?                  |
| 7.  | Did you feel anxious because of the changing information on the display?  |
| 8.  | Do you want to use it in the future?  |

**Fig. 2.** Experimental environment

proposed system, a web camera, and a camera for taking images of the participants. The proposed system assumed that a driver uses a smartphone as a car navigation system, and thus a smartphone was shown on the display as though it were installed on the right side of the dashboard of the driver (Fig. 3). We used a driving simulator from Forum Eight Co., Ltd., and created the road environment using UC-win/Road.



**Fig. 3.** Example of the experiment course and an example of implementation of the proposed system

### 4.3 Overview of the System Used in the Experiment

The system that we constructed displayed a virtual front vehicle separately from the system of the driving simulator. Figure 4 shows an example of the system screens. When a user started driving, a screen simulating a mobile device was displayed. The background photograph of the driving simulator was applied to the background of the mobile-device screen to make it appear as though the virtual vehicle was running forward.

Because this system was independent from the driving simulator, it needed to obtain information on the driving speed. Thus, we used the web camera to determine the driving speed of the image from the display of a driving simulator. The speed data were obtained by converting the image into numbers via image recognition. The display for operation was used to obtain speed information via the web camera, and the system was set to show the same information as that shown by the driving simulator. Whenever the value for speed was determined from the web camera, the result was transmitted to the PC for the operation of this system and was also saved as a log.

When this system detected the running speed of the driving simulator, the illustration of the virtual vehicle was shown on the screen of the system. The system calculated the size of the illustration of the vehicle according to the driving speed acquired by the web camera and updated the illustration. The illustration of the vehicle became larger as the driver accelerated and became smaller as the driver decelerated.

In this experiment, to verify the speed control effect of the proposed system, the size of the illustration of the vehicle was not changed in proportion to the speed. Rather, the size was changed with consideration of the speed limit set in the experiment. An example of the display pattern is shown in Fig. 5. The size was changed in proportion to the driving speed when the speed was slower than the limit speed (i.e., 80 km/h). Whenever the speed limit was exceeded, the size was doubled every 20 km/h to make it easy to understand the change, and to give a feeling of approach and anxiety. Whenever the speed exceeded the speed limit significantly, the virtual vehicle was displayed as being larger than the screen size of the device. As described in Sect. 4.1, the size of

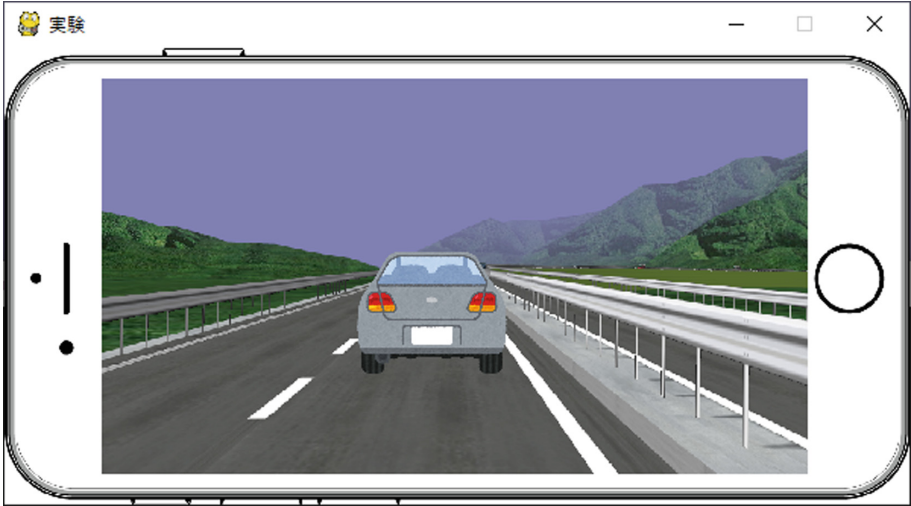


Fig. 4. System screen

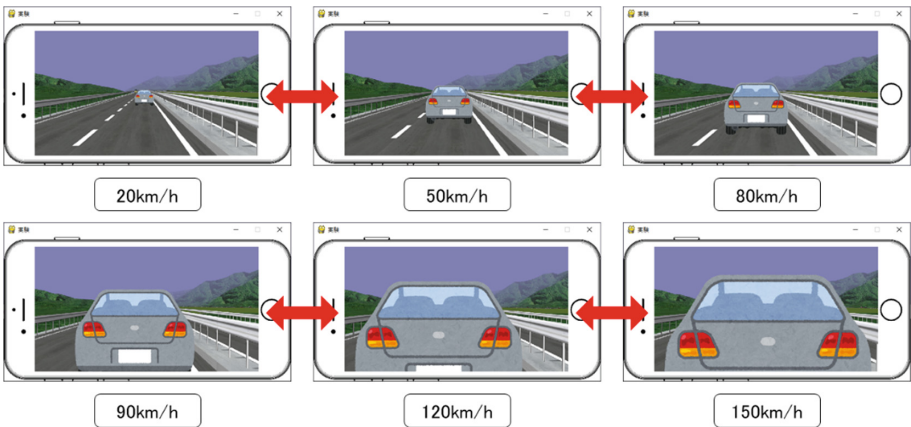


Fig. 5. Display patterns of the system (Above: Under the speed limit, Below: Over the speed limit)

the virtual vehicle did not change in the 1st drive, that is, the size was fixed to the size corresponding to 80 km/h shown in Fig. 5.

#### 4.4 Experimental Course

In this experiment, we configured roads with good visibility, such as expressways and bypass roads. This course was composed of an eight-course track 20 km long. In this experiment, the participants drove up to 15 km from the starting point of this course. We set various vehicles in the left lane with traveling speeds of 80 km/h and no vehicles in the right lane.

## 4.5 Experimental Procedure

Before the experiment, we explained the purpose and outline of the experiment to each of the participants, and we obtained his or her consent to participate in the experiment. To grasp the driving characteristics of the participants, we then used the Driving Style Questionnaire (DSQ) and Workload-Sensitivity Questionnaire (WSQ) created by Ishibashi et al. [8]. Afterward, we explained the procedure and precautions of this experiment. Each participant then performed a simple practice drive to get used to the driving operation of the driving simulator and the feeling of steering and accelerator. In the practice drive, the participant ran on the opposite lane of the course starting from the start position of this experiment, under the condition that there were no other vehicles on the road.

The participants then performed a driving experiment under the patterns of conditions for each of the experiments listed in Table 1. After each experiment, we asked the questions from the questionnaire shown in Table 2 to the participants. After the 2nd drive ended, we asked the question, “Did you understand the mechanism of or the reason for the changing information on the display?” to the participants. Afterward, before the 3rd drive started, we explained the mechanism of the display. After the completion of all experiments, we conducted interviews on their driving behaviors during the experiments and based on the questionnaire results.

## 5 Experimental Results

### 5.1 Questionnaire Results

We conducted the experiment with 21 participants (20 males and 1 female) aged 20 to 25 years (average age: 22.8 years old). Each of the participants has a driver’s license. The attributes of the participants are listed in Table 3.

Table 4 outlines the results of the questionnaire for the three experiments. In the 1st drive, the average score and standard deviation for question No. 1 are 3.95 and 1.00, respectively. The average answers to questions No. 2 and No. 3, asking if the participant confirmed or cared about the presence of the system, are 2.52 and 2.38, respectively, which are both low scores.

Among the results of the 2nd drive, the average score for question No. 1 is 3.81. Both the average score and standard deviation are almost the same as for the 1st drive. On the other hand, the average score for question No. 2 is 3.38 and for question No. 3 is 3.48. These results indicate that the degrees of confirmation and concern for the system in the second experiment are higher than for the third one. The average score for question No. 4, asking how the change of the display was recognized by the participant, is high, at 4.05, whereas the average scores for questions No. 5, No. 6, and No. 7 are 3.52, 3.81, and 3.29, respectively. However, the standard deviations for questions No. 4, No. 5, and No. 6 are slightly higher than those for the other questions. These results indicate that the participants had the various reactions to the system. Table 5 lists a summary of the answers to the question “Did you understand the mechanism of or the reason for the changing information?”, which was asked after the 2nd drive. Of the 21 participants, those who understood the mechanism (Yes) numbered at 13 persons (62%), whereas those who did not (No) numbered at 8 persons (38%).

**Table 3.** Attributes of the participants

| Participant | Gender | Age | Licensed driver for how long? | Driving frequency            | Uses expressway? |
|-------------|--------|-----|-------------------------------|------------------------------|------------------|
| A           | Male   | 23  | 2–3 years                     | Once a month                 | Yes              |
| B           | Female | 22  | 2–3 years                     | Almost every day             | Yes              |
| C           | Male   | 23  | 3 years or more               | Almost every day (Motorbike) | Yes              |
| D           | Male   | 23  | 3 years or more               | Once a month                 | Yes              |
| E           | Male   | 25  | 3 years or more               | Once a month                 | Yes              |
| F           | Male   | 22  | 3 years or more               | Once a month                 | Yes              |
| G           | Male   | 22  | 2–3 years                     | Once a week                  | Yes              |
| H           | Male   | 25  | 3 years or more               | Once a year                  | Yes              |
| I           | Male   | 21  | 3 years or more               | Once a month                 | Yes              |
| J           | Male   | 24  | 2–3 years                     | Almost every day (Motorbike) | Yes              |
| K           | Male   | 23  | 3 years or more               | Once a year                  | Yes              |
| L           | Male   | 24  | 3 years or more               | Almost every day (Motorbike) | Yes              |
| M           | Male   | 22  | 1–2 years                     | Once a month                 | Yes              |
| N           | Male   | 22  | 2–3 years                     | Once a month                 | Yes              |
| O           | Male   | 20  | Less than 1 year              | Once a year                  | Yes              |
| P           | Male   | 21  | Less than 1 year              | Once a month                 | Yes              |
| Q           | Male   | 21  | 3 years or more               | Once a month                 | Yes              |
| R           | Male   | 23  | 1–2 years                     | Almost every day (Motorbike) | Yes              |
| S           | Male   | 24  | 3 years or more               | Once a month                 | Yes              |
| T           | Male   | 22  | 2–3 years                     | Once a year                  | No               |
| U           | Male   | 24  | 1–2 years                     | Once a year                  | Yes              |

Meanwhile, among the results of the 3rd drive, the average scores for questions No. 1, No. 2, and No. 3 are 4.00, 4.10, and 3.95, respectively. These scores are the highest, and the standard deviations are the lowest, compared with for the other experiments. That is, there are little variations in the responses, and many of the participants gave the positive responses to the questions. Answers to question No. 4 exhibit a similar tendency to the results for the 2nd drive. The average score for question No. 7 is 3.38, which is almost equal to that in the 2nd drive. Finally, the average score for question No. 8 is 2.71.

The results of the after-experiment interviews with each participant are then analyzed. The question contents of the interviews are different among different participants because

**Table 4.** Summary of the results of the questionnaire

| No. | 1st drive  |      | 2nd drive  |      | 3rd drive  |      |
|-----|------------|------|------------|------|------------|------|
|     | Evaluation | SD   | Evaluation | SD   | Evaluation | SD   |
| 1.  | 3.95       | 1.00 | 3.81       | 0.96 | 4.00       | 0.62 |
| 2.  | 2.52       | 1.10 | 3.38       | 1.13 | 4.10       | 0.81 |
| 3.  | 2.38       | 1.09 | 3.48       | 1.10 | 3.95       | 0.90 |
| 4.  |            |      | 4.05       | 1.17 | 4.29       | 0.76 |
| 5.  |            |      | 3.52       | 1.37 | 3.95       | 1.09 |
| 6.  |            |      | 3.81       | 1.14 | 4.00       | 0.93 |
| 7.  |            |      | 3.29       | 0.93 | 3.38       | 0.90 |

**Table 5.** Summary of the answers regarding about understanding of system structure

|                       | Yes       | No       |
|-----------------------|-----------|----------|
| Number of respondents | 13 people | 8 people |
| Percentage            | 62%       | 38%      |

we asked the questions depending on the driving behaviors noticed by the experimenter noticed and on the questionnaire result. According to the interviews, impressions and actions felt in the 2nd and 3rd drives differed depending on whether the mechanism of the system was understood or not. Some of the 13 participants who replied “I understood (Yes)” said, “From the 2nd drive, I was surprised and slightly scared by the system display,” “Checked the speedometer after checking the system display,” and “I tried to find a speed that resulted in a good illustration size.” On the other hand, some of the 8 participants who replied “I didn’t understand (No)” said, “I was aware that the display of the system had changed in the 2nd drive, but I did not understand the mechanism, so I continued driving,” and “I started to realize the mechanism after I heard the explanation on the system.” As a common opinion, the improvement of display size, change of illustration, change of the lane where the virtual vehicle runs when the drivers change their lanes, and dealing with the experience of using the system are pointed out as opportunities for improvement. Furthermore, the degrees of the sense of, or the anxiety from, approaching the vehicle are different among the participants.

## 5.2 Analysis of Experimental Data

Table 6 lists the average speeds of the participants. The 3rd drive resulted in the slowest average speed, whereas the 1st drive resulted in the fastest. We performed a t-test at the 5% level to confirm the differences in the average speeds of the 1st and 2nd, 2nd and 3rd, and 1st and 3rd experiments. Table 7 lists the results of the t-tests. Each comparison uses  $p < 0.05$ , and significant differences are said to exist.

**Table 6.** Average speeds of the participants (km/h)

| Participant | 1st drive | 2nd drive | 3rd drive |
|-------------|-----------|-----------|-----------|
| A           | 105.23    | 109.33    | 107.74    |
| B           | 110.55    | 88.4      | 83.61     |
| C           | 109.3     | 80.44     | 94.94     |
| D           | 113.62    | 117.12    | 108.26    |
| E           | 90.47     | 88.3      | 64.62     |
| F           | 98.65     | 115.2     | 90.92     |
| G           | 95.66     | 56.55     | 71.49     |
| H           | 118.95    | 114.62    | 77.94     |
| I           | 129.35    | 116.94    | 103.75    |
| J           | 131.17    | 129.32    | 125.74    |
| K           | 73.13     | 43.7      | 42.16     |
| L           | 89.87     | 90.62     | 87.52     |
| M           | 96.01     | 75.08     | 78.73     |
| N           | 76.62     | 102.63    | 79.7      |
| O           | 117.78    | 107.08    | 93.5      |
| P           | 92.06     | 88.34     | 79.55     |
| Q           | 95.86     | 71.6      | 67.78     |
| R           | 114.05    | 115.45    | 121.37    |
| S           | 94.64     | 90.87     | 91.73     |
| T           | 84.66     | 85.05     | 86.07     |
| U           | 122.4     | 114.23    | 116.53    |
| Average     | 102.86    | 95.28     | 89.22     |

**Table 7.** Results of t-tests for average speed

|         | 1st and 2nd drive | 2nd and 3rd drive | 1st and 3rd drive |
|---------|-------------------|-------------------|-------------------|
| p-value | 0.03843           | 0.04351           | 0.0001623         |

For each experiment, we then confirm, using Pearson correlation analysis, whether there is a relationship between the results of the answers to the questionnaire and the average speeds. Among questions No. 1 to No. 7, we performed the analysis on questions No. 1, No. 3, No. 4, No. 6, and No. 7, which are considered to be related to controlling the speeds of the car driven by the participants. Table 8 outlines the results of the correlation analysis between the response results to the questionnaire and the average speeds. In this table, strong correlation is denoted by  $\bigcirc$ , weak correlation by  $\Delta$ ,



and no correlation by  $\times$ . For the 1st drive, a negative correlation is observed for question No. 1, and a weak negative correlation is observed for question No. 3. For the 2nd drive, a weak negative correlation is observed for question No. 1, and negative correlations are observed for the other questions. Finally, for the 3rd drive, no correlation is observed for question No. 1, whereas weak negative correlations are recognized for the other questions.

**Table 8.** Results of correlation analysis

| No. | 1st drive               |         | 2nd drive               |         | 3rd drive               |         |
|-----|-------------------------|---------|-------------------------|---------|-------------------------|---------|
|     | Correlation coefficient | Results | Correlation coefficient | Results | Correlation coefficient | Results |
| 1.  | -0.63                   | ○       | -0.39                   | △       | -0.07                   | ×       |
| 3.  | -0.26                   | △       | -0.69                   | ○       | -0.32                   | △       |
| 4.  |                         |         | -0.53                   | ○       | -0.29                   | △       |
| 6.  |                         |         | -0.57                   | ○       | -0.25                   | △       |
| 7.  |                         |         | -0.43                   | ○       | -0.33                   | △       |

### 5.3 Consideration

According to the results of the questionnaire in Sect. 5.1, for question No. 1, which asked whether the participants were concerned with their speeds, the average score increased in the 3rd drive. However, this score was almost the same for the 1st and 2nd drives. Furthermore, for questions No. 2 and No. 3 on the consciousness of the participants about the system, the scores increased from 1st drive to 2nd drive and from 2nd drive to 3rd drive. The reason for the increase in score from the 2nd to 3rd drives seems to be the explanation of the proposed system being given after the 2nd drive. Because the score also increased for the 2nd drive compared with that for the 1st drive, the consciousness about the system is said to have been improved by the change in image size to large and small.

The results for questions No. 4, No. 5, and No. 6 were high for both the 2nd and 3rd drives. From these results, the design is said to be intuitively easy to understand. However, the resulting score for question No. 5, regarding the information that the vehicle is going away, was low. The reason for this low score is considered to be from the participants seemingly not needing to step on the brake in the expressway course because of the ability to slow down by releasing the accelerator pedal instead.

According to the results for question No. 7, which asked about the sense of unease, the scores were almost same, and were both low, for the 2nd and 3rd experiments. As a result, although the proposed system can provide a sense of approaching the vehicle in front, the virtual approach to the preceding vehicle don't produce a sense of uneasiness. Thus, design considerations should be made to create this necessary sense of uneasiness.

We then consider the average speeds in the experiment. The average speeds in the 2nd and 3rd drives were lower than in the 1st drive, with a significant difference. From this result, the change in size of the virtual front vehicle is said to be effective for speed control. In addition, because the speed in the 2nd drive was slower than that in the 1st drive, the speed will become slower even when no explanation about the mechanism of the system has been given. However, the average speed decreased in the 3rd drive, and the speed is said to be further decreased by explanations on the proposed system.

Afterward, we explain the relation between the results of the answers to the questionnaire and the average speeds. For the 2nd drive, there were negative correlations for most of the questionnaire items and the average speed, i.e., the average speeds became slow for many participants, as the scores for the questionnaire became high. Based on this observation, adjustments in the speed appeared to be due to the participants confirming changes in the image size. The degree of the correlation became lower for question No. 1 for the 2nd drive compared with that for the 1st drive. This result is considered to be due to the participants getting used to the 2nd experiment operation. Moreover, in the 3rd drive, the correlation between average speed and question score became low, which was caused not only by the driving operation but also by habituation to the display of the system.

In summary, we consider the evaluation points listed in Sect. 4.1.:

- “Do you feel the approximate speed when the size of the virtual vehicle changes to large or small?”: According to the results of the questionnaire, the participants were paying attention to the change of size. They were concerned about their speed, and thus their average speeds decrease. Furthermore, in the interviews, we obtained opinions such as “The size showed how fast I was driving,” and “I found out I can set the image of this size by a certain speed.” Therefore, we consider that the drivers were able to estimate their approximate speeds using the proposed system.
- “Do you feel proximity and anxiety when the size of a virtual vehicle changes to large or small (become particularly large)?”: From the results of the questionnaire and interviews, we consider the participants feeling a sense of approaching the visualized vehicle to be a given. However, not many experiment participants felt anxiety from this part of the experiment. Therefore, an examination of the design is needed to trigger the required anxiety.
- “Is it possible to make the participants follow safe speed using this system? (This question is evaluated from the behavior of the participants and the driving log.)”: From the driving movie data, we confirmed that the participants looked at the speedometer after looking at the system. In addition, their average speeds significantly decreased in the 2nd drive and 3rd drive, compared with the 1st drive, whenever the size of the virtual front vehicle changed. For this reason, it is concluded that the system was able to encourage drivers to decrease their speeds.

## 6 Conclusion

In this research, we proposed a system that uses a mobile device installed in a car to visualize information as though a vehicle was running forward in front of the car.

Specifically, we proposed a system that expresses a sense of approaching the visualized vehicle by changing the size of the illustration based on the driving speed. In a verification experiment using a driving simulator, we conducted the driving tests with 21 participants. The results of our questionnaire and an analysis of the experimental data showed speed control as an effect of the visualization by the proposed system.

As a future plan, we consider varying the content according to the actual driving environment to make the visualization more realistic. In addition, we need to conduct experiments with enhancement expressions, where approaching the virtual vehicle is emphasized by different images or by changes in the background color.

**Acknowledgment.** This work is supported in part by KAKENHI no18H03483.

## References

1. National Police Agency (Japan): Statistical Data of Police White Paper 2019 (2019)
2. Ministry of Land, Infrastructure, Transport and Tourism (Japan): Road Bureau: Safety. [http://www.mlit.go.jp/road/road\\_e/safety.html](http://www.mlit.go.jp/road/road_e/safety.html). Accessed 18 Jan 2020
3. Prime Minister of Japan and His Cabinet (Japan): Public-Private ITS Initiative/Roadmap 2019 (2019)
4. Hartman, E.: Driver Vision Requirements. Society of Automotive Engineers Paper 700392, pp. 629–630 (1970)
5. Kondoh, T., Yamamura, T., Kitazaki, S., Kuge, N., Boer, E.R.: Identification of visual cues and quantification of drivers' perception of proximity risk to the lead vehicle in car-following situations. *J. Mech. Syst. Transp. Logist.* **1**(2), 170–180 (2008)
6. Shimizu, S., Shidoji, K., Matsuki, Y., Uekusa, O., Kato, N.: In-vehicle effect measurement for a system to prevent dangerous driving. *J. JSAE* **42**(3), 789–794 (2011)
7. Takada, S., Hiraoka, T., Saito, A., Fujii, T., An, S.: Experimental study about effectiveness of expressway driving game based on gamenics theory on driver behavior. *J. JSCE Ser. D3 (Infrast. Plan. Manage.)* **73**(5), 971–980 (2017)
8. Ishibashi, M., Okuwa, M., Akamatsu, M.: Development of driving style questionnaire and workload sensitivity questionnaire for drivers' characteristic identification. *Proc. JSAE* **55**(2), 9–12 (2002)



# Roles on Corporate and Public Innovation Communities: Understanding Personas to Reach New Frontiers

Maximilian Rapp<sup>1,3</sup>(✉), Niclas Kröger<sup>2,4</sup>, and Samira Scheerer<sup>3</sup>

<sup>1</sup> Skolkovo Institute of Science and Technology, Moscow, Russia

maximilian.rapp@de.ey.com

<sup>2</sup> HYVE, Munich, Germany

niclas.kroeger@hyve.net

<sup>3</sup> EY (Ernst & Young), Munich, Germany

samira.scheerer@de.ey.com

<sup>4</sup> HYVE: University of Salzburg, Salzburg, Austria

**Abstract.** Innovation communities have been a focus of research over the last decade to foster and analyze how businesses and the public sector can engage customers or citizens to co-create products, strategies or services. While an applied and professional community management is seen as key, most initiators still do not understand the different personas and characters about their communities to adjust their management on an individual level and to embrace further potentials. In this paper, we analyzed 64 communities to pinpoint 11 different personas, which can be found repeatedly on either public or industry driven innovation communities. Based on our findings about their characteristics, motivation triggers and behavior on innovation communities, we offer managerial implications to enhance strategies and community management systems of innovation communities.

**Keywords:** Community management · Community roles · Open innovation · Open government · Innovation communities · Crowdsourcing · Public participation · Motivation and participation

## 1 Introduction

Scholars, public institutions and corporations alike have done extensive work analyzing, testing and implementing innovation communities. Online communities or social media platforms are used strategically to integrate customer voices [1], to listen to citizens and their opinions, to collaboratively work on products or policies [2] and more. However, recent publications as well as the experience of international brands and influential governmental organizations show that the potential of Open Innovation platforms still has not reached its limits and leaves different areas with high potential uncovered [3]. While the number of initiatives in the public and private sector and the level of professionalism is increasing, still many innovation communities are shut down, the implementation rate is inefficient, or the efforts are not seen as sustainable worthwhile compared with the

outcome [4]. The reasons for a partial unproductive use can be explained several different ways, including the following:

**Topic Progress:** Innovation communities and their use through crowdsourcing to filter trends, share experiences and submit ideas have been part of the corporate and public world for over a decade. The quantity of initiatives has led to a numbing effect within the recipients, picking merely the most interesting initiatives with the highest rewards - intrinsic and/or extrinsic - to participate. Therefore, a decrease in quality and quantity of contributions can be detected - at least within those initiatives, which processes have not changed over the years. Innovative approaches like using a social crowd just to give feedback overnight or within sprints to spice up the challenge (gamification, shorter runtime, short descriptions) can increase the time participants spend in the community and make week-long interactions on simple idea contests as well as evaluating each other's concepts more attractive [5].

**Implementation Process:** Organizations have always struggled with Open Innovation projects and especially in crowd-based initiatives to evaluate, prioritize and eventually implement the gathered insights for final products, strategies, concepts or policies [6]. The progression of cultural change, adoption of lean management and Open Innovation frameworks have cut deeply into all kinds of organizations and leave the responsible managers in sheer desperation of how to transfer and adopt citizen or consumer centric perspective into the innovation process. Innovation communities and their insights are no exception as their management and processes within the organizations are usually decoupled from the innovation management process which leads unwillingly to an implementation gap.

**Community Management:** The success of innovation communities has always been linked to the management of the communities and consequently to the handling of individual profiles, personas and the social behavior of the crowd within a given framework [7–9]. While many publications use cases and practitioners have highlighted the importance and explained tools to deliver professional community management, the exact roles and recurring personas within communities have not been understood holistically. By analyzing existing research and executed projects, it becomes evident that organizations - private and public alike - have focused on a professional, but mainly generalist community management; an in-depth understanding of the individual behavior of the crowd is still missing. The community is a fragile social environment, mirroring needs, behavioral patterns but also moods of society-at-large. To deliver customized community management, motivation, dynamics and behavior of the recurring characters (personas) is key.

The aforementioned frontiers of innovation communities are significant bottlenecks and therefore a matter deemed worthy of analysis. As a starting point and focus of this research, we want to give a broad coverage by diving deep into the latter frontier, namely the challenge of how to deliver customized community management for individual personas and profiles to engage in a more sustainable way and to get the maximum out of community-based innovation initiatives.

Firstly, we cluster and explain the different existing roles and personas on social media collaboration platforms driven through crowdsourcing. We show clusters as well as patterns of behavior of 11 different roles and personas. Additionally, we compare those in the scope of the public and the private sector to deliver insights about potential learnings from each other's sectors and to derive management implications for stakeholders from the practical field on how to handle and manage communities. This improves the outcome and chances of implementation by specifically understanding the personas and roles of the participants.

Basis for our discussed insights is a data set from 64 innovation communities from political and public institutions as well as from companies around the world. The research team has accompanied each one of those (Participatory Action Research) during recent years and conducted various interviews with stakeholders, citizens and customers. To summarize, all cases have been fundamentally analyzed and are a fruitful source of data to answer our research questions:

- 1) Which roles and personas are existing on social media collaboration platforms?
- 2) What is their primary motivation to participate?
- 3) What differences exist between public and industry led innovation communities regarding those roles and personas and the frequency of their appearance?

As we have already indicated in the introduction, many other frontiers can be listed. Consequently, we desire to provide a longer-term perspective concerning further research efforts and potential follow-up analysis about this topic after highlighting first managerial implications in the scope of our findings.

## 2 Theoretical Background

Online communities are used to generate innovation for several years now. In the private sector they are used to create new products and services or to solve problems [10, 11]. The public sector also has been using innovation communities for a while now for different purposes like policy creation and party programs or even entire constitutions [12, 13]. While this approach works very well in an open setting involving customers and external experts, communities are also used within organizations to connect employees virtually and create innovations [14]. Literature is using a variety of names to describe these internal and external as well as private and public sector communities. In the scope of this research study, we will use the term innovation communities to describe communities where people join an online platform with Web 2.0 functionalities to submit, discuss and evaluate ideas. Other terms that may be used in this context in a similar manner include crowdsourcing contests, social media collaboration platforms, call for ideas, innovation contest, crowdsourcing, E-Participation, Open Government or Open Innovation communities.

The framework in which the community is established (i.e. internal vs. external, public vs. private sector) also determines the kind of participants attracted to participate and hence, the results that can be generated within the different communities. The community members of the different initiatives are also motivated by varying factors which results in different measures to be taken in order to manage and motivate the community [15]. Some research has already been done to identify different roles in communities.

Hutter et al. [9] have shown, there are four major kinds of participants (competitors, community members, observers, co-operators) active in industry innovation communities. However, only limited suggestions on how to manage them was provided. In the public sector, the knowledge about the different user groups is limited to so-called lurkers (i.e. inactive participants) [15]. A major trigger for their participation is their underlying motivation. Research indicates different results for the motivation of participation when looking into specific communities [16–18]. There are various motives why people participate in innovation communities and often times, it is a combination of different factors that lead to participation. These can be divided into extrinsic and intrinsic factors [19]. Some love to solve problems, participate in order to learn, want to have a positive impact on others or be part of a community and are therefore driven by intrinsic motivation [20]. At the same time, also extrinsic motivation factors like the chance to win prizes, to use the new product or service or to get recognized, play an important role [16, 21]. In contrast, in the public sector the motivation may differ. Most participants are motivated intrinsically by political interest [22] or to meet one's own needs [23] as prize money is rarely offered to incentivize participation [24].

Besides the motivation for participation, other factors may also influence how participants behave in the community. Group norms and the social identity of group members have an influence on their participation behavior [25]. At the same time, the dynamics within groups may have an influence on the behavior [8] leading to a great variety of different types of community members. In software product development, the issue of addressing the needs of specific groups is done through personas which provide a better understanding of the characteristics, needs and values of different user types [26, 27]. In communities, the value does not lay in the right features but rather how it is managed to provide the desired outcome for the initiators [28]. Combining the concepts of personas and community management leads to the need of a persona-specific management of innovation communities.

### 3 Methodological Approach

To gain an in-depth understanding about potential roles on innovation communities and how they differ between private sector and public sector, we use the Participatory Action Research Approach (PAR) [29]. In PAR, the researchers become active participants in the cases they want to study, thereby leveraging the full potential of the case study method [30]. Through the deep immersion into the topic they gain valuable insights and get a better understanding of the focus topic. This qualitative method allows approaches to new research fields in an exploratory manner. In this research project, the research team has been managing the innovation communities for or with the public or private entities mostly in the scope of project work. In addition, interviews with users have been conducted to get a better understanding of the underlying motives for their participation and action/in-action within the communities. Besides and fundamentally this qualitative approach is backed by a large data set as 64 innovation communities have been accompanied by the researchers. The number of innovation activities (e.g. contests) is even higher as the researcher accompanied different activities in several of the communities. While 29 of them were hosted by public entities, 35 of them were hosted by companies. These

communities can further be distinguished as 23 internal (with own employees or specific group) and 41 external driven initiatives (innovation communities) (see Table 1). Due to the lower market acceptance and penetration for internal communities in the public sector as well as more difficult access to internal industry communities, the number of cases is lower in these areas.

**Table 1.** Distribution of the sample according to sector and target group

|          | Industry | Public | Total |
|----------|----------|--------|-------|
| Internal | 13       | 10     | 23    |
| External | 22       | 19     | 41    |
| Total    | 35       | 29     | 64    |

The innovation communities span a variety of different topics, but the technical functions of the underlying software platforms were almost identically. In internal innovation communities, only employees of the respective companies could participate and could log in with their company accounts. For the external innovation communities, users could create an account on their own and there was no pre-selection of the participants. All participants were then able to submit own ideas, comment on other participants’ contributions and evaluate these. Ideas could be text only or include further attachments like images or PDF files. On the platforms, all submissions were visible to all users who could get inspiration from the content that has already been submitted. Each user had a personal profile where they were able to share some basic personal information. In most cases, the participants also had a pin wall on which other participants could leave personal messages separated from idea contributions.

The innovation communities displayed in Table 1 show some examples of the cases that have been used as the basis for this analysis. Their explanation provides a better understanding of the research subjects. Further organizations that are part of the study are for example the Wirtschaftskammer Tirol (chamber of commerce Tyrol), the German ministry of economics or the Austrian ministry of finance in the public sector and companies like BMW, Lufthansa, Intel or P&G in the private sector (Table 2).

**Table 2.** Exemplary selection from different fields of cases used in the research project

|          | Industry         | Public |
|----------|------------------|--------|
| Internal | Zodiac Aerospace | CSU    |
| External | Vodafone         | ÖVP    |

**Zodiac Aerospace - Open Innovation Challenge:** The Open Innovation Challenge was a community that was built by the French aerospace company Zodiac Aerospace to leverage the creative power of their employees regarding the airplane interior of the



future. The participants were free to submit any idea regarding the cabin interior of commercial aircrafts. The platform was open to all Zodiac Aerospace employees worldwide. Prizes for the participation were two Apple Watches. Within the contest, 2,746 employees registered on the platform. They submitted 610 ideas and 2,051 comments.

**Vodafone - Connected X Challenge:** The Connected X Challenge was a community that was built by the telecommunication provider Vodafone in Germany to find new use cases for an existing technology that had been developed by Vodafone. On the platform, the participants were asked to submit ideas about new use cases for Vodafone's new Narrowband IoT hardware platform. The platform was globally open to everyone who was interested in the topic. Prizes for the participants included money, a visit to a trade fair where the winners were awarded and non-monetary prizes like smartphones. In the scope of the contest, 556 participants registered in the community and submitted a total of 406 ideas, 1,254 comments and 1,300 evaluations.

**ÖVP - Ideenwand:** The Ideenwand platform and community was built by the Austrian people's party ÖVP to gather ideas and opinions to build the basis for their new party program within their Evolution Volkspartei initiative. On the platform, participants were asked to submit ideas and discuss political topics that were relevant to them. The platform was initially open to everyone. In a second voting phase, only party members could vote to determine a certain short list of guidelines for the program. During the campaign, 5,337 citizens registered, submitted 2,949 ideas, wrote 6,537 comments and evaluated almost 12,000 likes and dislikes.

**CSU - WikoNet:** The WikoNet was a community that was built by the Christian Social Union (CSU) in Bavaria to connect experts from politics, economics and scientific experts within the political party. On the platform, participants could submit own ideas and discuss the presented content. The platform was only open to party members of the CSU, who discussed the future of their economic strategy. Over 100 submissions by 52 selected experts were discussed several hundred times.

## 4 Findings and Discussion: Community Roles and Comparative Insights on Patterns and Motivation Within Internal/External as Well as Public/Industry Collaboration Platforms

After analyzing 64 internal and external organized initiatives with close connection to the project sponsors/initiators and users/employees (e.g. through interviews), we could identify clear and recurring behaviors as well as patterns which we clustered in 11 different roles and personas. In the following paragraphs we present them in detail by highlighting their characteristics, uncovering their motivation for participation and by illustrating the major differences between public and industry led innovation communities targeting an internal and external audience:

**Community Managers** are responsible for leading and operating the platform as well as activating the community regarding idea generation. They are the binding element between organizations and the virtual group. In this role, they are communicating and

de-escalating with the main purpose to keep the community values and netiquette. Every community should have multiple and professional community managers, as they usually also review the content, forward ideas to experts within the organization and track the data, movements and dynamics on the platform for potential counteracting. The number and efforts of community managers varies from case to case, being tightly correlated to topic, number of employees, recruiting efforts or fan base; ergo, to the size of the (expected) community. By accompanying and evaluating the different initiatives in this paper, we found that political external driven communities need the highest amount of community managers (max. value: crowdsourcing political party program of the ÖVP in Austria with up to 15 community managers), while internal platforms (industry and public) and external collaboration platforms in the private sector level out at three to maximum five community managers (thumb rule each with 2 h efforts a day besides their daily jobs). If quality and quantity is missing on that end, it usually results in sustainable deterioration.

**Opinion Leaders** have a great interest in the topic and are highly motivated (intrinsic). Next to their active engagement on the platform, they even support the community managers by fostering discussions, demonstrating values and motivating others (socially engaged). Usually those highly value-oriented participants can be found more on external public communities due to the strong correlation to intrinsic triggers of participation. Additionally, our analysis shows that there is also a difference within the public sector. They are more likely to engage in political party driven communities than on those initiated by ministries or public institutions. Interestingly, the sight of this persona on internal platforms is very rare, as questions would be raised about their enormous time spent on the platform on top of the core task, to submit ideas and discuss. In a way, they can be seen as entertainers, as they market the purpose and initiative through various channels (this is the case on internal and external communities). This starts f.i. with employees communicating their own ideas in meetings, coffee breaks, lunch breaks and phone calls (offline push regarding internal platforms) and continues with writing blog articles, sharing social media postings or fostering group discussions as external participant for vitalizing and energizing the community of their beloved brands (online push regarding external platforms).

**Power Ideators** are widely welcoming characters on a community as they generate a huge number of qualitative ideas or further develop already existing idea sparks and concepts. They usually engage in discussions, but mainly focus on their own contributions. This is due to their intrinsic but also extrinsic motivation to participate. They try to publish as many ideas as possible and in a very early stage to occupy whole topic fields. Other contributors can just enhance the ideas via comments but cannot publish similar ideas as this would seem to be simply copycats. They usually have a broad knowledge across industry, so they simply apply ideas from other industrial areas and adjust them to the use case of the innovation community. Regarding this role we see a significant difference between the two sectors, not in frequency of their appearance, but for reasons of their motivation. Public communities are inclined to attract those power ideators who are driven by a far more intrinsic calling to change the status quo than those participating in communities across industry who usually share this high number of ideas to raise their chances to win prizes or to get acknowledged by the initiators to gain improved access

and connections. However, the number of power ideators is very small in general, but even in thin quantity (under 20 power ideators on average on the analyzed communities) they can be accountable for more than 40% of all the submitted ideas in a single innovation community. At least that has been the result in more than 25% of all the analyzed communities. Internal communities show even lower numbers of power users due to the lack of expertise of emerging technologies (many internal ‘call for ideas’ are based around those topics), lack of time (users on external communities use their free time, employees usually use their work time as technological access to the platform is limited to the workplace) or simply due to missing incentives.

**Supporters** are driven by intrinsic interests and join the community to promote a certain topic. They actively engage and promote the ideas of others with constructive comments to take their input to the next level. Most winning ideas are usually reworked several times during a contribution phase as comments and critics as well as evaluation push the ideator to react and enhance the idea continuously. Interestingly, supporters predominantly do not have own ideas but try to share their expertise to push other contributions and to create value. Parallel to the reasons for the lack of power ideators on internal collaboration platforms, this support role is hardly found both for public and private organizations. Reading and commenting on other ideas is time intensive which means that regular work must get done later or even after the regular office hours. Most internal platforms can be accessed only through the companies’ or organizational networks as remote accessibility is ranked as risky. Consequently, user focus on their own ideas and evaluations. This leads to an exciting finding, namely the significant lower number of exchange and discussions on internal communities compared to those with an external crowd, even though a homogenous and already familiar (and perceived safer) community and environment of employees seems to be more interactive than a community mixed with strangers across the globe. Moreover, our data shows that supporters on external platforms have declined in numbers over the past years, while this role can still be seen frequently on public communities. The main reasons for this development are the rising number of international brands using community-based approaches to foster customer voices and its consequence of a non-exclusive status, in simpler words - too much competition for interested users and ideators. In contrast, we have seen that the number of participants in general on internal platforms is usually higher than those on external ones (still depends on the size of the organization which initiates the community) even though the target group is limited as no outsiders are allowed. It seems that internal communication and CEO messages to the employees has a magnetic impact and is highly efficient for including the staff, while compared gathering a global crowd needs a lot more community building and communication efforts as well as channels.

**Experts** are actual luminaries for the field in demand. With their background knowledge, they make a valuable contribution to the community and their specialist knowledge provide high-quality ideas and comments. The involvement of experts is one of the main goals for external communities as variously and often quoted in innovation, ‘most of the smartest people work for someone else’ and outside-in perspective is the main reason for pushing forward Open Innovation programs. The acquisition and recruitment of those high-end contributors is difficult and elaborate, albeit critical for the qualitative input. Simply described - the more external experts the merrier. On the contrary,

regarding internal platforms we can summarize through our data that experts are usually also feedback providers for the whole community and part of the respective department running the topic/call for ideas on the platform. Here, the number of experts is usually low (around 5–10 on average), but their part is key as they drive the evaluation and prioritization process by picking out the best ideas which will be further worked on for implementation.

**Profit-oriented** participants are also submitting ideas (more than discussing them as ideas submissions are usually better incentivized), especially on external platforms. The reason for the engagement is more extrinsic than intrinsic, as they usually have a certain goal. In most cases analyzed, they simply try to win the offered prize money or other exclusive prizes by sharing their thoughts and ideas. But not all motivation is triggered by monetary elements, many professionals also seek a possibility to get connections to the organizer of the community and convince them about their expertise as they are on the hunt for a new job. While this persona is very common in the industry, public participants usually try to profit in a different way as in most of all cases, public organizations do not give away material or monetary prizes anyways. Here, the citizens - if motivated extrinsically - rather hope for political influence, being part of the decision-making process or access certain networks. This phenomenon is pretty much mirrored within internal communities. Corporations tend to give away phones, extra vacation days, or very exclusive prizes for ideators, wherefore the penetration of participation is usually higher than in the public sphere, where no incentives at all is common. Furthermore, and not surprising, we analyzed that in both areas many profit-oriented driven users hope for leaving a 'digital business card' to apply for a better or new job through their online performance.

**Disturbers** are mainly characterized by negative input, such as provocative comments and bad evaluations about other ideas. In doing so, they cause damage in content and attempt to develop conflict within the community. However, our research has shown that this is usually not happening for simple reasons to hijack the community and spread harm within the social collaboration platform (the number of cases of competition hijacking the community to damage the brand or political organization is very low with under 5% of the disturber cases in general). In fact those personas usually cast a negative shadow over other ideas (worse evaluation) to spotlight their own contributions for extrinsic reasons (in many analyzed cases they even asked other community members to down-vote ideas, so their average score is better and gets more attention from the initiator). Moreover, this phenomenon can be detected mainly on external public communities (especially from political parties) and sometimes bigger brands but is nearly non-existing on internal platforms. The explanation is simple as no individual would risk his job by neglecting the netiquette and wider purpose of the initiative of his or her employer and leave a negative image.

**Idea-Thieves** who are driven by an egoistic motivation are rare. They are characterized by copying existing ideas and simply describe them in other words or 'restructure' them but not really improving or develop them any further. This means that they try to find the easiest way to receive incentives or other extrinsic elements (contacts to brands etc.) and frustrate the actual idea provider. This is a double-edged sword as people can also do it unwillingly. Especially within internal communities from organizations or

companies, people post ideas which have already been submitted without knowing as no research on other ideas have been made so far (such would not be tolerated internally and is easily trackable). It often ends in a dispute on who was first and community managers must de-escalate. Thieves are mainly active on external communities in the private sector, while stealing ideas on political and socially driven platforms would not make much sense as extrinsic incentives are usually not provided as well as manifold opinions are more emotionally driven and therefore not an interesting subject to copy.

**Explorers** join a community because they are interested in the topic and would like to learn from the community. Even if they do not bring much expertise, they try to get involved partially (more through evaluations and comments than ideas). On internal platforms, this role is typically taken by interns, working students and young professionals to broaden their horizons or to find interesting topics they can identify and maybe later work on. On external platforms (especially industry), many students, trainees or apprentices register and take their chances to connect with their favorite or interesting brands, to get in touch with experts from that brand and to learn more about its strategy and purpose. Public communities are usually not a place for those kind of explorers, as interested and younger people tend to engage more actively and share their ideas and opinions and follow a different kind of intrinsic motivation (in this case more power ideator, supporter or passive member).

**Passive Members** register in the community and are highly interested in the topic, yet they do not share their own input or participate in discussions. Rather, they let themselves be inspired by the content and comments of others or try to learn something new by just reading them. As this is still connected to a semi-active part (registering, logging in again and reading of contributions) they still invest plenty of time. This applies especially for communities focusing on emerging technologies, where a limited amount of knowledge has yet been gathered in general and major learnings can be achieved through connecting with the experts on the community. In parallel, political and social driven communities spread opinions, provide information and exchange comments with a local, federal or national perceived importance for the individual. The invested time is shared by participants from internal communities (industry and public alike), however, daily work is omnipresent and hinders employees to participate even if it is just passive. Hereby, their frequency is usually a bit lower than on external innovation communities. The tendency shows that whenever employees find the time to register and log-in, they usually try to engage actively. On the contrary to external initiatives they also cannot expect insights and knowledge from outside of the organizational boundaries, which is one of the most important reasons for external members to participate by just reading completely different perspectives from a crowd across the globe.

**Lurkers** are usually the vast majority of internal and external platforms due to time limitations, fading interest due to the goals and purpose of the initiator, social dynamics, unsatisfying feedback, other work etc. Most of the visitors of the platform stay lurkers and do not become active. Now, one could state that non-active members are not a part of the community, which is only partially correct. However, a substantial number of quantitative traffic (clicks) are coming from users, who do not contribute in any way and who often just visit the platform once or register and visit the platform rarely without taking any actions. Those statistics are used to market and back the success of

the initiative (internal: in front of the board to get more funding for further and similar projects; or external: to show how many people have been included in a public decision process/strategy as well as to show client centricity and relationship) (Table 3).

**Table 3.** Identified community characters clustered by tendencies of motivation and engagement

|                      | Active  | Passive                              |
|----------------------|---|--------------------------------------|
| Intrinsic Motivation | Community manager<br>Opinion leader<br>Power ideator<br>Supporter<br>Expert | Explorer<br>Passive member<br>Lurker |
| Extrinsic motivation | Power ideator<br>Profit-oriented<br>Disturber<br>Idea thief                 |                                      |

The probability of occurrence of the identified characteristics, varies in the different sectors (public vs. industry). In addition, the frequency of the characters is very different between internal and external platforms. Based on our experience, we give an estimate below (scale from 0 to 3) of how likely it is that the respective character will be encountered on the various platforms. The probabilities of occurrence are dummy variables coded in the following way: 0 means that these characters are not encountered, 1 means rare occurrences, 2 sometimes present and 3 that these characters occur (Table 4).

## 5 Managerial Conclusions and Further Research

In each community, different characters interact with various motivation drivers, intentions, interests and thus different engagements. In the public sector, many participants have an actual interest in the development of certain topics and thus act intrinsically, open innovation participants in the industry sector are more extrinsically driven. A community is only as good as its management. To reach the most constructive discussions and qualitative output, it is necessary to manage these different characters on the platform and deal with their wide-range of characteristics.

In this research, we have shown our analysis and findings about the different roles and behaviors of personas on internal and external innovation communities initiated by both - public and industry - players. Even more, we have shown differences between those patterns and projects as well as highlighted the primary motivation triggers for each persona. Those findings should be of major value for managers and decision makers in the public and private field, driving innovation communities in the present and the future, as many conclusions and derivations can thrive from it. Furthermore, we believe this paper should be viewed as a starting point for going beyond the qualitative data and backing the findings with further quantitative insights to answer correlation questions between

**Table 4.** Overview of community characters and their occurrence

|                   | Internal Platforms |                 | External Platforms |                 |
|-------------------|--------------------|-----------------|--------------------|-----------------|
|                   | Public Sector      | Industry Sector | Public Sector      | Industry Sector |
| Community manager | 3                  | 3               | 3                  | 3               |
| Opinion leader    | 1                  | 1               | 2                  | 1               |
| Power ideator     | 1                  | 2               | 3                  | 3               |
| Supporter         | 1                  | 1               | 3                  | 2               |
| Expert            | 1                  | 1               | 3                  | 3               |
| Profit-oriented   | 1                  | 2               | 2                  | 3               |
| Disturber         | 0                  | 0               | 3                  | 2               |
| Idea thief        | 1                  | 1               | 1                  | 2               |
| Explorer          | 2                  | 2               | 1                  | 3               |
| Passive member    | 2                  | 2               | 3                  | 3               |
| Lurker            | 3                  | 3               | 3                  | 3               |

users and different types of initiated innovation communities. Of interest is also a deeper evaluation of the current development of concepts and processes of communities which was stated as one of the reasons for the declining participation in general at the beginning of this paper. Formats like overnight feedback, sprints or digital community twins are spreading but are rarely analyzed. An open question is also how communities should evolve technologically, thinking about the users' demands, the continuous change in features as well as in emerging and possible enhancements.

Finally, we also see a major benefit in transferring the detected insights into guidelines for practitioners on how to handle and successfully enhance existing and future communities (practical how-to manual with managerial implications). Consequently, and as a final outlook, we would like to start by giving a quick overview and sneak-peak on the latter, meaning that we would highlight some selected managerial implications from our findings.

**Get inspired Outside of Your Own Sphere:** We have seen that users' behavior differs not just between public and industry led initiatives, but also strongly between internal and external communities. So far, many can say that is no surprise. However, we give proof that really all defined personas are existing on each platform, just excluding the role of the disturber on internal initiatives. Prior hypotheses might have headed in the opposite direction, suggesting that especially between public and industry different roles and personas might exist. Managers should not just investigate similar innovation communities from the competition or potential industry partners, but really look beyond the industries, including the public sector. In the scope of co-creation and innovation ecosystems, meeting with potential cooperation partners or, simply a sparring partner would be helpful prior to the kick-off of a project to learn from others' mistakes and gain

important insights. Responsible managers might raise the question, why other companies or organizations should meet up and help with their time when they cannot deliver anything in return. The answer is rather easy: share your insights after the project as well, so they can improve their already existing platform and its management. Thinking about doing this with multiple partners also for other problems or questions gives a perfect starting point for a cross industrial network (or innovation community) for co-creation in an open innovation ecosystem. A shared community initiative with joined forces and multiple questions for the customers/citizens is also an option (e.g. in the smart city context municipality areas and local business can work close together).

**Individualize Your Community Management:** Managers nowadays are aware of heterogeneity as well as social dynamics on online communities and consequently the efforts they must master in order to be successful. However, until today, still many underestimate how much work has and should go into a professional community management. Surely, you can ‘survive’ and go through an initial phase with limited resources, but every hour invested in the community means a reward in the form of more users, more ideas, more dynamics and a more healthy and vivid social community. Here, be aware that as our research has shown, different roles and personas are active. And like raising or treating your children, it is important to understand that users have different characters and should be managed in different and customized ways. As an example, we identified the issue of many internal driven communities dealing with ‘stealing’ ideas. Make sure your community management is de-escalating and bring the solution from the online to the offline world. Not every thief is one on purpose as stated above and you do not want to lose potential ideators; track the timing on the exact submission through the platform, ask if both might share and link the idea so they can work further in a team. If such cooperation is not desired, make sure the second in line (timewise) should have a clear improvement on the existing status. With our matrix shown in this paper you can already anticipate roughly what kind of personas are to be expected as well as in which ratio they usually appear (internal vs. extern; public vs. industry). Your incentive strategy (e.g. intrinsic vs. extrinsic), recruiting channels (e.g. online vs. offline), IP and legal strategy (e.g. exchange ideas for prizes vs. annex everything), evaluation process (e.g. pitching favorite ideas vs. jury voting) etc. should be adapted accordingly. Just to pick out one of those mentioned: if you see that many supporters are existing or more importantly missing on your platform, adjust your incentive structure. It is also important to ensure to incentivize valuable comments or enhancements of existing ideas through a community award or give away prizes for certain groups (e.g. student awards or internships if many of those personas are registered).

**Adapt Your Technological Platform:** Communities are technologically driven, but the importance of technology is reduced by the rising importance of social management and dynamics over time. However, evolving technology through the years and community initiatives can help to foster better ideas and attract different players and users. Many companies and organizations buy either out of the box solutions or on the contrary over engineered (usually by customizing) their platform without a sustainable effect. Make sure which features you really need for the expected fellowship on the community. Our matrix of personas on the different community levels should come in handy



to preselect the right features. For instance, make sure when including your employees on an internal platform that solutions are prioritized where people can also work on the way back to their homes (most participants with remote access did it during the public transportation time). Technology is only successful when the management behind is taking the appropriate decisions. So, make sure to guarantee free time for participating in internal platforms or combine it with gamification approaches like design thinking workshops or credits on the platform for good ideas which can be exchanged for spare time, lunch meals, etc. In addition, make sure to keep moving forward with your platform and do not stop the efforts even though the pilot project was not 100% satisfying. Alongside the evolvement of your platform comes the understanding of its management - and that takes time. With time, also more questions like internal policies, stakeholder management and KPIs will arise. Here, managerial implications of the data above can help to avoid mistakes and wrong directions - one reason more for us to publish further research on this topic.

## References

1. Chesbrough, H., Brunswicker, S.: Managing open innovation in large firms. In: Survey Report UC Berkeley & Fraunhofer Institute for Industrial Engineering, Stuttgart (2013)
2. Schmidhuber, L., Hilgers, D., Rapp, M.: Political innovation, digitalisation and public participation in party politics. *Policy Polit.* **47**(3) (2019)
3. Saebi, T., Foss, N.J.: Business models for open innovation: Matching heterogeneous open innovation strategies with business model dimensions. *Eur. Manag. J.* **33**(3), 201–213 (2015)
4. Kohler, T., Nickel, M.: Crowdsourcing business models that last. *J. Bus. Strategy* **38**(2), 25–32 (2017)
5. Yang, K.: Research on factors affecting solvers' participation time in online crowdsourcing contests. *Future Internet* **11**(8), 176 (2019)
6. Simula, H.: The rise and fall of crowdsourcing? In: 2013 46th Hawaii International Conference on System Sciences, pp. 2783–2791. IEEE, January 2013
7. Koch, G., Füller, J., Brunswicker, S.: Online crowdsourcing in the public sector: how to design open government platforms. In: Ozok, A.A., Zaphiris, P. (eds.) OCSC 2011. LNCS, vol. 6778, pp. 203–212. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-21796-8\\_22](https://doi.org/10.1007/978-3-642-21796-8_22)
8. Shen, X.L., Lee, M.K., Cheung, C.M.: Exploring online social behavior in crowdsourcing communities: a relationship management perspective. *Comput. Hum. Behav.* **40**, 144–151 (2014)
9. Hutter, K., Hautz, J., Füller, J., Mueller, J., Matzler, K.: Communitition: the tension between competition and collaboration in community-based design contests. *Creativity Innov. Manag.* **20**(1), 3–21 (2011)
10. Füller, J., et al.: Consumer empowerment through internet-based co-creation. *J. Manag. Inf. Syst.* **26**(3), 71–102 (2009)
11. Brodie, R.J., et al.: Consumer engagement in a virtual brand community: an exploratory analysis. *J. Bus. Res.* **66**(1), 105–114 (2013)
12. Sæbø, Ø., Rose, J., Flak, L.S.: The shape of eParticipation: characterizing an emerging research area. *Government Inf. Q.* **25**(3), 400–428 (2008)
13. Lee, G., Kwak, Y.H.: An open government maturity model for social media-based public engagement. *Government Inf. Q.* **29**(4), 492–503 (2012)

14. Albers, A., Maul, L., Bursac, N.: Internal innovation communities from a user's perspective: how to foster motivation for participation. In: Abramovici, M., Stark, R. (eds.) *Smart Product Engineering*, pp. 525–534. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-30817-8\\_51](https://doi.org/10.1007/978-3-642-30817-8_51)
15. Edelman, N.: Lurkers as actors in online political communication. *Comunicazione Politica: Il perimetro della democrazia in rete. A che punto è il dibattito* (2012)
16. Lakhani, K.R., Wolf, R.G.: *Why hackers do what they do: understanding motivation and effort in free/open source software projects* (2003)
17. Bilgram, V., Bartl, M., Biel, S.: Getting closer to the consumer—how Nivea co-creates new products. *Market. Rev. St. Gallen* **28**(1), 34–40 (2011)
18. Füller, J., Hutter, K., Faullant, R.: Why co-creation experience matters? Creative experience and its impact on the quantity and quality of creative contributions. *R&D Manag.* **41**(3), 259–273 (2011)
19. Deci, E.L., Ryan, R.M. (eds.): *Handbook of Self-Determination Research*. University Rochester Press (2004)
20. Brabham, D.C.: Moving the crowd at Threadless: motivations for participation in a crowdsourcing application. *Inf. Commun. Soc.* **13**(8), 1122–1145 (2010)
21. Acar, O.A.: Motivations and solution appropriateness in crowdsourcing challenges for innovation. *Res. Policy* **48**(8), 103716 (2019)
22. Hutter, K., Füller, J., Koch, G.: Why citizens engage in open government platforms? In: *Lecture Notes in Informatics (LNI)-Proceedings*, vol. 192 (2011)
23. Koch, G., Rapp, M., Hilgers, D.: Open Innovation für Parteien—Wie politische Parteien von neuen Formen der Mitglieder-und Bürgerpartizipation profitieren können Internet und Partizipation, pp. 203–222. Springer, Wiesbaden (2014). <https://doi.org/10.1007/978-3-658-01028-7>
24. Mergel, I.: Opening government: designing open innovation processes to collaborate with external problem solvers. *Soc. Sci. Comput. Rev.* **33**(5), 599–612 (2015)
25. Dholakia, U.M., Bagozzi, R.P., Pearo, L.K.: A social influence model of consumer participation in network-and small-group-based virtual communities. *Int. J. Res. Market.* **21**(3), 241–263 (2004)
26. Pruitt, J., Grudin, J.: Personas: practice and theory. In: *Proceedings of the 2003 Conference on Designing for User Experiences* (2003)
27. Friedman, B., Kahn, P.H., Borning, A., Huldtgren, A.: Value sensitive design and information systems. In: Doorn, N., Schuurbiers, D., van de Poel, I., Gorman, M.E. (eds.) *Early engagement and new technologies: Opening up the laboratory. PET*, vol. 16, pp. 55–95. Springer, Dordrecht (2013). [https://doi.org/10.1007/978-94-007-7844-3\\_4](https://doi.org/10.1007/978-94-007-7844-3_4)
28. Gebauer, J., Füller, J., Pezzeri, R.: The dark and the bright side of co-creation: triggers of member behavior in online innovation communities. *J. Bus. Res.* **66**(9), 1516–1527 (2013)
29. Baskerville, R.L., Wood-Harper, A.T.: A critical perspective on action research as a method for information systems research. *J. Inf. Technol.* **11**(3), 235–246 (1996)
30. Yin, R.K.: *Case Study Research: Design and Methods*. Sage Publications (2009)



# Qualitative Evaluation of the Usability of a Web-Based Survey Tool to Assess Reading Comprehension and Metacognitive Strategies of University Students

Luis A. Rojas P.<sup>1</sup>, Maria Elena Truyol<sup>2</sup>(✉), Juan Felipe Calderon Maureira<sup>3</sup>, Mayron Orellana Quiñones<sup>2</sup>, and Aníbal Puente<sup>3</sup>

<sup>1</sup> Universidad Central de Chile, Santiago, Chile  
lrojas.larp@gmail.com

<sup>2</sup> Universidad Andres Bello, Santiago, Chile  
maria.truyol@unab.cl

<sup>3</sup> Universidad Andres Bello, Viña del Mar, Chile

**Abstract.** Survey tools are the main support for the development of reading comprehension evaluation instruments. The functionalities provided in the questionnaires design are the fundamental pillar to elaborate studies to evaluate in depth the reading comprehension and metacognitive strategies used. Survey tools are the main support for the development of reading comprehension evaluation instruments. Questionnaires design in available survey tools are based primarily on open, closed, multiple choice, or mixed questions. The functionalities provided in the questionnaires design are the fundamental pillar to elaborate studies to evaluate in depth the reading comprehension and metacognitive strategies used. However, current survey tools do not provide the required functionalities (i.e., types of questions) to design studies/questionnaires that allow the development of reading tasks that examine the students' strategic potential.

**Keywords:** Usability · Qualitative evaluation · Survey tool · Reading comprehension

## 1 Introduction

To know the level of reading comprehension of the pre-university students is a relevant information to help them to be successful in their university learning process. This information can be obtained with the help of proper evaluation instruments. Even more efficient would be to be able to obtain information on students' reading comprehension skills, systematize and manage that information for large numbers of students. This could be done with a web-based survey tool designed for such purposes. This instrument should be able to capture superior cognitive processes and to allow an objective evaluation, all of this in an environment of easy interaction for the student.

Reading is a central part of the learning process and becomes more important as learners progress from “learning to read” to “reading to learn”. The central objective of

reading at university levels is to integrate the information of the text, even if it is not complete, with the knowledge provided by the reader, in order to interpret the global and deep meaning. These require not only reading strategies but also cognitive processes highly related to metacognition that are generally not fully developed in students of the first semesters of university. Survey tools are the main support for the development of reading comprehension evaluation instruments. Modern survey tools include at least three main components: questionnaire design, distribution, and reporting. In general, the questionnaires design in available survey tools are based primarily on open, closed, multiple choice, or mixed questions. The functionalities provided in the questionnaires design are the fundamental pillar to elaborate studies to evaluate in depth the reading comprehension and metacognitive strategies used. However, current survey tools do not provide the required functionalities (i.e., types of questions) to design studies/questionnaires that allow the development of reading tasks that examine the students' strategic potential. It is necessary then the construction of new functionalities that allow the implementation of these cognitive processes.

Usability plays an important role for the success of any educational software. In particular, and especially, any survey tools. This is so important because if a survey tool is not usable enough, it obstructs the main objective of the designed measuring instrument. That is, students would spend more time learning how to use the tool instead of concentrating on the tasks proposed to assess reading comprehension. For that, educational researchers should not devalue usability testing of these instruments.

There are several tests that quantify how end-users accept technology but there is a lack of detailed findings, with a holistic view of that experience. Also, educational software has characteristics that are usually left out of usability measurement instruments. In this way, qualitative research can help to obtain detailed information related to the usability of the survey tool, for educational use, and identify specific improvement opportunities in the interaction of the users (students) with its tasks.

The main objective of this study was to evaluate the usability of a new web-based survey tool designed for the assessment of reading comprehension and metacognitive strategies in university students, using a qualitative setting. For that, the following specific objectives were stated:

- to explore user interaction with the different functionalities.
- to study their perceived benefits from the use of the different functionalities.
- to formulate suggestions for improvement of the web-based survey tool.

## 1.1 Research on Usability

For successful in a survey tool application, usability has an important role as a foundational and desirable characteristic. To obtain a more precise and objective measurements with a survey tool, user interface should be designed with focus on an ease of use and absence of cognitive obstructions [1].

In usability design and evaluation, it is important to evidence the occurrence of some indicators related to good attributes. Some trending attributes in usability research are related to easy to learn, efficient to use, easy to remember, few errors and pleasant to use [2], learnability, user satisfaction, ease of use [3], user experience, engagement,

functionality [4], among others. Related to these attributes, there are several instruments to evaluate the indicators associated to them. These instruments are focused primarily on a quantitative measurement of certain indicators from an isolated, structured and observational perspective. Nevertheless, it is impossible to establish some relationships between indicators, or to obtain more detailed findings [5].

To obtain detailed usability information and improvements opportunities of a system a holistic perspective is needed. This perspective can be provided by qualitative research. A well-known qualitative instrument is the interview [6], the semi-structured approach [7], which provide versatility and flexibility. Semi-structured interviews provide lines of enquiry to be pursued within the interview, to follow up on interesting and unexpected avenues that emerge [8].

To provide a specific guide to construct a semi-structured interview as a usability evaluation instrument, some hints must be obtained from valid usability tests. As mentioned before, trends in usability evaluation presents as common categories related to learnability and ease of use, how user is satisfied, and other specific associated to functionality.

The literature provides some examples of usability tests formulated to evaluate general systems. An example is the usability tests formulated to evaluate a web-based tool, focused on cognitive development [9]. The instrument of usability evaluation was formulated according to several phrases and sentences, with a positive or negative answer. These sentences are specific to this tool, and are focused on learnability, functional complexity, reliability and handling, among others. On the other hand, the work of Parshall & Harmes [10] provides a checklist about usability attributes for computer programs. Regarding to ease of use, satisfaction and usefulness, the works of [11, 12], provide specific questionnaires to evaluate these categories. Particularly, David's works presents a perspective based on sentences focused on perception of users, with emphasis in the benefit or disadvantage for respective subject, with a dichotomous answer. Lin, Choong, & Salvendy [13] focused on formulate a universal scale of usability, proposes a methodology to compare usability of different software system, with the instrument Purdue Usability Testing Questionnaire. This instrument identifies eight human factors in software usability: compatibility, consistency, flexibility, learnability, minimal action, minimal memory load, perceptual limitation and user guidance.

Review of the literature suggests that some usability tests can be used as a framework to qualitatively explore users' experiences. A semi-structured interview must be constructed to specifically obtain information about user interaction with our system of interest: a web-based survey tool to evaluate reading comprehension and metacognitive strategies.

The main objective of this study was to evaluate the usability of a new web-based survey tool designed for the assessment of reading comprehension and metacognitive strategies in university students, using a qualitative setting. For that, the following specific objectives were stated:

- to explore user interaction with the different functionalities.
- to study their perceived benefits from the use of the different functionalities.
- to formulate suggestions for improvement of the web-based survey tool.

## 1.2 Measurement Instruments on Reading Comprehension

In recent decades there has been a notable growth in the need to improve reading performance standards both in developed and developing countries. The influence of the PISA (Program for International Students Assessment) model promoted by the OECD [14] was being the engine that develops and activates the new reading tools [15]. In the English-speaking language, the study of reading performance has a long history. In contrast, in the context of Spanish speech reading comprehension, instruments are newer and focused on specific reading aspects, such as semantic representation and previous knowledge [16].

In the literature some evaluation instruments can be found with an adequate psychometric quality. For example, PROLEC-R [17] is a set of evaluations aimed to detect reading difficulties. ESCOLA (The Reading Consciousness Scale) evaluates metacognition and executive functions [18]. EDILEC [19] dynamically and automatically evaluates reading skills for secondary school students. In the case of educational Chilean market, Dialect [20] is a platform for the timely diagnosis of reading skills, with evaluation of reading comprehension for primary students with a good grade of validity. Other tests are focused on reading speed, related with online reading courses (e.g., [21]).

Some of the previous mentioned tests have a digital or an online version. These versions have some features, such as immediate evaluation, automated feedback, strengths and weaknesses of the students, and pedagogical recommendations (e.g., Dialect [20]). Regarding to the variety of items in traditional tests, closed and open items are used to get information about reading comprehension. Open items (such as word completion or essay-faced items) could be complex to grade without an exact and objective answer or a very accurate evaluation rubric. Nevertheless, in digital and automated tests is complex to evaluate open answers, emphasizing the use of closed items for a fast an objective grading process.

Digital platforms provide a wide variety of media inputs and output, such as touch-screens, shared screens, audio support, etc. However, they are sub-utilized. Multiple-choice items are used to facilitate grading and stimulate discrimination capability in answer selection [22]. In this line, dynamics textual organizers such as conceptual maps, word clouds, etc., could be used for reading comprehension tasks that can be evaluated in an objective way [23]. These kinds of task could be incorporated in digital platforms, to take advantage of the cognitive processes involved in them, but it is not used in the mentioned tests. To evaluate reading comprehension and metacognitive strategies with a web-based survey tool, some special functionalities are required. In order to incorporate superior cognitive processes and allow an objective evaluation, the functionalities must enable to:

- organize and represent knowledge in a meaningful way according to the selected topic.
- use dynamic modes of text selection.
- to take advantage of visual hierarchy of elements.

## 2 Description of the Functionalities of a Web-Based Survey Tool to Assess Reading Comprehension and Metacognitive Strategies

According to the proposed objectives, a web-based survey tool is developed for the specific purposes to this study. A set of Spanish texts from a wide range of themes were selected. Each text has items to evaluate reading comprehension and metacognitive strategies. As mentioned before, in order to incorporate superior cognitive processes and allow an objective evaluation, the functionalities must have specific characteristics that go beyond the traditional multiple-choice surveys. These special tasks are formulated through the following functionalities designed:

- *Conceptual Map (CM)*: users must build a conceptual map from a closed list of words, using a combo-box item for each word to be completed. The structure of conceptual map was previously defined, with the corresponding grammar connectors and reading direction (See Fig. 1).

Berko, J. (2010). Desarrollo del lenguaje. Una revisión y una vista preliminar. En Jean BerkoGleason y NanBernsteinRatner, El desarrollo del lenguaje. Madrid: Pearson.

Lea el texto que se presenta a continuación y luego realice las siguientes actividades.

Se ha demostrado que algunos insectos, como las abejas, tienen elaborados sistemas de comunicación. El etólogo Karl von Frisch (1950) empezó a estudiar las abejas en la década de los años veinte y ganó el premio Nobel en 1973 por sus estudios de la comunicación entre estos insectos tan sociales. A diferencia del maullar expresivo de un gato hambriento, en muchos sentidos el sistema de comunicación de las abejas es referencial: comunica a otras abejas algo sobre el mundo externo. Una abeja que regresa a la colmena tras haber encontrado flores llenas de néctar agrupa una audiencia y a continuación realiza una danza que indica la dirección y la distancia aproximada del néctar con respecto a la colmena. Otras abejas miran, se

1.- Ubique las siguientes palabras en el mapa conceptual para representar la idea principal del texto.

```

graph TD
    A[abejas] -- poseen --> B[comunicación]
    B -- de --> C[Seleccione]
    C -- del tipo --> D[Seleccione]
    D -- informa --> E[Seleccione]
    E -- del --> F[Seleccione]
    
```

**Fig. 1.** Screenshot Conceptual Map (CM) functionality.

- *Explanatory Concepts Selection (ECS)*: users must select some short sentences or words sets from a paragraph, which explain a proposed affirmation. Users drag the mouse to select the corresponding words and then they must confirm the selection using a specific button (See Fig. 2).

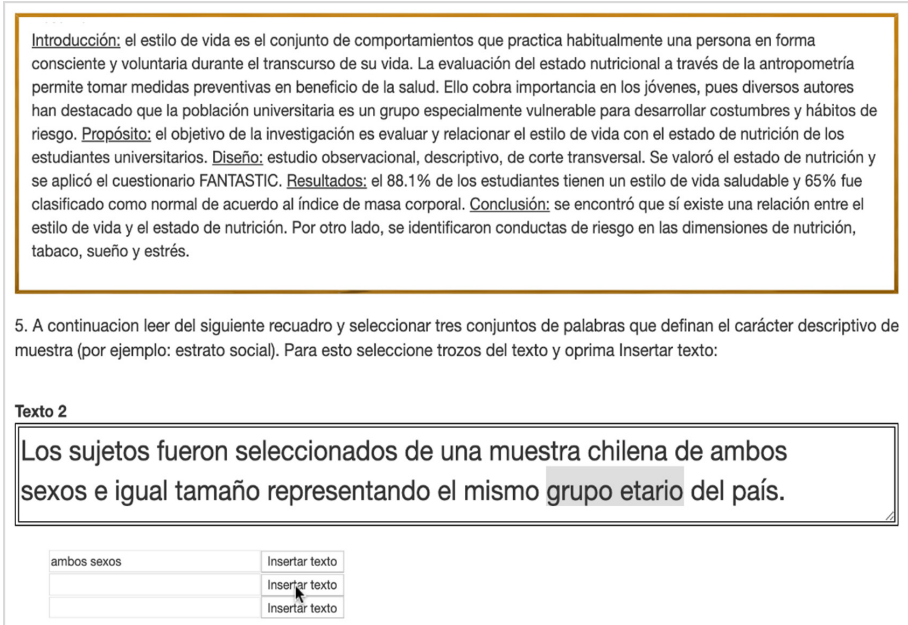


Fig. 2. Screenshot Explanatory Concepts Selection (ECS) functionality.

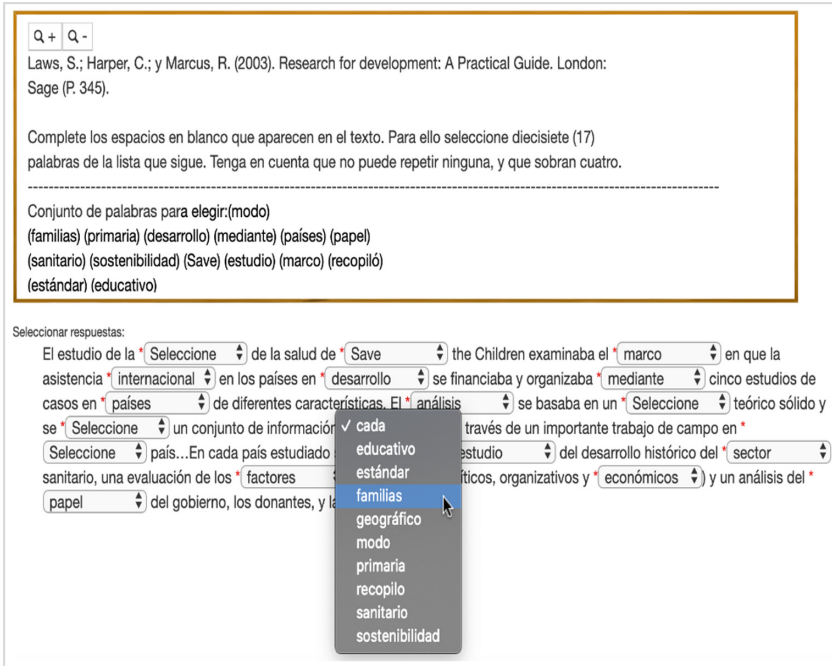
- *Sentence Completion (SC)*: users must complete sentences or paragraphs with some hidden words, following a pattern. A closed list of words was previously defined to fill in the blanks, without repetition. A combo-box item for each hidden word is used to do the selection (See Fig. 3).
- *Hierarchical Sentences Sorting (HSS)*: used to evaluate metacognitive strategies, users face a reading situation. Then they must sort several sentences with specific reading strategies, according to a relevance-prioritization sorting. Users must drag-and-drop each sentence, placing in upper positions the sentence with the highest priority or relevance (See Fig. 4).

### 3 Materials and Methods

This work studied the usability of a new web-based survey tool for the assessment of reading comprehension and metacognitive strategies in university students through semi-structured interviews.

To study usability a useful method is user testing. It was intended to include participants of early university courses to work with a similar profile of people that is the target



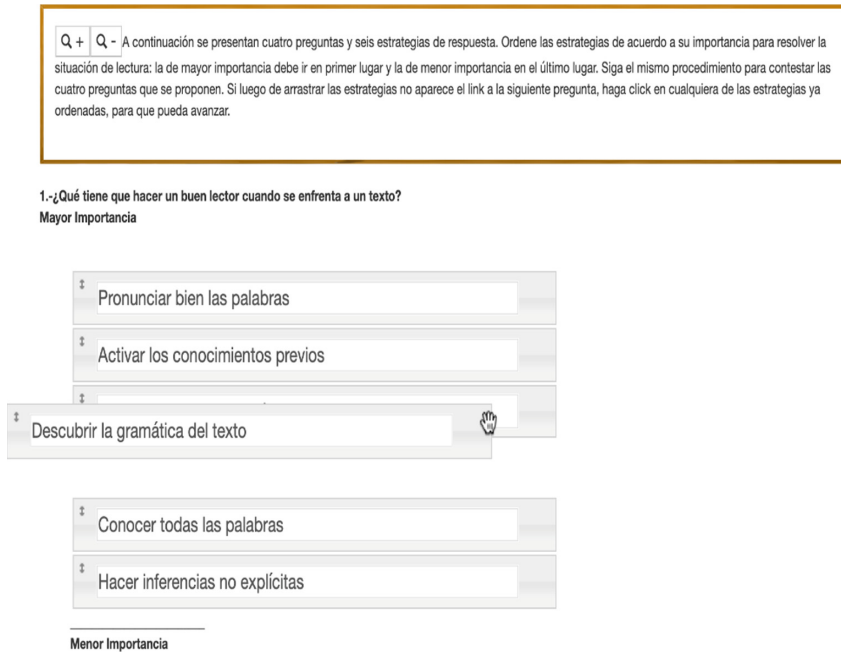


**Fig. 3.** Screenshot Sentence Completion (SC) functionality.

of the designed tool. The inclusion criteria were university students, minimum 19 years old. No special experience in technology was asked. The participants were recruited via massive email with help of professors of basic university courses, on a large private university on Chile. Interviews were scheduled between June and July 2019. They were done in a classroom, with all the technical resources needed and following the same protocol: brief presentation of the research objectives; explanation and signature of participant’s consent; explanation of the task sheet; participant’s task fulfilment; qualitative interview and SUS questionnaire at the end of the session.

Different actions were carried out to ensure the validity of our qualitative study. As we mentioned before, we obtained a signed consent to use and release the data of the interviews and any other instrument used to obtain data for this study. Participants were assured that the use of the information obtained would be carried out without any connection with their person.

Semi-structured interviews provided participants the privacy to interact with the functionalities of the web-based survey tool and the freedom to elaborate their ideas during the interview. The semi-structured interview guide was constructed based on some usability tests: Usability [9]; Usability checklist for computer-based testing programs [10, 24]; Usefulness, Satisfaction, and Ease of use [11]; Purdue usability [13]; Perceived usefulness and perceived ease of use [12]. Questions were adapted specially for this study. The interview guide included core questions and clarification/elaboration questions,



**Fig. 4.** Screenshot Hierarchical Sentences Sorting (HSS) functionality.

which enabled participants to reflect on all the functionalities during the interview. For details, see Table 1.

The interviews were digitally recorded, transcribed and analysed by three persons independently, following a thematic analysis. Thematic analysis is a qualitative analytic method that it is aimed to identifying, analysing and reporting themes or patterns within all the data. It helps to organizes de data set in some detail, but it often goes further: it helps to interpret some important aspects of the research topic [25]. A theme represents an important aspect about the data, related to the research question. It represents some pattern or meaning within the data that it is worthy to identify, code and analyse. In this work, the coding was done without a previous definition of them, but with an underlying theory, usability, that oriented the coding of the data to that feature. This resulted in several codes not only about usability, but also about user experience [26], expanding the limits and interests of the study. Initial coding was performed by two researchers to ensure consistency and establish the definition of the set of codes. A third researcher was used to check the code definitions.

In the next section, the results are presented based on the defined codes and the themes that organize them, exploring how users interacted with the functionalities designed for the web-based survey tool. Suggestions for improvement and perceived benefits are presented.

**Table 1.** Interview guide.

| Question   | Elaboration question   | Tests   |
|--|--|---------|
| Do the functionalities of the web application satisfy you completely?  | Which of the functionalities did you find satisfactory? Why? Which of the functionalities was not satisfactory for you? Why? At some point you didn't know what to do? In what functionality? What was the difficulty you had? | [11–13] |
| Would you use the web application again for similar tasks?             | What functionalities would you use again? Why? Which ones would you not use again? Why?  | [11–13] |
| Does the web application work the way you wanted?                      | What functionality worked as you expected? Why? Which functionality did not work as you expected? Why?   | [11–13] |
| Were you able to complete the nodes of the concept map?                | At some point you did not know how to use this functionality? Why? Do you feel that the functionality distracted you from the specific task of identifying concepts? Why?  | [24]    |
| Were you able to complete the proposed texts with the suggested words? | At some time, you did not know how to use this functionality? Why? Do you feel that the functionality distracted you from the specific task of selecting the suggested words? Why?   | [24]    |
| Were you able to select the descriptive word group?                    | At some time, you did not know how to use this functionality? Why? Do you feel that the functionality distracted you from the specific task of identifying the word group? Why?  | [24]    |
| Were you able to perform the hierarchical order requested?             | At some time, you did not know how to use this functionality? Why? Do you feel that the functionality distracted you from the specific task of ordering hierarchically? Why?   | [24]    |
| Did you find the design pleasant and functional?                       | What elements of the design (for example: colours, buttons, cursors, font size, available windows, etc.) did you find most effective and suitable for the tasks to be performed?   | [9, 24] |

*(continued)*

**Table 1.** (continued)

| Question   | Elaboration question   | Tests          |
|--|--|----------------|
| Was it easy for you to learn how to use the different functionalities?     | In which task were the functionalities more difficult/easy to learn? What was complicated? What difficulties did you find to learn how to perform the task of CM/ECS/HSS/SC? | [9, 11–13]     |
| Did you find it quick to learn how to use the different functionalities?   | In which task were the functionalities faster to learn?  | [9, 11–13]     |
| Did the functionalities allow you to correct mistakes easily?              | What difficulties did you find to correct errors in the task of CM/ECS/HSS/SC?   | [9, 11–13, 24] |
| Did the functionalities allow you to check before finishing the task?      | What difficulties did you find to check before completing the task in the CM/ECS/HSS/SC task?  | [24]           |
| Did you find any inconsistency that prevents progress in tasks?            | Do all commands work properly? Do you always consider having controlled the functioning of the functionalities? Was external assistance necessary to solve any problems?     | [11–13, 24]    |
| Was it easy for you to understand the information presented on the screen? | Was the information presented enough? Was it necessary to request external assistance? In some task/s the information on the functionalities were insufficient? In which?    | [9, 24]        |
| Did you find the use of the functionalities simple?                        | Could you properly use the mouse to complete the tasks? Was the navigation through the web application friendly?   | [9, 11–13]     |

## 4 Results

We interviewed 14 participants, in an age bracket between 19 and 23 years. Interviews were completed in 25 min on average.

During the analysis process some codes were proposed to organize the information. As this process progressed, the codes were refined and showed that the level of detail reached by the interviewees in their responses allowed not only to obtain information about usability but also about their user experience. For that, we decided to work with the seven factors that describe user experience, according to Morville [26]: Useful, Usable, Findable, Credible, Desirable, Accessible and Valuable. Furthermore, we included two additional factors that were considered central for a web-based survey tool designed for student's assessment: Learnability and Error prevention/recovery. Moreover, it was possible to associate positive, negative and neutral value to the codings based on the detailed explanations provided by the participants (Positive/Neutral/Negative). We differentiate

the users' explanations between the different functionalities tested: Explanatory Concepts Selection (ECS), Conceptual Map (CM), Sentence Completion (SC), Hierarchical Sentences Sorting (HSS). We also coded the explanations to differentiate whether the user comment is about the functionality or the task the functionality is used for (Functionality/Task). We included a category that we called General to include all the explanations that were done about the complete set of functionalities, without specific mention of one.

This coding scheme can be organized for clarity into five themes named as "Easy of Use", "Appreciation" (as the recognition of the good qualities of someone or something), "Satisfaction" (as the fulfilment of one's wishes, expectations, or needs, and the pleasure derived from this), "Learnability" and "Error Prevention/Recovery". The codes were grouped by themes as follows:

- Easy of use: *Usable*, *Findable* and *Accessible*.
- Appreciation: *Useful* and *Valuable*.
- Satisfaction: *Credible* and *Desirable*.
- Learnability: *Learnability*.
- Error Prevention/Recovery: *Error Prevention/Recovery*.

#### 4.1 General Descriptive Analysis

Analysing the codification of all the participant's mentions versus the value assigned and sorting out by functionalities, Fig. 5 shows that *Explanatory Concepts Selection (ECS)* functionality is the one with the higher quantity of negative comments, followed by the *Conceptual Map (CM)*. Approximately 42% of the negative comments were for *ECS* and of all the mentions for *ECS*, about 72% were negatives. The other two functionalities and the *General* category received more positive than negative comments.

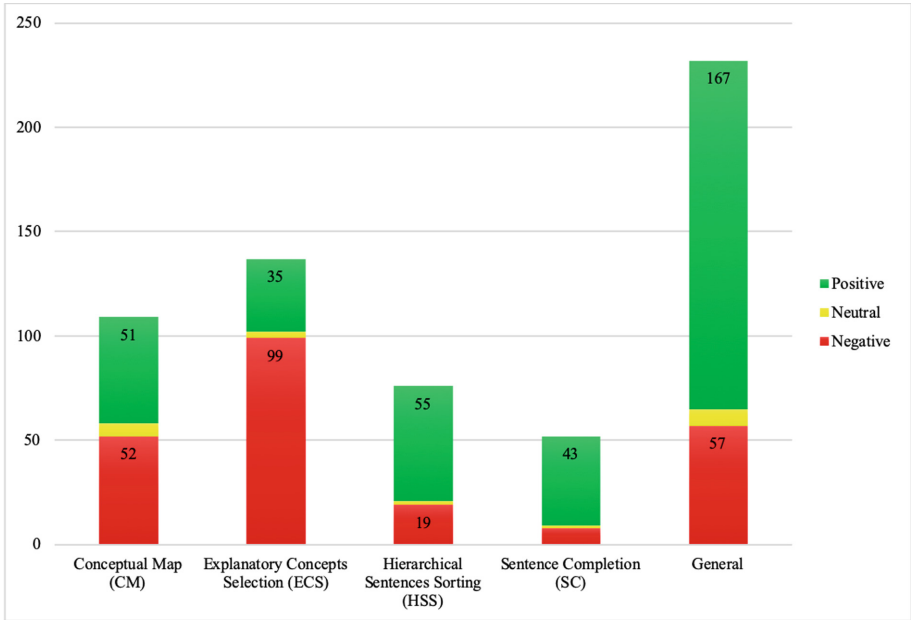
In the case of the codes versus the value assigned, Fig. 6 shows that *Useful* received 33.6% of the total negative mentions, followed by *Learnability* and *Findable* with 18.5% each.

So far, the results obtained clearly show that the *Explanatory Concepts Selection (ECS)* functionality was perceived by users as the one that needs more improvements. In relation to codes or factors, the one that presented more suggestions for improvement was the *Usable* factor.

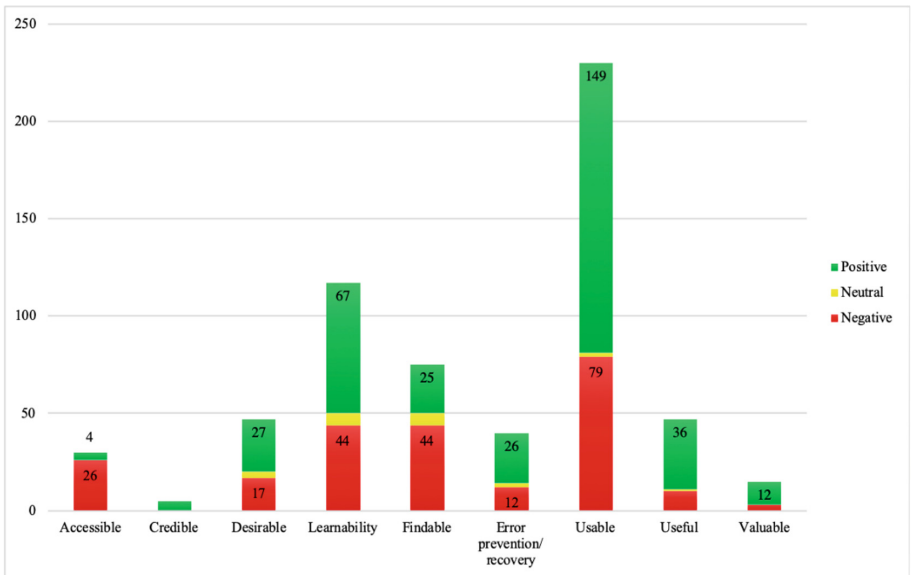
#### 4.2 User Experiences

Now we are going to go into detail about the suggestions for improvements and the perceived benefits mentioned by the users. For this, the information is organized in relation to the main themes: "Easy of Use", "Appreciation", "Satisfaction", "Learnability" and "Error Prevention/Recovery".

**Easy of Use.** This theme attempts to group user comments regarding the general use of functionalities. When it comes to using a functionality, is it usable, findable and accessible? What are the improvements that should be introduced to the functionality? What are the positive aspects of the use of functionality?



**Fig. 5.** Mentions filtered for functionality. By value and sorted by functionality type.



**Fig. 6.** Mentions filtered for functionality. By value and sorted by factor.

It was mentioned before that the Explanatory Concepts Selection (ECS) functionality presented the higher quantity of negative comments of all the functionalities tested. This

fact can be interpreted as the ECS was perceived by users as the one that needs more improvements.

Most interviewees indicated difficulty in the way they should select the words. They emphasized that it was not intuitive and that the way of presenting the task led to the writing of the words directly in the blank spaces of the text:

*“[...] I had to first select the text and then put “select” and there it was saved, then perhaps it was not so intuitive for the person.” [I3; ECS; Usable; Negative]*

*“I thought I had to write or insert text ... something was going to appear, and then I realized that it was like that ... [...]” [I10; ECS; Usable, Negative]*

Some other negative comments had to do with the way in which the different parts of the task presented in that functionality were structured and with the information provided:

*“It was very clear, but I didn’t know if it was possible to put on more than one word or an exact word ... that was it.” [I11; ECS; Findable; Negative]*

*“Of course, one does not realize quickly that it is a button, then perhaps if it could be highlighted.” [I4; ECS; Findable; Negative]*

One issue to mention in this theme is regarding the Accessible factor. That is a factor that has almost no positive mentions. These few are related to the utility perceived by the users of the zoom tool in the text boxes. The negative mentions, they are all related to the font size used and the difficulty in detecting the scroll bar of text boxes:

*“[...] in fact, what was useful to me was that zoom option.” [I1; General; Accessible; Positive]*

*“Ah, yes, really ... here it says zoom ... ah ... but only those in the (box) text.” [I7; General; Accessible; Negative]*

*“I cannot visualize well that I had a bar to download the text [...] and then I thought that the text was just there, and I got a little lost.” [I6; ECS; Accessible; Negative]*

On this theme, the functionalities that turned out to be the best evaluated by users were Conceptual Map (CM) and Hierarchical Sentences Sorting (HSS). These two functionalities were highlighted for being intuitive:

*“And that on the concept map that marked in asterisk what you had already put ... that helps a lot.” [I1; CM; Usable; Positive]*

*“As for the part of the hierarchical order, just making a simple move and changing from ... it worked well.” [I4; HSS; Usable; Positive]*

*“Dragging from top to bottom ... that was good.” [I9; HSS; Usable; Positive]*

*"[...] the part of the map and that of placing the words, yes. It is quite intuitive that where the arrow is here to place something. It guides you to what you have to do." [I14; CM; Usable; Positive]*

**Appreciation.** This theme attempts to group user comments regarding their perception of utility and value. When using the functionalities, are these relevant to the user in relation to its objectives? These functionalities offer the user some added value, something worthwhile.

Perceived benefits from the use of the different functionalities were mentioned by the users. Some of them were related to the usefulness and added value of the general proposal. Comments included:

*"I find that it is more entertaining to do a test in this way than to have to use one of the typical ones that everyone already knows." [I1; General; Valuable; Positive]*

*"The zoom part of the text, I found that very useful." [I7; General; Useful; Positive]*

*"[...] if you use this system for, for example, for tests, I find it very good, because it is also more interactive." [I11; General; Valuable; Positive]*

Some other comments were oriented to highlight some specific characteristics of a functionality that were appreciated by the users:

*"[...] the same as just selecting and adding (text), I found that super innovator, because I had never seen it." [I1; ECS; Valuable; Positive]*

*"It is the one that stands out [...] the most because it looks a bit like what you do when you scratch a text in a book and then remember it." [U2; ECS; Valuable; Positive]*

*"[...] but yes, I find it good, because if we do it on paper, poor teacher [...] then bringing that paper comfort here is super good." [I14; ECS; Useful; Positive]*

**Satisfaction.** This theme integrates the information obtained about the fulfilment of user wishes, expectations, or needs, and the pleasure derived from this. This has to do with comments associated with the *Credible* and *Desirable* factors, which consider aspects of trust for the use of the functionalities and the emotional impact generated on the user.

These mentions are undoubtedly modulated by personal tastes and previous user experiences. Therefore, it is not uncommon to find a wide variety of comments on functionalities, both positive and negative. As an example, some comments about the *Sentences Completion (SC)*, expressing different user preferences:

*"Already, the Sentence Completion. That was very entertaining, it looked a lot like the PSU (test for university selection), which they gave you a text and you have to choose the word that according to you is the one that most suited." [I1; SC; Desirable; Positive]*



*"[...] although I would have preferred, I don't know ... instead of having the words to select in each one of them, having fixed them and dragging them down the text". [I12; SC; Desirable; Negative]*

In general, talking about all the functionalities that integrate the web-based survey tool, some users mentioned the colours used as something negative :

*"Maybe for me it's a bit dark, but under that scope it's fine." [I9; General; Desirable; Negative]*

*"I would put it as more vivid colours to make it more striking, because I feel that black and white is like very, I don't know, it doesn't call me ... I don't know, a visual theme." [I12; General; Desirable; Negative]*

But also, there were users that found pleasant the minimalist design:

*"So, in general I like ... this ... I give my approval, if I can say anything." [I6; General; Desirable; Positive]*

*"I liked the design". [I8; General; Desirable; Positive]*

Always considering that the amount of mentions associated with the Credible and Desirable factors are few compared to the other factors (See Fig. 2), the functionality that appeared as not suiting the preferences of the users is the Conceptual Map (CM). They mentioned:

*"I think the most correct thing would be that once one selects the option and it is already selected to highlight it, that is, as to highlight it with a colour and also highlight where it was already selected so that one realizes that It was already really selected." [I3; CM; Desirable; Negative]*

*"So that is another technical side like me with the maps, for anyone it is different, but in my case, it was like that. I felt it very close, I could not connect well, but I intuited more than anything." [I14; CM; Desirable; Negative]*

On the contrary, the functionality that appears as the users favourite was the Hierarchical Sentences Sorting (HSS):

*"[...] because I feel that it is easier for me to sort things out, than to be... I don't know, for example, to make the concept map, or the sentence completion, yes, I think that hierarchical order." [I7; HSS; Desirable; Positive]*

*"I would use for example that of underlining because I liked." [I2; ECS; Desirable; Positive]*

**Learnability.** As mentioned earlier, it is relevant to detect those difficulties associated with learning the use of functionalities since they will be part of a web tool for evaluation. It was also shown in Fig. 5 that the functionalities that presented the greatest

amount of negative comments turned out to be *Explanatory Concepts Selection (ECS)* and *Conceptual Map (CM)*.

The main difficulties revealed by the users in the case of the functionality *Explanatory Concepts Selection (ECS)* were related to the way in which they interacted with the text for the selection of the words. Participants expressed:

*“At first I didn’t know where to choose the text to underline to add in the test.”* [12; ECS; Learnability; Negative]

*“The functionality of selecting text was the most complicated because it is the least common [...]”* [15; ECS; Learnability; Negative]

Similarly, the difficulties mentioned in relation to the functionality of *Conceptual Map (CM)* were related to the way in which the users needed to interact with the boxes to complete. They mentioned that:

*“[...] but still, at the beginning I didn’t understand how to select.”* [11; CM; Learnability; Negative]

*“[...] On the concept map as I did not know that it had to be pressed instead of selecting ... that was the one that cost me the most [...]”* [16; CM; Learnability; Negative]

The *Hierarchical Sentences Sorting (HSS)* is the functionality with the most quantity of positive comments associated with the *Learnability* factor. As an example, the users mentioned that:

*“The two that were faster were the Hierarchical Sentences Sorting and the Sentence Completion. It is very explanatory what to do [...]”* [14; HSS/SC; Learnability; Positive]

*“Ah yes, that is what I actually understood without reading, [...] and I read it the same way, but apart from that I dragged and without problems.”* [16; HSS; Learnability; Positive]

**Error Prevention/Recovery.** As mentioned before, it is very important to design a user interface with focus on an ease of use and absence of cognitive obstructions. It turns out to be relevant the possibility of preventing usage errors and making simpler the correction of mistaken responses during tasks.

In general, users indicated that they had no problems correcting errors while using the functionalities. They mentioned that:

*“I did not notice that much, but I made several changes in terms of answers and yes (I could correct the errors).”* [11; General; Error prevention/recovery; Positive]

*“Well, there was no problem. Once you delete something in the checkbox, in the selection, there is no problem in changing it ... none.” [I10; General; Error prevention/recovery; Positive]*

In particular, the functionality Sentences Completion (SC) resulted the one with some negative comments. These comments were about the way of changing the selected word:

*“[...] because I had to change the answers twice, instead of being directly excluded, for example, having an option to leave it empty and then change it [...].” [I6; SC; Error prevention/recovery; Negative]*

*“The third, the words ... How to go back I get a little complicated [...].” [I14; SC; Error prevention/recovery; Negative]*

On the other hand, the functionality Conceptual Map (CM) received both negative and positive comments about how they should proceed to make changes to the words already included in the concept map. As an example, they mentioned:

*“[...] in the concept map I had to change one to another and I had no problems [...].” [I12; CM; Error prevention/recovery; Positive]*

*“On the concept map, not ... because I was confused by the asterisk, and suddenly I couldn't and I had to put anyone to put it back there, that complicated me a little [...].” [I13; CM; Error prevention/recovery; Negative]*

## 5 Discussion

The aim of this study was to evaluate the usability of a new web-based survey tool designed for the assessment of reading comprehension and metacognitive strategies in university students, using a qualitative setting. Traditionally, tests are used as a way of obtaining information on how users use the system. However, it is possible to affirm that the data obtained in this qualitative usability study provide in-depth and detailed information necessary to improve the functionalities designed for the web-based survey tool.

Related to “Easy of Use”, data from the sample of early-course university users suggest that the best functionalities, according to what they perceived, turned out to be Hierarchical Sentences Sorting (HSS) and Sentences Completion (SC). Users highlighted the intuitive design, which simulates manual actions performed by people to sort things and to complete texts, and only mentioned possibilities for improvements associated with personal tastes and visual design.

On the other hand, the functionality that received the most negative comments was Explanatory Concepts Selection (ECS). In this case, the comments were mainly related to the difficulty in understanding how the method of text selection was. The structure of the functionality and the information provided received negative comments as well. All the

specific and detailed comments and suggestions were really helpful for the improvement of the web-based survey tool.

Regarding “Appreciation”, the data obtained allow us to point out that the users found potential in the proposed functionalities. They highlighted the novelty of the proposal, the intuitive of the general use and those characteristics of the functionalities that simulate real actions that the subjects perform in pencil and paper tasks.

“Learnability” plays a very important role in the use of the designed web-based survey tool. This is because if the functionalities are not easy enough to learn to use the users would spend more time in that process instead of focusing on the task that assess their reading comprehension. A first inspection indicated that the amounts of negative and positive comments about Learnability factor were similar. A more detailed analysis showed that functionalities Explanatory Concepts Selection (ECS) and Conceptual Map (CM) turned out to be the ones that obtained the most negative comments. Users highlighted the difficulty found in both functionalities to learn the dynamics of interaction for the selection of the necessary texts.

It is important to note that the methodology designed for the execution of this qualitative usability evaluation proved to be very helpful and with potential for improvements. The construction of a careful semi-structured interview guide and the fact of using the selected factors to encode the interviews allowed an organization of the data obtained capable of providing an orderly and efficient access to the information provided by the users. The descriptive analysis of the factors in association with the functionalities allowed us to quickly detect the main problems experienced by the users and to which functionalities they were associated. This also allowed to generate traceability towards the comments made by the users in the interview and therefore to the information that is of our interest in improvement of the web-based survey tool. Obviously, in the same way, we were able to detect the perceived benefits from the use of the different functionalities.

A possible limitation to this work methodology lies in the existence of the possibility that the negative comments are in an excessive number due to the realization of questions to deepen the information associated with aspects of the functionalities to be improved. As future work, an analysis associated with this problem is planned, in which the negative comments generated freely by users and those generated by in-depth questions made by the interviewer are differentiated.

Another important issue to mention is that for the present work the information obtained on the approach of the tasks was not used. It is possible to think that some comments made by users associated with the functionalities could have been related to the design of the task or problems of the user to perform the specific task (and not with the use or design of the functionality). It is foreseen as future work to inquire if the task had any modulating effect on the perception generated by the user on the functionality.

## 6 Conclusion

This study explored users’ interactions with a web-based survey tool specifically designed to evaluate reading comprehension and metacognitive strategies. Detailed information about their user experiences was obtained through semi-structured interviews and a qualitative analysis.

The analysis performed allowed us to distinguish users' perceptions of different functionalities. Also, it was possible to differentiate the users' experiences associated with the tasks the functionalities were developed for and the experiences of using the functionalities themselves. Moreover, it was possible to associate positive, negative and neutral effects with the codes and their detailed explanations. The methodology carried out was of great benefit and evidenced the potential of this type of analysis in a user evaluation.

With semi-structured interviews, it was possible to achieve improvements in the information obtained with respect to quantitative usability tests. They provided insight into usage, benefits and improvements needed for the web-based survey tool's functionalities. We use the results of this study to recommend changes in the design of the web-based survey tool with the proposed tasks to assess reading comprehension.

## References

1. Carvalho, A.: Usability testing of educational software: methods, techniques and evaluators. *Actas 3º Simpósio Int. Informática Educ.*, pp. 139–148 (2001)
2. Nielsen, J.: *Usability Engineering*. Morgan Kaufmann (1994)
3. Kumar, B.A., Mohite, P.: Usability of mobile learning applications: a systematic literature review. *J. Comput. Educ.* **5**(1), 1–17 (2018)
4. Anderson, K., Burford, O., Emmerton, L.: Mobile health apps to facilitate self-care: a qualitative study of user experiences. *PLoS ONE* **11**(50), e0156164 (2016)
5. Hertzum, M.: A usability test is not an interview. *Interactions* **23**(2), 82–84 (2016)
6. Taylor, C.: *Interviewing in Qualitative Research in Healthcare*, Maidenhead Berks. Open (2005)
7. DiCicco-Bloom, B., Crabtree, B.F.: The qualitative research interview. *Med. Educ.* **40**(4), 314–321 (2006)
8. Blandford, A.E.: *Semi-structured qualitative studies*. Interaction Design Foundation (2013)
9. Pirnay-Dummer, P., Ifenthaler, D., Spector, J.M.: Highly integrated model assessment technology and tools. *Educ. Technol. Res. Dev.* **58**(1), 3–18 (2010)
10. Parshall, C.G., Harnes, J.C.: Improving the quality of innovative item types: four tasks for design and development. *J. Appl. Test. Technol.* **10**(1), 1–20 (2009)
11. Lund, A.M.: Measuring usability with the use questionnaire. *Usability Interface* **8**(2), 3–6 (2001)
12. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **13**, 319–340 (1989)
13. Lin, H.X., Choong, Y.-Y., Salvendy, G.: A proposed index of usability: a method for comparing the relative usability of different software systems. *Behav. Inf. Technol.* **16**(4–5), 267–277 (1997)
14. OECD: *Students, Computers and Learning* (2015)
15. Martínez, T., Vidal-Abarca, E., Sellés, P., Gilbert, R.: Evaluación de las estrategias y procesos de comprensión: el Test de Procesos de Comprensión. *Infancia Aprendiz* **31**(3), 319–332 (2008)
16. Vieiro, P., Gómez, I.: *Psicología de la lectura*. Esp. Editor. Pearson (2004)
17. Arribas, D., Cuetos, F., Rodríguez, B., Ruano, E.: *PROLEC-R Batería de evaluación de los procesos lectores, revisada*. TEA Ediciones, Madrid (2010)
18. Rodríguez, V.J., Ferreras, A.P., Izquierdo, J.M.A., Durante, L.A.: Medición de estrategias metacognitivas mediante la Escala de Conciencia Lectora: ESCOLA. *Electron. J. Res. Educ. Psychol.* **7**(2), 779–804 (2009)

19. Ávila Clemente, V., Gil Pelluch, L., Gilbert Pérez, R., Maña Lloria, A., Llorens Tatay, A.C., Vidal-Abarca Gamez, E.: Método de evaluación dinámica automatizado de competencias lectoras para educación secundaria (EdiLEC). *Univ. Psychol.* **15**(1), 219–232 (2016)
20. Dialect: [En línea] (2013). <http://www.diamas.cl/#brief1>. Accedido 28 ene 2020
21. Optimus: [En línea] (2019). <http://nilvem.com/optimus3/>
22. Cain, K., Oakhill, J.: Assessment matters: issues in the measurement of reading comprehension. *Br. J. Educ. Psychol.* **76**(4), 697–708 (2006)
23. McKay, T.: More on the validity and reliability of C-test scores: a meta-analysis of C-test studies. Georgetown University (2019)
24. Liu, J.: The assessment agent system: design, development, and evaluation. *Educ. Technol. Res. Dev.* **61**(2), 197–215 (2013)
25. Braun, V., Clarke, V.: Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**(2), 77–101 (2006)
26. Morville, P.: Experience design unplugged. In: *ACM SIGGRAPH 2005 Web Program*, pp. 10 (2005)



# Automatic Versus Manual Forwarding in Web Surveys - A Cognitive Load Perspective on Satisficing Responding

Arto Selkälä<sup>1</sup>(✉), Mario Callegaro<sup>2</sup> , and Mick P. Couper<sup>3</sup>

<sup>1</sup> University of Lapland, Rovaniemi, Finland  
arto.selkala@ulapland.fi

<sup>2</sup> Google, London, UK  
callegaro@google.com

<sup>3</sup> University of Michigan, Ann Arbor, MI, USA  
mcouper@umich.edu

**Abstract.** We examine the satisficing respondent behavior and cognitive load of the participants in particular web survey interfaces applying automatic forwarding (AF) or manual forwarding (MF) in order to forward respondents to the next item. We create a theoretical framework based on the Cognitive Load theory (CLT), Cognitive Theory of Multimedia Learning (CTML) and Survey Satisficing Theory taken also into account the latest findings of cognitive neuroscience. We develop a new method in order to measure satisficing responding in web surveys. We argue that the cognitive response process in web surveys should be interpreted starting at the level of sensory memory instead of at the level of working memory. This approach allows researchers to analyze an accumulation of cognitive load across the questionnaire based on observed or hypothesized eye-movements taken into account the interface design of the web survey. We find MF reducing both average item level response times as well as the standard deviation of item-level response times. This suggests support for our hypothesis that the MF interface as a more complex design including previous and next buttons increases satisficing responding generating also the higher total cognitive load of respondents. The findings reinforce the view in HCI that reducing the complexity of interfaces and the presence of extraneous elements reduces cognitive load and facilitates the concentration of cognitive resources on the task at hand. It should be noted that the evidence is based on a relatively short survey among university students. Replication in other settings is recommended.

**Keywords:** Web surveys · Interactivity · Human-computer interaction · Interface design · Cognitive load theory · Cognitive Theory of Multimedia Learning · Survey satisficing theory · Sensory memory · Working memory · Long-term memory · Satisficing responding · Automatic forwarding · Manual forwarding · Item-level response times · Standard deviation · Cognitive burden · Nondifferentiation · Straightlining · Intrinsic cognitive load · Extraneous cognitive load · Eye-tracking · Eye-movements · Proactive attention · Contiguity principle · Spatial orienting · Attentional capture · Group-mean centering · Log-linear model

## 1 Introduction

The e-environment consists of various technologies, applications and functionalities like web pages, web browsers, multimedia presentations and web surveys, all of them involving human-computer interaction (HCI). In this article, we apply *Cognitive Load theory* (CLT), *Cognitive Theory of Multimedia Learning* (CTML) and *Survey Satisficing Theory*. CLT and CTML are widely applied in human computer interaction research when Survey Satisficing theory is applied in web survey methodology, the field which has much in common with human-computer interaction research. We create a theoretical framework based on these theories taken also into account the latest findings of cognitive neuroscience in order to understand human-computer interaction in a more coherent way in order to apply this understanding in a particular web survey interface employing automatic page forwarding. In addition, we develop a new method to measuring satisficing responding in web surveys which we introduce in this article applying it in the web survey data.

We investigate a particular functionality of web surveys, the page forwarding procedure and its implications for satisficing respondent behavior in an interaction of other task completion elements. Web surveys are a widely-used data gathering method in social science and market research representing an important advancement in the evolution of self-administered questionnaires making large samples affordable to a wide range of researchers (Tourangeau et al. 2013). However, the prevailing lack of national e-mail registries appears as a major challenge in applying web surveys as a scientific data gathering method making it difficult to constitute a representative sample at the general population level yet overcome by online panels and weighting/post-stratification methods (Callegaro et al. 2015). In addition, web surveys in particular have been found to be subject to declining participation rates (Callegaro et al. 2015). This challenge has been overcome by advanced invitation methods (Bandilla et al. 2012; Callegaro et al. 2015; Dillman 2019; Kaplowitz et al. 2012; Selkälä et al. 2019). It should also be noted that a poor participation rate of any given survey data does not necessarily imply poor statistical representativeness given that a nonresponse rate correlates only weakly with a nonresponse bias (Davern 2013; Groves and Peytcheva 2008).

Computer and web-based survey technology has made possible question formats, layouts and functionalities that would be impossible or difficult to implement using traditional paper questionnaires. Interactivity is then a key element of web surveys, unlike paper and pencil surveys (Couper 2008). One element of interactivity is automatic forwarding, which advances respondents automatically to the next item without the need to click on a “next” button (as in manual forwarding). Another well-known example of interactivity in web surveys is a drop-down question; a format which is impossible to conduct as such on a paper and pencil survey.

In the perspective of a human-computer interaction, a major challenge becomes how to manage a cognitive burden of the participants of any given task. This becomes essential given that in most cases a poor task performance is found to occur due to an excessive cognitive burden of the participants. One of the most influential theories to understand the formation of a cognitive burden in different contexts is the cognitive load theory (CLT) largely adopted in the field of human-computer interaction (HCI)



(Hollender et al. 2010), in usability research (<https://www.nngroup.com/articles/minimize-cognitive-load/>), in internet psychology (Sundar 2007), in the field of instruction science (Clark and Mayer 2016), in the research of multimedia learning (Mayer and Fiorella 2014) and in the web survey methodology.

Given that our aim in this study was not only to analyze how the individual questions affect satisficing responding in different interfaces, automatically (AF) and manually (MF) forwarded web surveys but also to analyze how these interface features interact with the individual questions generating satisficing responding we argue that the conventional approach to understand a cognitive survey response process (Tourangeau 1984; 2018) is in this respect insufficient. This is due to the fact that the different versions of this description start at the level of interpreting the meaning of each question or at the level of comprehension. In both of these cases, the response process starts at the level of working memory turned to be insufficient in terms of analyzing the interface features. In order to understand the cognitive response process as a whole in web surveys the process should be interpreted starting at the level of *sensory memory* (Mayer 2014). As a cognitive response process starts at the level of sensory memory it occurs by selecting appropriate elements into the working memory based on spatial orienting and attentional capture in the preparation of eye movements further generating intrinsic or extraneous load of respondents (Theeuwes 2014). In addition, we expect the total cognitive load of respondents accumulating as responding proceeds across the questionnaire based on the interaction of intrinsic and extraneous loads and the proactive nature of recall (Heideman et al. 2018; Nobre and Stokes 2019).

## 2 Theoretical Background

### 2.1 Automatic Versus Manual Forwarding

Auto forwarding or automatic advance in web surveys is a functionality that can be used when a question type has mutually exclusive answers (such as radio buttons) where clicking on a radio button response, for example, takes the respondent directly to the next page (question) without the need to click on a “next” button. Question types like check-all-that-apply and open-ended questions cannot use automatic forwarding. The main arguments in support of auto forwarding are that it (1) reduces respondent burden (number of clicks) and (2) serves as a “forcing function,” requiring the selection of a response to proceed (Selkälä and Couper 2018). With the recent rise in the proportion of respondents completing web surveys on mobile devices (specifically smartphones; see Couper et al. 2017) and the corresponding finding that surveys completed on smartphones take longer to complete than those completed on personal computers (PCs) (Couper and Peterson 2016), researchers are trying to find ways to make such surveys more efficient, especially for the sets of questions with similar response options (see de Bruijne et al. 2015; de Leeuw et al. 2012; Klausch et al. 2012).

As Selkälä and Couper (2018) have noted, while there have been many arguments for and against auto forwarding, empirical research on the topic is scarce. In the first known study to examining auto forwarding, Rivers (2006) reported significantly ( $p < .001$ ) fewer break offs in the AF (40.6%) than in the MF (49.4%) version. He also reported significantly ( $p < .01$ ) shorter completion times for the AF (median = 19.3 min) than

the MF (median = 23.1 min) version and higher levels of user satisfaction with the AF version ( $p < .001$ ). Both of these results were found also in the Selkälä and Couper study (2018). Hays et al. (2010) also found that the survey took about 50% longer ( $p < .025$ ) in the MF version: mean completion times were 9.1 min for AF and 13.5 min for MF. Missing data, reliability, and mean scale scores were similar across the groups. Somewhat similar findings were reached by Giroux et al. (2019) given that they did not find significant differences in survey duration time, straight-lining, breakoff rates, or item nonresponse (for mobile users) between the two experimental groups, but desktop users without the automatic advancement feature had higher item nonresponse.

The research findings regarding item nondifferentiation<sup>1</sup> also called straightlining between auto and manual forwarding are not entirely consistent. Auto forwarding has been shown to increase non-differentiation and primacy effect (Hammen 2010) but also decreasing it in the case of particular horizontal scrolling matrix” (HSM) version (de Leeuw et al. 2012) leading the authors considering findings as an evidence of deeper processing in the AF version. On the other hand, automatic forwarding has been almost consistently shown to decrease web survey completion times in comparison with manual forwarding in the Hays et al. (2010), Rivers (2006) and Selkälä and Couper (2018) studies but not in the Giroux et al. (2019) study. When it comes to item level response times Selkälä and Couper (2018) found AF respondents taking on average 0.4 s ( $p < .001$ ) longer to provide an initial answer to each item. They (2018, p. 11) argue this suggests support for the hypothesis that by simplifying the response process, auto forwarding allows the respondent to focus more fully on the item under consideration.

From the more specific perspective focusing on an opportunity to change an already given answer auto forwarding and manual forwarding represent different solutions. Respondents in the MF condition could change answers before proceeding, whereas AF respondents would need to return to the item to make a change. Respondents can also return to review previous items without making changes. Selkälä and Couper (2018) found that MF respondents change responses significantly more on experimentally manipulated items conveying a low information accessibility or a consistency requirement in comparison to corresponding neutral items in the control groups: 15.5% of the respondents exposed to the low information accessibility version changed answers to this item, compared with 3.3% for the control groups; similarly, 14.9% of those exposed to the consistency requirement changed answers, compared to 2.9% for the control group. They did not find such differences in the AF group; overall changes were very low (0.6% to zero). Taking into account the response time findings that experimentally manipulated items took longer to complete, Selkälä and Couper (2018) conclude that the questions conveying low accessible information or consistency requirement increase cognitive burden of respondents. To some extent, this leads respondents to revisit those items and change their responses. However, no evidence was found to support their hypothesis that respondents in the AF groups return more to experimentally manipulated items in order to change their responses. Instead, they found higher rates of returns for both MF and AF groups to experimentally manipulated items and higher rates of changed answers to the MF groups but not the AF groups. Giroux et al. (2019) found similar results given that

<sup>1</sup> Item nondifferentiation is a response style where the answers to a battery of questions with the same response options (e.g. a table or grid) are the same or very similar among each other.

in their study respondents receiving the automatic advancement treatment on average changed about 50% fewer answers across the survey instrument than those who did not receive the automatic advancement design.

## 2.2 Cognitive Load Theory, Cognitive Response Process, and an Expected Accumulation of Cognitive Load in the AF and MF Interfaces

Cognitive load theory (CLT) provides a theoretical framework addressing individual information processing and learning (Paas and Sweller 2012). CLT is concerned with the learning of complex cognitive tasks, in which learners are often overwhelmed by the number of interactive information elements that need to be processed simultaneously (Paas et al. 2010). CLT is based on the definition of different types of cognitive loads: intrinsic, extraneous, and germane (Paas et al. 2003). Intrinsic load is the load caused by the complexity of the materials to be learned and therefore the complexity of the schemas that must be acquired (Paas et al. 2010). Extraneous load is caused by inadequately designed instructional procedures that interfere with schema acquisition. Germane load is generated as a result of beneficial instructional design factors that support schema creation, learning, instructional task performance, and transfer (Ayres and van Gog 2009; Hollender et al. 2010; Leppink et al. 2013; van Merriënboer et al. 2006; Paas et al. 2010).

CLT is based on understanding how these different types of loads interact with each other in any learning process or a task completion process. An essential clarification in this respect is offered by Paas et al. (2010). They argue that intrinsic load is dependent upon element interactivity, the number of elements that need to be processed simultaneously by the learner. If element interactivity is high, learning becomes difficult and WM-resource intensive [WM: working memory], whereas for low element interactivity material, learning is easier, requiring fewer WM resources. They also argue (2010) that when instructional material is poorly constructed, extraneous load is generated because the learner is diverted away from schema acquisition and uses up precious WM resources by trying to deal with a suboptimal learning environment. Because intrinsic and extraneous cognitive load are additive, an increase in extraneous cognitive load reduces the WM resources available to deal with intrinsic cognitive load and hence reduces germane cognitive load. On the other hand, when intrinsic cognitive load is high, it becomes important to decrease extraneous cognitive load; otherwise, the combination of both might exceed the maximum cognitive capacity and thus prevent effective, or germane, processing activities to occur.

In most cases when CLT is applied an analysis of a cognitive process is based on intrinsic and extraneous types of loads. This is probably due to practical reasons given that these concepts offer a necessary but also sufficient theoretical basis to understand most of the cognitive processes. As discussed above, intrinsic elements refer to the task completion elements that are essential to the task and cannot be separated of it without jeopardizing the accomplishment of the task. In other words, they are necessary to learn in terms of the task completion. On the other hand, the elements capable to generate extraneous load are in most cases irrelevant in terms of the task completion. They consist of technical procedures or information which is redundant or overlapping with the intrinsic information of the task. An excessive total cognitive load also understood

as a working memory overload can occur either as a result of an interaction between simultaneously occurring intrinsic elements or an interaction between intrinsic and extraneous elements. In a detail the relationship of intrinsic and extraneous loads should be understood as follows. Because intrinsic and extraneous cognitive load are additive, an increase in extraneous cognitive load reduces the working memory resources available to deal with intrinsic cognitive load (Paas et al. 2010). Therefore, if intrinsic load is high, extraneous cognitive load must be lowered. Inversely, if intrinsic load is low, a high extraneous cognitive load may not be harmful because the total cognitive load occurs within working memory limits (van Merriënboer and Sweller 2005).

From the perspective of different memory types CLT focuses on an interaction between working memory (WM) and long-term memory (LTM) which is the key to understanding how learning takes place and how complex problems get solved (Ayres 2018). Like *Cognitive Theory of Multimedia Learning*, CLT argues that new information needs to be first processed and integrated with prior knowledge in WM before it is encoded in LTM as new knowledge (Ayres 2018). These theories also share an understanding that the role of participants in any type of task of Web should be understood as the active participants rather than passive recipients of communication (Sundar 2007). In web surveys this feature of the task completion environment is introduced as an interactive principle (Couper 2008).

However, what differs between CLT and CTML from the perspective of memory processes is that CLT focuses on interaction between working memory and long-term memory when CTML additionally recognizes the importance of *sensory memory* (Mayer 2014). The sensory memory operates at the level of spatial orienting and attentional capture that participate in the preparation of eye movements, in turn responsible for cognitive load formation (Mayer 2014; Theeuwes 2014). This process as a whole is concentrated as a *selection* of relevant information transferred to *working memory* further organizing it in order to create mental representations that are integrated with a prior knowledge of long-term memory (Mayer and Moreno 2003). From this perspective learning or a task completion based on any visual stimuli starts at the level of *sensory memory* proceeding towards a *working memory* and *long-term memory* interaction enabling a deeper information processing. Regarding the common descriptions of a cognitive survey response process the organizing and integrating information processing levels are well represented, unlike the sensory memory level. We therefore suggest that these descriptions should be completed by the selection process occurring at the level of sensory memory.

The widely applied description of a cognitive survey response process is as follows (Tourangeau 1984, 2018):

1. Comprehension (interpreting the meaning of each question)
2. Retrieval (searching and retrieving information stored in memory)
3. Judgment and estimation (integrating the information into an opinion or judgment)
4. Reporting an answer (expressing this opinion appropriately).

A more recent description of the cognitive survey response process with the addition of the level of sensory memory is as follows:

1. Selection (transfer information from sensory memory to working memory based on eye movements)
2. Comprehension (interpret the intended meaning of question)
3. Retrieval (retrieve relevant information from memory)
4. Judgment and estimation
5. Reporting an answer.

As a consequence of the addition introduced above, it becomes possible to analyze the web survey response process from the perspective in what extent the interface features likely generate a cognitive load of the respondents as information is transferred from sensory memory (SM) to working memory (WM). This framework turns out to be beneficial given that in most cases when web survey responding is evaluated, the interface features are excluded from the analysis, focusing instead on the cognitive response process in terms of the substantive nature of individual questions. What is needed instead when trying to understand the web survey response process from the perspective of cognitive burden or cognitive load is to take into account the interface features together with the individual questions.

With regard to the interaction between participants and an interface it has been found that greater interactivity of users with a website engenders greater navigational - and hence cognitive load on users (Sundar 2007). This relationship can also be expected to occur in web surveys in a way that greater effort navigating through the web survey interface increases the total cognitive load of the participants. In addition, an excessive total cognitive load should be expected to occur either as a result of an interaction of intrinsic elements or an interaction of intrinsic and extraneous elements.

A more elaborate view of how the navigation occurs on the interface can be reached by using eye tracking method measuring the eye movements of the participants (Kim et al. 2016; Krejtz et al. 2018; Zagermann et al. 2016). Eye-tracking can be used to detect intrinsic as well as extraneous cognitive load (Makransky et al. 2019). An extraneous load is generated for instance, when two elements, both necessary for learning, are located visually separated on the interface, making it difficult for the participants to reach a coherent understanding about their interrelationship. In order to improve their understanding, the participants are therefore forced to scan back and forth wasting precious cognitive processing capacity (Mayer 2017). This kind of interface design violates *a contiguity principle* generating an extraneous *sensory memory* load further leading to working memory overload (Clark and Mayer 2016; Makransky et al. 2019). The harmful consequences of this design can be explained by the law of proximity referring to a phenomenon fostering learning when related representations are spatially integrated or close to each other (Beege et al. 2019; Clark and Mayer 2016). With regard to surveys it is shown that when items are presented in proximity to each other, the likelihood for an assimilation effect increases (Couper et al. 2001; Tourangeau et al. 2013). The reason for this lies in the law of proximity, which causes items to be perceived as a group (Toepoel et al. 2009).

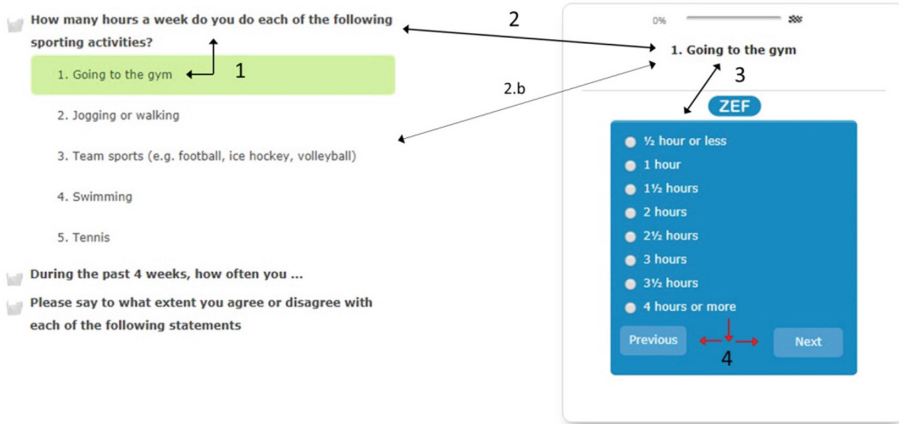
In a typical situation when CLT is applied to an evaluation of usability of any given interface the interface elements affecting usability remain as permanent across the task completion. This is the case for instance regarding navigation through an ordinary web page not including any interactive elements. However, when CLT is applied to the web

survey response process its role should be understood differently. What makes the web survey task completion process different is the repetition. In the web survey response process the same type of intrinsic-, and extraneous elements repeatedly follow each other instead of occurring once. Taken into account the particular nature of a web survey responding process, we expect the total cognitive load to be generated across the response process based on the permanent elements of the process like interface features and non-permanent elements like individual questions and their interaction. We also expect the total cognitive load of respondents accumulating across the responding occurring at the highest level at the end of the questionnaire assuming that the capacity of individual web survey questions to generate a cognitive load of respondents remains somewhat equal across the survey. The cognitive load accumulation is recognized in CLT and measured by subjective measures like rating scales as well as objective measures like the amount of time the learner spends on completing the task or navigation behavior (Antonenko and Keil 2018).

Given that we were not able to use an eye tracking method we nevertheless offer a hypothesis with regard to the eye movements on the studied interfaces AF and MF. The hypothetical framework regarding eye movements makes it possible to understand how the studied interface features contribute and interact in the formation of different types of cognitive loads during the response process. It should be noted that the respondents in both studies were encouraged to use PCs or tablets but were not prevented from using a smartphone. Despite this opportunity, only 16% of the respondents completed the survey using a smartphone (Selkälä and Couper 2018). Thus, although the interfaces for PC and smartphone respondents differ considerably (Appendix 2) we find the proportion of smartphone respondents so low that we do not expect it to affect the results.

The expected eye movements on a manually forwarded interface are illustrated in Fig. 1. Given that the interface in both versions (AF/MF) is divided in two parts, on the left side of the screen appears the list of the items when on the right side of the screen appears the particular question and the response options of it. As the response has been given to the particular question the next question on the list is activated with a shaded background. At the same time, it appears visible on the right side of the screen. It should be noted as well that the individual items are shared into groups. The headline of the question group appears visible on the left side of the screen when the items within it appear visible below of the headline. We expect the eye movements occurring in the response process as follows. Firstly (1) respondents are likely to focus on the relationship of the headline of a question group and the first item within it. Secondly (2) we expect them to focus on the relationship of the headline of a question group and the headline of the first item on the right side of the screen. Thirdly (2.b) we expect the respondents focusing on the relationship of the headline of the first item on the right side of the screen and other items on the item list. We expect these particular eye movements to generate lower cognitive load than other expected eye movements given that their intention is more to become aware of the task completion elements as a whole than focus on the particular question under the consideration. We expect the third (3) particular eye movements occurring between the headline of the first item on the right side of the screen and the response options below of it. In the automatically forwarded interface (AF) we do not expect any other notable eye movements to occur but in the manually forwarded version (MF) we

expect the fourth (4) major eye movement to occur between the response options of the responded question and previous and next buttons below them. We expect the first, second and the third eye movements generating intrinsic load in both of the versions. In addition we expect the fourth (4) eye movements in manually forwarded (MF) version generating extraneous load.



**Fig. 1.** The expected eye movements on manually forwarded interface.

In addition to permanent interface features discussed above the repetition of individual questions should take into account in order to understand how a cognitive load accumulates in detail in the AF and MF conditions. As illustrated in the popular descriptions of the cognitive response process (Tourangeau 1984, 2018) the retrieval is an essential part of responding to individual questions. However, this part of the response process should be understood differently when trying to achieve a coherent understanding about an accumulation of cognitive load in terms of the individual questions and the interface as a whole. The recent findings from the field of neuroscience show that recall cannot be considered just as a passive operation to retrieve something from the long-term memory. It should instead be considered as an active process especially in terms of anticipating upcoming events.

An increasing variety of experimental approaches is being used to explore how long-term memory (LTM) content is used proactively to guide adaptive behavior. The approaches share the notion that the brain uses LTM information constantly, proactively, and predictively (Nobre and Stokes 2019). Information in working memory has been considered the major source of top-down proactive attention. Even before the target stimuli appear, these memory traces influence the pattern of brain activity in a proactive fashion to facilitate the processing of signals associated with likely relevant items (Chelazzi et al. 1993; Kastner et al. 1999; Stokes et al. 2009). In addition, a recent neurophysiological study of sequential learning in a serial response task has also revealed a proactive anticipation of upcoming stimuli and associated responses based on learned spatiotemporal expectations (Heideman et al. 2018; Nobre and Stokes 2019).

These findings receive support from the survey satisficing studies showing that satisficing respondent behavior and straightlining occur more likely towards the end of the questionnaire than toward the beginning (Knowles 1988; Krosnick and Alwin 1988). When the retrieval process is understood in a proactive fashion as discussed above, anticipating the upcoming stimuli, it becomes easier to understand why satisficing respondent behavior becomes more likely towards the end of the questionnaire. When the respondent retrieves an activated material from long term memory integrating it with the mental representation based on the present question the invested cognitive effort depends also on other sources to increase total cognitive load. These sources generate intrinsic or extraneous load, or the load based on their interaction. The anticipatory nature of retrieval can explain why respondents tend to relieve excessive load through satisficing in repetitive tasks like web surveys. It can be expected to occur when the repetitive task includes certain permanent interface elements increasing the extraneous load on the respondents. This is the case in the manually forwarded interface. Under these circumstances, the respondents can easily anticipate the burdening nature of the upcoming task and adjust the invested cognitive effort across the individual items (see Heideman et al. 2018; Nobre and Stokes 2019). A crucial part of this process in the MF interface generating extraneous load is the selection between the previous and next buttons illustrated in Fig. 2.

As a consequence of an interaction between the individual questions and the MF mechanism we expect a total cognitive load of respondents accumulating and turning excessive as individual items follow each other. The excessive load is then relieved through satisficing. Inversely arguing, if we expect the accumulation of total cognitive load not occurring, we shouldn't be able to observe an increased satisficing towards the end of the questionnaire. In this case, each load in relation with the individual items should completely be relieved after giving a response. What follows is that in this case a respondent should be able to start a response process on each question without an accumulated load originated from previous questions. Given that the empirical findings and theoretical reasoning discussed above suggest otherwise we accept the accumulation hypothesis.

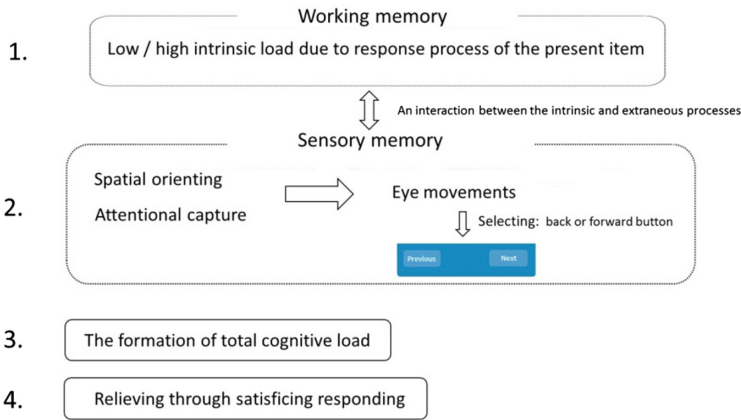
### 2.3 Survey Satisficing and Response Time Based Measurements

Krosnick's survey satisficing theory (Krosnick 1991, 1999) is probably the most influential theory regarding satisficing respondent behavior in surveys. It is based on Tourangeau's (1984) description of the cognitive process taken place in survey responding borrowing also from Simon's (1957) more general theory of decision-making. Survey satisficing occurs when instead of investing a sufficient effort to respond thoughtfully or optimally respondents take shortcuts in order to minimize cognitive effort (Kim et al. 2019; de Rada and Dominguez-Alvarez 2014; Zhang and Conrad 2018). It is this resort to a satisfactory rather than an optimal decision strategy that gives satisficing theory its name (Hamby and Taylor 2016).

Krosnick introduced the idea of "weak" and "strong" forms of satisficing. Weak satisficing occurs when the four cognitive stages of survey responding—comprehension, recall, retrieval, and judgment (Cannell et al. 1981; Tourangeau et al. 2000; Tourangeau 2018) are undertaken but less thoroughly than they might be. Strong satisficing occurs when one or more of these stages are skipped entirely. Examples of weak satisficing are



The selection of back or forward buttons in the MF interface



**Fig. 2.** The hypothesized process selecting between previous and next buttons in the MF interface.

acquiescence, where respondents show a tendency to agree with statements in attitude questions, selecting the midpoint on opinion questions with odd numbers of response options, and selecting the first reasonable option from a list rather than considering all options and selecting the most appropriate (Krosnick and Alwin 1987). The strong form of satisficing occurs for instance when respondents select “*Don’t Know*” when they could provide a substantive answer, when they select a substantive response option randomly or when they do not differentiate their responses on a battery of scale items (Hamby and Taylor 2016; Lipps 2007; Kaminska et al. 2011; Vannette and Krosnick 2014).

Non-differentiation—in other words, straightlining—a tendency to give the same answers across several items is a widely applied satisficing measure in survey methodology. It can be detected by Cronbach alpha, simple nondifferentiation, mean root of pairs, maximum identical rating, standard deviation of battery and scale point variation (Kim et al. 2019). Non-differentiation is more common among respondents with less education and low verbal ability and it is more common toward the end of a questionnaire than toward the beginning (Krosnick and Alwin 1988; Vannette and Krosnick 2014).

Survey satisficing is often suggested as occurring as a consequence of an excessive cognitive burden on respondents. Following this hypothesis Vannette and Krosnick (2014) suggest that in order to minimize the likelihood of satisficing, questionnaire designers should take steps to maximize respondent motivation and minimize task difficulty. This can be reached by making it easy for respondents to interpret questions, to retrieve information from memory, to integrate the information into a judgment, and to report the judgment (Vannette and Krosnick 2014). Regarding to measure the cognitive burden of respondents, response times are widely applied (Zhang and Conrad 2014). Their popularity in this respect is based on the assumption that if a survey question is in some way difficult or complex, it takes more time to answer due to greater thought and attention to determine a response (Turner et al. 2015).

In terms of satisficing, shorter item-level response times have been suggested to indicate sub-optimal responding or satisficing (Di et al. 2016; Zhang and Conrad 2014; Conrad et al. 2017). Zhang and Conrad (2014) showed that straightlining is significantly associated with speeding, a tendency to answer faster than is necessary in order to offer a response processed at least at the minimum level of cognitive effort. The findings of Callegaro et al. (2009) suggest support for these findings given that in their study the participants (job applicants) who inherently were expected holding stronger motivation towards the task, spent more time to accomplish the task than their counterparts, the less motivated participants. Thus, faster task accomplishment indicates satisficing task performance in their study as well. However, regarding the response time based satisficing measurements, Turner, Sturgis, and Martin (Turner et al. 2015) found the opposite results given that a higher proportion of “*Don’t Know*” answers and a tendency for rounding—typical satisficing measures—were associated with *longer* response latencies.

Using response latencies as a satisficing measure is not unproblematic given that the item level response times do not vary just because of the motivation and ability of respondents (Krosnick 1991; Vannette and Krosnick 2014) but other individual level factors as well. One of these factors is related with the information accessibility of respondents. It is well known that respondents with strong attitudes tend to offer their initial answer faster in attitude questions compared with their counterparts with less strongly attitudes (Fazio 1990). Regarding these respondents, shorter item level response times do not indicate satisficing but a tendency to answer faster due to more easily accessible information in their working memory. On the other hand when a question is complex and relevant information more difficult to recall, answering takes longer due to a greater attention and a cognitive process to formulate an answer (Turner et al. 2015; Yan and Tourangeau 2008). It is also well known that satisficing respondent behavior becomes more prevalent as responding proceeds towards the end of the questionnaire (Vannette and Krosnick 2014). What follows is that the item level response times should be decreased correspondingly towards the end of the questionnaire if shorter item-level response times are interpreted to indicate satisficing responding. On the other hand, as the responding proceeds towards the end of the questionnaire respondents become more fatigued due to accumulated cognitive burden. This should be detected as longer item level response times in case an appropriate amount of cognitive effort is invested in each item. However, when the respondents take shortcuts (satisficing) the item level response times should to decrease instead.

As discussed above, applying item level response times as a measure of satisficing is not as straightforward as suggested in the previous literature. This is mainly due to various factors affecting item level response times including individual level factors like motivation and ability or an information accessibility but also questionnaire level factors like a length of the questionnaire, an interface design in web surveys as well as question level factors like a substantial complexity of survey questions. It seems obvious that the interrelationship of these different sources affecting item level response times requires a clarification.

As the traditional survey satisficing theory is extended by the cognitive load theory, the interrelationship of its core elements; respondent ability, respondent motivation and task difficulty can be understood in a more advanced fashion. The starting point to

understand the relationship of these factors with an understanding how the cognitive load is generated is to realize that in case when the task difficulty is increased to a very high level generating correspondingly high cognitive load, the respondents inherently susceptible to satisficing respondent behavior tend to relieve an excessive load through satisficing. According to survey satisficing theory, these respondents are more likely less educated (individuals with lower abilities) and less motivated in comparison with their counterparts. At a more detailed level, the satisficing can be expected to occur more frequently in relation with more complex and cognitively demanding questions given that despite whether satisficing occurs at the conscious or unconscious level of responding it can be explained by an increased pressure to relieve an excessive cognitive load through it. This becomes most effectively executed regarding the most demanding questions generating the highest cognitive load because by taking shortcuts (skipping the entire stages of a response process) with regard to these questions the highest amount of excessive load can be relieved with minimum effort.

Regarding the less demanding questions in the survey, the satisficers can either be expected to follow the similar satisficing pattern than with the more complex questions or invest their major cognitive capacity in order to respond to these questions in particular. The latter option becomes more likely when the satisficers are truly less motivated individuals with lower abilities as the survey satisficing theory claims given that these types of individuals probably find the easiest questions the most convenient to answer. It should also be noted that from the perspective of an accumulation of cognitive load across the survey the less motivated respondents with lower abilities should become even more susceptible to satisficing respondent behavior towards the end of the questionnaire. Consequently, this tendency becomes even stronger when the individual questions are substantially complex or the web survey interface elements increase an extraneous load of the participants.

**Table 1.** The examples of satisficing responding patterns on item level response times.

| Items | load | A | B | C |
|-------|------|---|---|---|
| 1     | -    | ↓ | ↑ | ↕ |
| 2     | +    | ↘ | ↘ | ↘ |
| 3     | -    | ↓ | ↑ | ↕ |
| 4     | +    | ↘ | ↘ | ↘ |
| 5     | -    | ↓ | ↑ | ↕ |
| 6     | +    | ↘ | ↘ | ↘ |

*Note.* The direction of arrow represents an expected change in direction of item item level response time. The size of the arrow represents an expected change in the magnitude of the item level response time.

As a result of an extension of survey satisficing theory discussed above at least three satisficing responding patterns can be expected to occur under cognitively burdening

circumstances (Table 1). As expected in the traditional survey literature, shorter item level response times could occur evenly spread across the survey items as a consequence of increased satisficing responding. This satisficing pattern results shorter average item level response times given that all or most of the individual response latencies occur shorter in comparison with less burdening situation generating weaker satisficing respondent behavior. In Table 1 column A represents this satisficing responding pattern. However, as discussed above it is actually more likely that satisficers do not invest less response time consistently in each item across the survey but vary their responding pattern item by item. This becomes likely in particular when there are different types of questions on the survey in terms of their tendency to increase a cognitive load of respondents. The columns B and C represent these satisficing patterns. In both of these cases satisficers invest their major cognitive capacity to focus more carefully on cognitively less demanding questions. As a result, the item level response times do not decrease evenly across the items but decrease in terms of cognitively demanding questions and increase in terms of cognitively less demanding questions. As a result, the difference in average item level response times between satisficers and non-satisficers diminishes and may even disappear with regard to certain question combinations. However, what remains is converging item level response times in each of the illustrated satisficing responding patterns A, B and C. Thus, even though the difference in average item level response times cannot be treated as a reliable satisficing measure, the variation of item level response times appears to be reliable given that in all of the illustrated cases it would decrease. This makes the standard deviation (SD) of item level response times a more prominent measure of satisficing when an average item level response times should be treated as a secondary, a complementary measure of satisficing.

### 3 Subjects

In the first study (the University of Lapland, Finland) 3,023 undergraduate students were randomly assigned to six independent experimental conditions. This survey was fielded from October 7 to October 28, 2015. In the second study (the Lapland University of Applied Sciences) 5,004 undergraduate students were randomly assigned to six independent conditions following the same procedures as Study 1. This survey was fielded from April 18 to May 8, 2016. Respondents in both studies were encouraged to use PCs or tablets, but were not prevented from using a smartphone. The breakoffs and response rates of the aggregated data of two samples are shown in Appendix 3. The final data was applied in the present study in combining the automatically forwarded (AF) and manually forwarded (MF) groups resulting in two groups; AF ( $n = 863$ ) and MF ( $n = 900$ ).

## 4 Method

### 4.1 Measuring Response Times

Both client-side and server-side paradata were captured, and response time was measured at both the respondent (survey) level and the item level (Yan and Olson 2013). The total response time (TRT) for a survey for a particular respondent was calculated by taking

the difference between the first and last time stamps in the survey. **Item-level response times were also calculated as the difference between mouse clicks on two radio buttons or between the mouse clicks of the forward/backward button and a radio button. This is a measure of the time to select an initial response for an item after a page has loaded. It does not include the time following this selection** (i.e., the time taken to change an answer or to click the next button in the MF version).

In order to avoid the drawbacks in measuring satisficing and to develop a methodological solution taking into account the individual-level influence we use a standard deviation (SD) of item-level response times (calculated within the individuals) as a measure of satisficing. We interpret a decreased standard deviation of item-level response times indicating increased satisficing. This becomes intelligible when realizing that substantially different items require different amount of cognitive effort to become comprehended, the necessary information retrieved, an appropriate judgment completed, and a response given. As the standard deviation of item-level response times decreases, it reveals respondents investing response time in a more similar fashion in different types of items. This suggests an increased total cognitive load occurring across the items further relieved through satisficing responding. Empirically it can be recognized as a decreased standard deviation. Given that as we expect the MF to be associated with a more cognitively burdening procedure due to the previous and next buttons, we should observe lower SD in the MF group in comparison with the AF group.

However, because the standard deviation of item level response times is affected by an individual tendency to respond slower or faster (respondent baseline speed), we took an average item level response times of individuals account within the examined question batteries as a nuisance variable. In other words, the individual tendency to respond slower or faster was removed from the estimate of standard deviation by modelling it as an independent variable. Given that the original relationship of an average response speed of individuals and the SD appeared to be curvilinear in order to take this into account we were able to achieve more accurate estimates as well as compare the three question batteries including different types of items.

We applied the standard deviation (SD) of item-level response times as a measure of satisficing in several log-linear regression models. The approach was conducted within three question batteries separately including the items varying in their expected tendency to increase intrinsic cognitive load. In order to control the confounded influence of average respondent-level response speed on SD we added it in the models as an explanatory variable with the dummy variable “AF/MF”. To make the explanatory variables uncorrelated we group-mean centered the values of the average respondent-level response speed within the AF/MF groups (see Bell et al. 2018; Dalal and Zickar 2012; Enders and Tofghi 2007; Paccagnella 2006). We excluded individuals from the analysis whose average item-level response times exceeded an upper outer fence in the boxplot<sup>2</sup>. The model is:

$$\log Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

<sup>2</sup> When answering web surveys some respondents leave a question “open” for a long time because they interrupt answering the survey due to a variety of reasons. On that specific screen or question, the time latency is very high and therefore such cases need to be excluded from the analysis.

$Y$  = The log of standard deviation of item-level response times

$X_1$  = An average respondent-level response speed based on item-level response times

$X_2$  = AF/MF.

## 5 Results

The examination of the first question battery (4 items) revealed (Table 2) that one unit change in the “AF/MF” - variable ( $X_2$ ) is associated with a 16% decrease ( $p < 0.001$ ) in the expected geometric mean of the SD indicating that we expect to see about a 16% decrease in the geometric mean of the standard deviation of item-level response times for the MF group ( $n = 863$ ) compared with the AF group ( $n = 846$ ). The corresponding decrease regarding the second question battery (9 items; AF,  $n = 855$ ; MF,  $n = 881$ ) was 1% (non-significant) when for the third battery (4 items; AF,  $n = 842$ ; MF,  $n = 873$ ), it was 6% (non-significant;  $p = 0.108$ ). A one-unit change in an average respondent-level response speed ( $X_1$ ) was associated with 0.02% to 0.03% increase in SD in all three question batteries.

**Table 2.** An average respondent-level response speed and AF/MF associated with log of standard deviation of item-level response times.

|                    | $\beta$   | Exp. $\beta$ | 95% CI    | $p$    |
|--------------------|-----------|--------------|-----------|--------|
| <i>Items 2–5</i>   |           |              |           |        |
| RT                 | 0.00026   | 1.00         | 1.00–1.00 | <0.001 |
| AF/MF              | – 0.17439 | 0.84         | 0.78–0.91 | <0.001 |
| <i>Items 6–14</i>  |           |              |           |        |
| RT                 | 0.00022   | 1.00         | 1.00–1.00 | <0.001 |
| AF/MF              | – 0.01189 | 0.99         | 0.94–1.04 | n.s    |
| <i>Items 15–18</i> |           |              |           |        |
| RT                 | 0.00015   | 1.00         | 1.00–1.00 | <0.001 |
| AF/MF              | – 0.06626 | 0.94         | 0.86–1.02 | n.s    |

*Note.* RT = An average respondent-level response time. Exp. $\beta$  - exponentiated coefficient represents the ratio of expected geometric mean difference in the standard deviation of item-level response times for a 1-unit change in an explanatory factor. Maximum likelihood estimation.

The results are consistent with the theoretical expectations given that the largest difference (16%) in the expected geometric mean of SD between the AF and MF groups was found in the first question battery (items 2–5) including the most demanding questions. The second largest difference (6%) was found in the third question battery (items 15–18), including cognitively demanding attitude questions. The smallest (non-significant) difference (1%) was found in the second question battery (items 6–14), including the easiest items to answer; the mood items. The direction of the observed effects occurs

as expected, suggesting support for the hypothesis that as the total cognitive load of respondents increases, the standard deviation of item-level response times decreases. Additionally, the results suggest indirect support for the hypothesized eye-movements in the examined interfaces given that in the MF group a number of hypothesized eye-movements to navigate through the interface is bigger (4) than in the AF group (3) due to the previous and next buttons (Fig. 1). As a result, the total cognitive load of manually forwarded respondents tends to exceed their working memory capacity especially when the content of the questions (the first and third question batteries) with the required eye movements to navigate through the interface generates a high intrinsic load. In an interaction with a generated extraneous load due to the previous and next buttons the total cognitive load of respondents in the MF condition increases to a very high level exceeding their working memory capacity. As a consequence, the respondents relieve the excessive load through satisficing responding which was detected by a decreased standard deviation of item level response times.

In the second question battery, the total cognitive load of respondents is smaller due to the content of items. What follows is that the working memory capacity is not exceeded even in the MF group despite the extraneous load generated by the previous and next buttons. Empirically this was observed as a very small difference in standard deviation between the AF and MF groups. One should note however that this finding indicates nothing with regard to the overall level of satisficing across the mood items. It reveals only that the generated cognitive load of the respondents between the AF and MF groups does not differ from enough in order to lead the second of the groups relieving it through satisficing respondent behavior.

Despite the fact that satisficing respondent behavior is in other studies found to occur more frequently towards the end of the questionnaire we did not find support for these findings regarding the standard deviation. Despite the burdening content of the questions in the third response battery, the decrease in standard deviation was smaller in the MF group (6%) than corresponding difference in the first question battery (16%). One possible explanation for this is that the questionnaire was quite short, and the respondents were university students with high cognitive skills.

On the other hand, the differences between the average item-level response times in the AF and MF groups (Table 3) support the previous findings of the satisficing occurrence towards the end of the questionnaire given that the differences in average item-level response times between the AF and MF groups occurred the smallest in the first question battery, the second largest in the second question battery and the largest in the third question battery. The corresponding decrease in percentages in MF group compared with AF group were in the first question battery 4.2%, in the second question battery 5.9% and in the third question battery 7.3%. This suggests the likelihood of satisficing responding increasing in MF group towards the end of the questionnaire if the difference in average item-level response times in comparison with AF group is considered as a measure of satisficing.

## 6 Conclusion

In the perspective of a human-computer interaction, a major challenge becomes how to manage a cognitive burden of the participants which is also interpreted as increasing

**Table 3.** The differences in average item-level response times between the AF and MF groups in milliseconds.

|             | AF   | MF   | AF – MF | <i>t</i> | <i>p</i> |
|-------------|------|------|---------|----------|----------|
| Items 2–5   | 3885 | 3720 | 165     | 1.932    | <0.05    |
| Items 6–14  | 4509 | 4241 | 268     | 3.801    | <0.001   |
| Items 15–18 | 6827 | 6328 | 499     | 3.855    | <0.001   |

*Note.* The mean values are the average item-level response times calculated separately within the three question batteries based on the combined 3 groups allocated to AF and the three groups allocated to MF. Extreme outliers excluded based on Tukey's method. One-tailed t-tests, equal variances not assumed.

satisficing in the survey literature. In order to estimate satisficing between the different types of interfaces (AF/MF) we take into account that satisficers do not necessarily invest response time consistently in each item across the survey. Consequently, we introduce a more elaborated method, the standard deviation of item level response times, to measure satisficing and estimate the amount of total cognitive load of respondents.

We find support for our hypothesis that the MF version increases satisficing responding given that MF reduces both average item level response times as well as the standard deviation of item-level response times. This suggests support also for the hypothesis that the MF generates higher total cognitive load of respondents due to a more complex interface design. On the other hand, AF has shown to reduce completion times across the whole questionnaire (Selkälä and Couper 2018). This is consistent with our findings given that a shorter completion time allows the respondent to focus on individual items more carefully which is observed as a longer average item level response times. The findings reinforce the view in HCI that reducing the complexity of interfaces and the presence of extraneous elements reduces cognitive load and facilitates the concentration of cognitive resources on the task at hand.

To test these ideas further, we need longer questionnaires to detect fatigue effects towards the end of the questionnaire. We also need questions with varying levels of cognitive effort. We also need eye-tracking studies to determine the effect of the back and the next button on cognitive load. In addition, should be noted that the evidence is based on a relatively short survey among university students. Replication in other settings is recommended. We caution that this applies primarily to highly repetitive tasks, such as answering a series of questions using the same response format (as is often the case in the batteries of psychological measures). The effect may be different in surveys that vary the format and content of items. In other words, AF may be suitable for some types of survey questionnaires, but maybe not others. Further research is needed to figure out under what survey conditions AF is optimal.



## Appendix 1. The Questionnaires of the Experiment

| Conditions: 1(MF) & 2(AFM)<br>(control groups)   | Conditions: 3(MF) & 4(AFM)<br>(low accessible information)  | Conditions: 5(MF) & 6(AFM)<br>(consistency requirement)   |
|--|---|---|
| <i>The first file (items 1-5)</i>  | <i>The first file</i>   | <i>The first file</i>   |
| How many hours a week do you do each of the following sporting activities?<br>(The response options: ½ hour or less, 1 hour, 1½ hours, 2 hours, 2½ hours, 3 hours, 3½ hours, 4 hours or more)    |   |   |
| 1. Going to the gym  | 1. Going to the gym   | 1. Going to the gym   |
| 2. Horse sport   | 2. Jogging or walking   | 2. Horse sport  |
| 3. Team sports (e.g. football, ice hockey, volleyball)   | 3. Team sports (e.g. football, ice hockey, volleyball)  | 3. Team sports (e.g. football, ice hockey, volleyball)  |
| 4. Swimming  | 4. Swimming   | 4. Swimming   |
| 5. Tennis  | 5. Tennis   | 5. All sports put together  |
| <i>The second file (items 6-14)</i>  | <i>The second file</i>  | <i>The second file</i>  |
| During the past 4 weeks, how often you...<br>(The response options: All of the time; Most of the time; Some of the time; A little bit of the time; None of the time)                             |   |   |
| 6. Have been happy?  | 6. Have been happy?   | 6. Have been happy?   |
| 7. Have felt that you can reach all the goals you have set for yourselves?   | 7. Have felt that you can reach all the goals you have set for yourselves?                            | 7. Have felt that you can reach all the goals you have set for yourselves?                            |
| 8. Have felt yourself as energetic   | 8. Have felt yourself as energetic  | 8. Have felt yourself as energetic  |
| 9. Have felt downhearted   | 9. Have felt downhearted  | 9. Have felt downhearted  |
| 10. Have felt calm and peaceful  | 10. Have felt calm and peaceful   | 10. Have felt calm and peaceful   |
| 11. Have felt full of pep  | 11. Have felt full of pep   | 11. Have felt full of pep   |
| 12. Have been nervous  | 12. Have been nervous   | 12. Have been nervous   |
| 13. Have felt so down in the dumps that nothing could cheer you up   | 13. Have felt so down in the dumps that nothing could cheer you up                                    | 13. Have felt so down in the dumps that nothing could cheer you up                                    |
| 14. Have felt tired  | 14. Have felt tired   | 14. Have felt tired   |
| <i>The third file (items 15-18)</i>  | <i>The third file</i>   | <i>The third file</i>   |
| Please say to what extent you agree or disagree with each of the following statements.<br>(The response options: Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree) |   |   |
| 15. Exercising has health benefits   | 15. Exercising increases body's ability to buffer lactic acid   | 15. Exercising has health benefits  |
| 16. Exercising becomes more difficult for students with low income than for students with high income  | 16. Exercising becomes more difficult for students with low income than for students with high income | 16. Exercising becomes more difficult for students with low income than for students with high income |
| 17. I can easily afford chargeable sporting activities   | 17. I can easily afford chargeable sporting activities  | 17. I can easily afford chargeable sporting activities  |
| 18. Students would benefit in many ways by increasing student grants   | 18. Students would benefit in many ways by increasing student grants                                  | 18. Increasing student grants would increase the time I will spend in chargeable sporting activities  |

## Appendix 2. Screen Shots of ZEF Interface

### a. Manually forwarded ZEF survey on PC.

Arvioi kuinka monta tuntia viikossa käytät seuraavien liikuntamuotojen harrastamiseen?

- Kuntosali
- Hevosurheilu
- Joukkuelajit (esim. jalkapallo, jääkiekko, lentopallo)
- Uinti
- Tennis

Kuinka usein viimeisen 4 viikon kuluessa olet...

Mitä mieltä olet seuraavista väittämistä?

0% 388

1. Kuntosali

ZEF

- ½ tuntia tai vähemmän
- 1 tunti
- 1½ tuntia
- 2 tuntia
- 2½ tuntia
- 3 tuntia
- 3½ tuntia
- 4 tuntia tai enemmän

Edellinen Seuraava

### b. Automatically forwarded ZEF survey on PC.

Arvioi kuinka monta tuntia viikossa käytät seuraavien liikuntamuotojen harrastamiseen?

- Kuntosali
- Hevosurheilu
- Joukkuelajit (esim. jalkapallo, jääkiekko, lentopallo)
- Uinti
- Tennis

Kuinka usein viimeisen 4 viikon kuluessa olet...

Mitä mieltä olet seuraavista väittämistä?

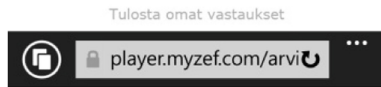
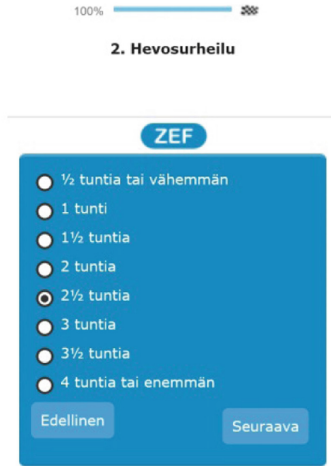
0% 388

1. Kuntosali

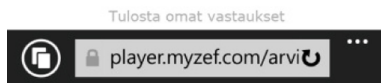
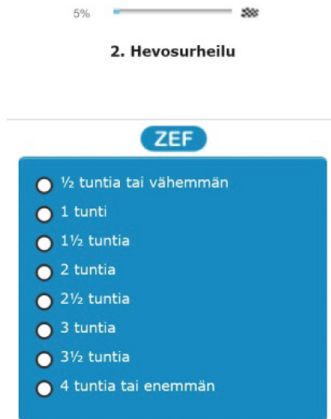
ZEF

- ½ tuntia tai vähemmän
- 1 tunti
- 1½ tuntia
- 2 tuntia
- 2½ tuntia
- 3 tuntia
- 3½ tuntia
- 4 tuntia tai enemmän

**c. Manually forwarded ZEF survey on smartphone.**

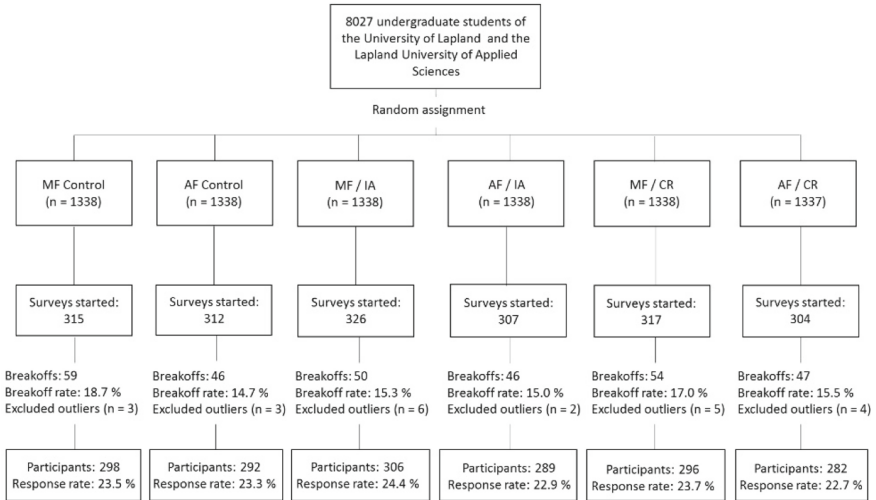


**d. Automatically forwarded ZEF survey on smartphone.**



## Appendix 3. Experimental Design of the Original Study

(Selkälä and Couper 2018)



*Note.* MF = Manual forwarding; AF = automatic forwarding; IA = information accessibility; CR = consistency requirement. The response rates are calculated by following the RR2 standard definition (AAPOR 2008, pp. 34, 48; Bethlehem and Biffignandi 2011, 439). The breakoff rates are calculated by the QBR definition (Callegaro et al. 2015) representing the combined proportion of breakoffs and partially completed questionnaires of all respondents starting the questionnaire.

## References

- Antonenko, P.D., Keil, A.: Assessing working memory dynamics with electroencephalography. Implications for research on cognitive load. In: Zheng, R.Z. (ed.) *Cognitive Load Measurement and Application: A Theoretical Framework for Meaningful Research and Practice*, 1st edn. Routledge, New York (2018)
- Ayres, P.: Subjective measures of cognitive load. What can they reliably measure? In: Zheng, R.Z. (ed) *Cognitive Load Measurement and Application: A Theoretical Framework for Meaningful Research and Practice*, 1st edn. Routledge, New York (2018)
- Ayres, P., van Gog, T.: State of the art research into cognitive load theory. *Comput. Hum. Behav.* **25**, 253–257 (2009). <https://doi.org/10.1016/j.chb.2008.12.007>
- Bandilla, W., Couper, M.P., Kaczmirek, L.: The mode of invitation for web surveys. *Surv. Pract.* **5**, 1–5 (2012). <https://doi.org/10.29115/sp-2012-0014>
- Beege, M., Wirzberger, M., Nebel, S., Schneider, S., Schmidt, N., Rey, G.D.: Spatial continuity effect vs. spatial contiguity failure. Revising the effects of spatial proximity between related and unrelated representations. *Front. Educ.* **4**, 86 (2019). <https://doi.org/10.3389/educ.2019.00086>
- Bell, A., Jones, K., Fairbrother, M.: Understanding and misunderstanding group mean centering: a commentary on Kelley et al.'s dangerous practice. *Qual. Quant.* **52**, 2031–2036 (2018)

- Callegaro, M., Yang, Y., Bhola, D.S., Dillman, D.A., Chin, T.-Y.: Response latency as an indicator of optimizing in online questionnaires. *Bull. Méthodologie Sociologique* **103**, 5–25 (2009)
- Callegaro, M., Lozar Manfreda, K., Vehovar, V.: *Web Survey Methodology*. Sage Publications, London (2015)
- Cannell, C.F., Miller, P.V., Oksenberg, L.: Research on interviewing techniques. *Sociol. Methodol.* **12**, 389–437 (1981)
- Chelazzi, L., Miller, E.K., Duncan, J., Desimone, R.: A neural basis for visual search in inferior temporal cortex. *Nature* **363**, 345–347 (1993)
- Clark, R.C., Mayer, R.E.: *e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning*, 4th edn. Wiley, Hoboken (2016)
- Conrad, F.G., Tourangeau, R., Couper, M.P., Zhang, C.: Reducing speeding in web surveys by providing immediate feedback. *Surv. Res. Methods* **11**, 45–61 (2017)
- Couper, M.P.: *Designing Effective Web Surveys*. Cambridge University Press, New York (2008)
- Couper, M.P., Peterson, G.: Why do web surveys take longer on smartphones? *Soc. Sci. Comput. Rev.* **35**, 357–377 (2016)
- Couper, M.P., Traugott, M.W., Lamias, M.J.: Web survey design and administration. *Public Opin. Q.* **65**, 230–253 (2001)
- Dalal, D.K., Zickar, M.J.: Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organ. Res. Methods* **15**, 339–362 (2012)
- Davern, M.: Nonresponse rates are a problematic indicator of nonresponse bias in survey research. *Health Serv. Res.* **48**, 905–912 (2013)
- de Bruijne, M., Das, M., van Soest, A., Wijnant, A.: Adapting grid questions on mobile devices. Paper presented at the European Survey Research Association Conference, Reykjavik, July 2015
- de Leeuw, E.D., Hox, J.J., Klausch, T., Roberts, A., de Jongh, A.: Design of web questionnaires: matrix questions or single question formats. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Orlando, FL, May 2012
- de Rada, V.D., Dominguez-Alvarez, J.A.: Response quality of self-administered questionnaires: a comparison between paper and web questionnaires. *Soc. Sci. Comput. Rev.* **32**, 256–269 (2014). <https://doi.org/10.1177/0894439313508516>
- Di, J.: Determining survey satisficing of online longitudinal survey data in the multicenter AIDS cohort study: a group-based trajectory analysis. *J. Med. Internet Res. Public Health Surveill.* **2**, 1–10 (2016)
- Dillman, D.: Web-push surveys; origins, uses and unsolved challenges. 2019 JPSM distinguished lecture, University of Maryland, 12 April 2019
- Enders, C.K., Tofighi, D.: Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychol. Methods* **12**, 121–138 (2007)
- Fazio, R.H.: Multiple processes by which attitudes guide behavior: the mode model as an integrative framework. *Adv. Exp. Soc. Psychol.* **23**, 75–109 (1990)
- Giroux, S., Tharp, K., Wietelman, D.: Impacts of implementing an automatic advancement feature in mobile and web surveys. *Surv. Pract.* **12**, 1–12 (2019). <https://doi.org/10.29115/sp-2018-0034>
- Groves, R.M., Peytcheva, E.: The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opin. Q.* **72**, 167–189 (2008)
- Hamby, T., Taylor, W.: Survey satisficing inflates reliability and validity measures: an experimental comparison of college and amazon mechanical turk samples. *Educ. Psychol. Measur.* **76**, 912–932 (2016). <https://doi.org/10.1177/0013164415627349>
- Hammen, K.: The impact of visual and functional design elements in online survey research. Paper presented at the General Online Research Conference, Mannheim, Germany, 26–28 May 2010 (2010)

- Hays, R.D., Bode, R., Rothrock, N., Riley, W., Cella, D., Gershon, R.: The impact of next and back buttons on time to complete and measurement reliability in computer-based surveys. *Qual. Life Res.* **19**, 1181–1184 (2010)
- Heideman, S.G., Rohenkohl, G., Chauvin, J.J., Palmer, C.E., van Ede, F., Nobre, A.C.: Anticipatory neural dynamics of spatial-temporal orienting of attention in younger and older adults. *Neuroimage* **178**, 46–56 (2018). <https://doi.org/10.1016/j.neuroimage.2018.05.002>
- Hollender, N., Hofmann, C., Deneke, M., Schmitz, B.: Integrating cognitive load theory and concepts of human-computer interaction. *Comput. Hum. Behav.* **26**, 1278–1288 (2010)
- Kaminska, O., McCutcheon, A.L., Billiet, J.: Satisficing among reluctant respondents in a cross-national context. *Public Opin. Q.* **74**, 956–984 (2011)
- Kaplowitz, M.D., Lupi, F., Couper, M.P., Thorp, L.: The effects of invitation design on web survey response rates. *Soc. Sci. Comput. Rev.* **30**, 339–349 (2012)
- Kastner, S., Pinsk, M.A., De Weerd, P., Desimone, R., Ungerleider, L.G.: Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron* **22**, 751–761 (1999)
- Kim, Y., Dykema, J., Stevenson, J., Black, P., Moberg, D.P.: Straightlining: overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Soc. Sci. Comput. Rev.* **37**, 214–233 (2019). <https://doi.org/10.1177/0894439317752406>
- Kim, S.-U., Lim, S.-M., Kim, E.-A., Yang, I.-H.: An analysis of eye movement and cognitive load about the editorial design in elementary science textbook. *Univ. J. Educ. Res.* **4**, 570–575 (2016). <https://doi.org/10.13189/ujer.2016.040314>
- Klausch, T., de Leeuw, E.D., Hox, J.J., Roberts, A., de Jongh, A.: Matrix vs. single question formats in web surveys: results from a large scale experiment. Paper presented at the General Online Research Conference, Mannheim, 5–7 March 2012 (2012)
- Knowles, E.S.: Item context effects on personality scales: measuring changes the measure. *J. Pers. Soc. Psychol.* **55**, 312–320 (1988)
- Krejtz, K., Duchowski, A.T., Niedzielska, A., Biele, C., Krejtz, I.: Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE* (2018). <https://doi.org/10.1371/journal.pone.0203629>
- Krosnick, J.A.: Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl. Cogn. Psychol.* **5**, 213–236 (1991)
- Krosnick, J.A.: Survey research. *Annu. Rev. Psychol.* **50**, 537–567 (1999)
- Krosnick, J.A., Alwin, D.F.: An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opin. Q.* **51**, 201–219 (1987). <https://doi.org/10.1086/269029>
- Krosnick, J.A., Alwin, D.F.: A test of the form-resistant correlation hypothesis: ratings, rankings, and the measurement of values. *Public Opin. Q.* **52**, 526–538 (1988)
- Leppink, J., Paas, F., Van der Vleuten, C.P.M., Van Gog, T., Van Merriënboer, J.J.G.: Development of an instrument for measuring different types of cognitive load. *Behav. Res. Methods* **45**(1058), 1072 (2013). <https://doi.org/10.3758/s13428-013-0334-1>
- Lipps, O.: Interviewer and respondent survey quality effects in a CATI panel. *Bull. Méthodologie Sociologique* **95**, 5–25 (2007)
- Makransky, G., Terkildsen, T.S., Mayer, R.E.: Role of subjective and objective measures of cognitive processing during learning in explaining the spatial contiguity effect. *Learn. Instr.* **61**, 23–34 (2019). <https://doi.org/10.1016/j.learninstruc.2018.12.001>
- Mayer, R.E.: Cognitive theory of multimedia learning. In: Mayer, R.E. (ed.) *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press, New York (2014)
- Mayer, R.E.: Instruction based on visualizations. In: Mayer, R.E., Alexander, P.A. (eds.) *Handbook of Research on Learning and Instruction*, 2nd edn. Routledge, New York (2017)

- Mayer, R.E., Fiorella, L.: Principles for reducing extraneous processing in multimedia learning: coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In: Mayer, R.E. (ed.) *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press, New York (2014)
- Mayer, R.E., Moreno, R.: Nine ways to reduce cognitive load in multimedia learning. *Educ. Psychol.* **38**, 43–52 (2003)
- Nobre, A.C., Stokes, M.G.: Premembering experience: a hierarchy of time-scales for proactive attention. *Neuron* **104**, 132–146 (2019)
- Paas, F., Renkl, A., Sweller, J.: Cognitive load theory and instructional design: Recent developments. *Educ. Psychol.* **38**, 1–4 (2003)
- Paas, F., Sweller, J.: An evolutionary upgrade of cognitive load theory: using the human motor system and collaboration to support the learning of complex cognitive tasks. *Educ. Psychol. Rev.* **24**, 27–45 (2012). <https://doi.org/10.1007/s10648-011-9179-2>
- Paas, F., van Gog, T., Sweller, J.: Cognitive load theory: new conceptualizations, specifications, and integrated research perspectives. *Educ. Psychol. Rev.* **22**, 115–121 (2010)
- Paccagnella, O.: Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. *Eval. Rev.* **30**, 66–85 (2006)
- Rivers, D.: Web surveys for health measurement. Paper presented at Building Tomorrow's Patient-Reported Outcome Measures: The Inaugural PROMIS Conference, Gaithersburg, MD, September 2006
- Selkälä, A., Couper, M.P.: Automatic versus manual forwarding in web surveys. *Soc. Sci. Comput. Rev.* **36**, 669–689 (2018). <https://doi.org/10.1177/0894439317736831>
- Selkälä, A., Reips, U.-D., Viinamäki, L., Suikkanen, A.: Demographics explaining a web survey entries election on the postal invitation letter. In: Conference Presentation, Session: Challenges and Opportunities of Switching to Web, Zagreb, Croatia, 15–19 July 2019. European Survey Research Association (2019)
- Simon, H.A.: *Models of Man*. Wiley, New York (1957)
- Stokes, M., Thompson, R., Nobre, A.C., Duncan, J.: Shape-specific preparatory activity mediates attention to targets in human visual cortex. *PNAS* **106**, 19569–19574 (2009)
- Sundar, S.S.: Social psychology of interactivity in human-website interaction. In: Joinson, A.N., McKenna, K., Postmes, T., Reips, U.-D. (eds.) *Oxford Handbook of Internet Psychology*. Oxford University Press, New York (2007)
- Theeuwes, J.: Spatial orienting and attentional capture. In: Nobre, A.C., Kastner, S. (eds.) *The Oxford Handbook of Attention*. Oxford University Press, New York (2014). <https://doi.org/10.1093/oxfordhb/9780199675111.013.005>
- Toepoel, V., Das, M., Van Soest, A.: Design of web questionnaires: the effects of the number of items per screen. *Field Methods* **21**, 200–213 (2009)
- Tourangeau, R.: Cognitive sciences and survey methods. In: Jabine, T., Straf, M., Tanur, J., Tourangeau, R. (eds.) *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. National Academy Press, Washington, DC (1984)
- Tourangeau, R.: The survey response process from a cognitive viewpoint. *Qual. Assur. Educ.* **26**, 169–181 (2018)
- Tourangeau, R., Conrad, F.G., Couper, M.P.: *The Science of Web Surveys*. Oxford University Press, New York (2013)
- Tourangeau, R., Rips, L.J., Rasinski, K.A.: *The Psychology of Survey Response*. Cambridge University Press, Cambridge (2000)
- Turner, G., Sturgis, P., Martin, D.: Can response latencies be used to detect survey satisficing on cognitively demanding questions? *J. Surv. Stat. Methodol.* **3**, 89–108 (2015)
- van Merriënboer, J.J.G., Kester, L., Paas, F.: Teaching complex rather than simple tasks: balancing intrinsic and germane load to enhance transfer of learning. *Appl. Cogn. Psychol.* **20**, 343–352 (2006). <https://doi.org/10.1002/acp.1250>

- van Merriënboer, J.J.G., Sweller, J.: Cognitive load theory and complex learning: recent developments and future directions. *Educ. Psychol. Rev.* **17**, 147–177 (2005)
- Vannette, D.L., Krosnick, J.A.: Answering questions: a comparison of survey satisficing and mindlessness. In: Ie, A., Ngnoumen, C.T., Langer, E.J. (eds.) *The Wiley Blackwell Handbook of Mindfulness*. Wiley, Chichester (2014). <https://doi.org/10.1002/9781118294895.ch17>
- Whitenton, K.: Minimize cognitive load to maximize usability (2013). <https://www.nngroup.com/articles/minimize-cognitive-load/>
- Yan, T., Olson, K.: Analyzing paradata to investigate measurement error. In: Kreuter, F. (ed.) *Improving Surveys with Paradata. Analytic Uses of Process Information*. Wiley, New York (2013)
- Yan, T., Tourangeau, R.: Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. *Appl. Cogn. Psychol.* **22**, 51–68 (2008). <https://doi.org/10.1002/acp.1331>
- Zagermann, J., Pfeil, U., Reiterer, H.: Measuring cognitive load using eye tracking technology in visual computing. In: *BELIV 2016: Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pp. 78–85 (2016). <https://doi.org/10.1145/2993901.2993908>
- Zhang, C., Conrad, F.G.: Speeding in web surveys: the tendency to answer very fast and its association with straightlining. *Surv. Res. Methods* **8**, 127–135 (2014)
- Zhang, C., Conrad, F.G.: Intervening to reduce satisficing behaviors in web surveys: evidence from two experiments on how it works. *Soc. Sci. Comput. Rev.* **36** (2018). <https://doi.org/10.1177/0894439316683923>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.


The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.







# A New Information Theory Based Clustering Fusion Method for Multi-view Representations of Text Documents

Juan Zamora<sup>1</sup>(✉)  and Jérémie Sublime<sup>2,3</sup> 

<sup>1</sup> Instituto de Estadística, Pontificia Universidad Católica de Valparaíso, 2340025 Valparaíso, Chile

[juan.zamora@pucv.cl](mailto:juan.zamora@pucv.cl)

<sup>2</sup> ISEP, Lisite Laboratory – DaSSIP Team, 10 rue de Vanves, 92130 Issy-Les-Moulineaux, France

[jeremie.sublime@isep.fr](mailto:jeremie.sublime@isep.fr)

<sup>3</sup> LIPN - CNRS UMR 7030, University Paris 13, 99 Avenue J.-B. Clément, 93430 Villetaneuse, France

[sublime@lipn.univ-paris13.fr](mailto:sublime@lipn.univ-paris13.fr)

**Abstract.** Multi-view clustering is a complex problem that consists in extracting partitions from multiple representations of the same objects. In text mining and natural language processing, such views may come in the form of word frequencies, topic based representations and many other possible encoding forms coming from various vector space model algorithms. From there, in this paper we propose a clustering fusion algorithm that takes clustering results acquired from multiple vector space models of given documents, and merges them into a single partition. Our fusion method relies on an information theory model based on Kolmogorov complexity that was previously used for collaborative clustering applications. We apply our algorithm to different text corpuses frequently used in the literature with results that we find to be very satisfying.

**Keywords:** Multi-view clustering · Information theory · Corpus analysis

## 1 Introduction

The goal of text corpus clustering is to partition a collection of text documents into several groups, such that texts inside the same groups (or clusters) are similar and share common themes or have a common style, while documents in different clusters are very distinct in nature. To achieve this goal, text documents must first be transformed using models such as the Vector Space Model (VSM) [20] in order to transform the original documents into numerical vectors that can be used by clustering algorithms such as K-Means or hierarchical clustering. One difficulty with the VSM model is the large number of existing methods to transform text documents into vector representations. Many representation

models exist, some are topic oriented, others focus on word embedding, while some methods are purely statistical representations. This abundance of methods in the literature allows for multiple vector representations of the same texts, all with different strengths and weaknesses. Applying clustering algorithms to these multiple representations can be seen as a multi-view clustering problem where the goal could be to find a consensus between the clustering partitions proposed under the various Vector Space Models [3]. Within this context, in this paper we propose a new method inspired from collaborative clustering, and which relies on the notion of Kolmogorov complexity to merge clustering partitions acquired from the clustering algorithms applied to different vector representations of text documents. Our proposed method is compared with state of the art methods applied to common text corpora that can be found in the literature.

The remainder of this paper is organized as follows: Sect. 2 focuses on various related works about both text mining and the information theory model used in this paper. Section 3 presents our algorithm. Section 4 features our experimental results and some comparisons with other methods. Finally, in Sect. 5 we draw some conclusion and give some ideas on possible extensions of this work.

## 2 State of the Art

Cluster ensembles is an overall framework in which multiple partitions are combined in order to obtain a consensus clustering. Multi-view clustering is one of the specific problems covered in this area [5].

The problem of combining multiple data partitions into a single one has been tackled at least by two approaches, namely Clustering Ensembles [4, 9, 10, 16, 21, 22, 24, 26] and Multi-View Clustering [3, 6–8, 11–13, 18, 25, 27], also known as data fusion.

In ensemble learning and ensemble clustering, several algorithms will work on the same data set with the goal of achieving a single result that should be better than the partitions learned from the different algorithms. As one can see, in ensemble clustering, several algorithms work on the same data and therefore the same view. However, in the case of multi-view clustering like in the present work, we have several algorithms and each of them works in a different view of the same data. And since we are dealing with several views, the goal with multi-view clustering is to merge them while taking into considerations that there might be multiple truths [30].

It is worth mentioning, that the distinction between ensemble clustering and multi-view clustering is not always obvious in the literature and some confusion may exist with different naming conventions depending on the field of application. In the following subsection, we make a quick review of the literature for both multi-view and ensemble clustering with a particular focus on text mining applications and methods that are close to the one presented in this work.

## 2.1 State of the Art on Combining Multiple Clustering Partitions

There are many different applications that require to combine multiple clustering partitions: In [9], the authors make a proposal for music clustering using partitions obtained from different music feature sets. Among these sets, they employ several word-level features. They pose the ensemble clustering problem as a binary clustering in a space induced by the multiple partitions. Additionally, they explore various optimization criteria for finding consensus partitions and propose a strategy for determining the final number of clusters. It is interesting to note that they apply this proposal for

In [7], the authors work specifically on text clustering. They propose to generate several partitions from each view by using different feature representations and then applying a clustering algorithm over each one. Then, similarity matrices are computed in three different ways, namely two based on partition memberships and another one based on feature similarity. Finally, a combined similarity matrix is obtained from those three previous ones and a standard clustering technique is applied to produce the consensus partition. In the same direction, [3] use more diverse text representations as views, more specifically LDA [1], Word2Vec [14] and TF-IDF [19] and then apply the same idea as the former work.

It is worth noting that multi-view text clustering shouldn't be confused with distributed clustering of texts [28], which mainly consists in distributing the clustering task without consideration for whether or not this is a multi-view task.

Another common application of multi-view clustering is multilingual clustering. In [18], for this specific application, the authors pose the multi-view clustering problem as a tensor decomposition as this approach was proven earlier to be theoretically efficient [11, 12].

## 2.2 Methods to Combine Multiple Partitions

In [21], the authors pose that an application of Cluster Ensembles is to combine partitions obtained from partial sets of features. As we have seen earlier, this is a case of multi-view clustering. Additionally, they pose that a motivation for using a cluster ensemble is to build a more robust solution that performs well over a wide range of data sets. Since the diversity of base partitions has a positive impact on the final consensus solution, it can be introduced mainly by using different sets of features in each partition, different parameter configurations of the same algorithm (values of  $k$  for k-means) and also using different and complementary base techniques. The authors also formulate consensus clustering as a hyper-graph cutting problem and solve in three different ways.

Co-association matrices are based on relative co-occurrence of two data points in the same cluster. They are another very common tool to tackle multi-view clustering. Several works exploit them in order to produce final partitions from several combinations of different data representations. [4] explore two strategies

for producing cluster ensembles: Using different views and using different clustering algorithms or parameter configurations. [26] address the problem from a similarity matrix completion problem in which missing values are associated to uncertain data pairs, this is pair of data points whose common membership in every partition is not consistent. In the same path, [16] propose to weight the contribution of each co-association matrix based on a novel reliability measure of each partition within the ensemble.

Some other contributions employ an utility function to measure similarity between partitions and then directly maximize an objective function to obtain the consensus [10, 22, 24].

In [8], a hybrid clustering method based on weighted linear combination of distance matrices for textual and bibliometric information is proposed.

### 2.3 Multi-view Clustering Applications and Kolmogorov Complexity

In the work of [17, 23], the notion of minimum description length (MDL) is introduced, with the *description length* being the minimal number of bits needed by a Turing Machine to describe an object. This measure of the minimal number of bits is also known under the name Kolmogorov complexity.

If  $\mathcal{M}$  is a fixed Turing machine, the complexity of an object  $\mathbf{x}$  given another object  $\mathbf{y}$  using the machine  $\mathcal{M}$  is defined as  $K_{\mathcal{M}}(\mathbf{x}|\mathbf{y}) = \min_{p \in \mathcal{P}_{\mathcal{M}}} \{l(p) : p(\mathbf{y}) = \mathbf{x}\}$  where  $\mathcal{P}_{\mathcal{M}}$  is the set of programs on  $\mathcal{M}$ ,  $p(\mathbf{y})$  designates the output of program  $p$  with argument  $y$  and  $l$  measures the length (in bits) of a program. When the argument  $\mathbf{y}$  is empty, we use the notation  $K_{\mathcal{M}}(\mathbf{x})$  and call this quantity the complexity of  $\mathbf{x}$ . The main problem with this definition is that the complexity depends on a fixed Turing machine  $\mathcal{M}$ . Furthermore, the universal complexity is not computable, since it is defined as a minimum over all programs of all machines.

In relation with this work, in [15], the authors solved the aforementioned problem by using a fixed Turing Machine before applying this notion of Kolmogorov complexity to collaborative clustering, which is a specific case of multi-view clustering where several clustering algorithms work together in a multi-view context but aim at improving each other partitions rather than merging them [2]. While collaborative clustering does not aim at a consensus, this application is still very close to what we try to achieve in this paper where we try to merge partitions of the same objects under multiple representations. For these reasons, we decided to use the same tool.

In the rest of this paper, just as the authors did in [15], we will consider that the Turing Machine  $\mathcal{M}$  is fixed, and to make the equations easier we will denote by  $K(\mathbf{x})$  the complexity of  $\mathbf{x}$  on the chosen machine. Then, we adapt the equations used in their original paper to our multi-view context for text mining and we use Kolmogorov complexity as a tool to compute the complexity of one partition given another partition. The algorithm to do so and how we use it is described in the next section.

### 3 Proposed Merging Method

#### 3.1 Problem Definition

Let us consider a data set  $\mathcal{X}$  of  $n$  data points and a measure of similarity  $S$  that allows to quantify the strength of the connection or closeness between any pair of data points in  $\mathcal{X}$ . The problem of data clustering can be stated as inducing an equivalence relation<sup>1</sup> on  $\mathcal{X}$  such that points  $\mathbf{a}, \mathbf{b}$  in the same equivalence class (that is a cluster) have a larger similarity value  $S(\mathbf{a}, \mathbf{b})$  in comparison with  $S(\mathbf{a}, \mathbf{c})$  or  $S(\mathbf{b}, \mathbf{c})$  for any other point  $\mathbf{c}$  in a different equivalence class.

The Multi-view clustering task considers that the information regarding to each data point in  $\mathcal{X}$  comes from multiple sources called views. After performing a clustering algorithm over each view several partitions are generated. Let us define this set of partitions as  $\mathcal{P}$ , and denote each of them with a capital letter (e.g.:  $A$ ).

A partition  $A$  is a set of  $|A|$  disjoint sets  $\mathbf{c} \in \wp(\mathcal{X})$  (the Power set of  $\mathcal{X}$ ) called clusters of  $\mathcal{X}$ . Let us define an agreement function  $\Omega$  between two clusters as a mapping  $\Omega : \wp(\mathcal{X}) \times \wp(\mathcal{X}) \rightarrow [0, 1]$  which attains lower values for clusters having a smaller overlap and higher values for clusters sharing more elements of  $\mathcal{X}$ . In this work we employ the Jaccard similarity function to measure agreement between two clusters.

For a point  $\mathbf{p} \in \mathcal{X}$ , its cluster in any partition  $A \in \mathcal{P}$  is denoted by  $\mathcal{N}_{\mathbf{p}}^A$  and it is defined as:

$$\mathcal{N}_{\mathbf{p}}^A = \{\mathbf{x} \in \mathcal{X} | \exists \mathbf{c} \in A \wedge \mathbf{p} \in \mathbf{c} \wedge \mathbf{x} \in \mathbf{c}\}$$

Given a cluster  $\mathbf{c}$  and a partition  $B$  the function that maps  $\mathbf{c}$  to the cluster in  $B$  with the largest overlap is called maximum agreement function and it is defined as follows:

$$\Phi_B(\mathbf{c}) = \underset{\mathbf{e} \in B}{\operatorname{argmax}} \Omega(\mathbf{c}, \mathbf{e}) \quad (1)$$

#### 3.2 The Algorithm

Our goal in this paper is to combine several partitions in order to build a final consensus. To this end, in our method we perform successive pairwise fusion procedures between partitions following a bottom-up strategy until we reach a single partition. This procedure is depicted in Algorithm 1.

Without loss of generality, when a fusion step is performed between two partitions  $A$  and  $B$ , a new partition  $C$  is created. Since the successive partition fusions are performed by following the maximum agreement criteria between clusters as stated in Eq. (1), it is possible that some data points do not fit to this rule and hence be marked as exceptions during the execution of the merge operation. The set of data points marked as exceptions before the creation of partition  $C$  is denoted by  $\xi_C$ , formally,

$$\xi_C = \{\mathbf{p} \in \mathcal{X} | \mathcal{N}_{\mathbf{p}}^A \cap \Phi_B(\mathcal{N}_{\mathbf{p}}^A) = \emptyset \cup \mathcal{N}_{\mathbf{p}}^B \cap \Phi_A(\mathcal{N}_{\mathbf{p}}^B) = \emptyset\} \quad (2)$$

<sup>1</sup> For the clustering task, the relation could be stated as “has the same label as”.

Thus, when partition  $C$  is created, each point  $\mathbf{p} \in \xi_C$  receives a weight  $W_C(\mathbf{p}, \mathbf{c})$  for every cluster  $\mathbf{c} \in C$ . This weight is made up by the relative weights that both source partitions  $A$  and  $B$  contribute, namely  $\omega_A(\mathbf{p}, \mathbf{c})$  and  $\omega_B(\mathbf{p}, \mathbf{c})$ . Without loss of generality, the contribution of each source partition is given by:

$$\omega_A(\mathbf{p}, \mathbf{c}) = \begin{cases} \Omega(\mathbf{c}, \mathcal{N}_{\mathbf{p}}^A) & \text{if } \mathbf{p} \notin \xi_A \\ \Omega(\mathbf{c}, \Phi_A(\mathbf{c})) \cdot W_A(\mathbf{p}, \Phi_A(\mathbf{c})) & \text{if } \mathbf{p} \in \xi_A \end{cases} \quad (3)$$

Thus, the final weight  $W_C(\mathbf{p}, \mathbf{c})$  for each point  $\mathbf{p} \in \xi_C$  in each cluster  $\mathbf{c} \in C$  is given by:

$$W_C(\mathbf{p}, \mathbf{c}) = \frac{\omega_A(\mathbf{p}, \mathbf{c})}{2} + \frac{\omega_B(\mathbf{p}, \mathbf{c})}{2}$$

A more detailed description of this merging process is depicted in Algorithm 2. It is important to indicate that once a point is marked as an exception, it remains so through all the subsequent fusions. After the last fusion, each of these exception data points are assigned to one of the final clusters by picking the one whose membership weight is the highest. This exception resolution is described between lines 7–9 in Algorithm 1 where  $K(A|B)$  is the kolmogorov complexity of partition A knowing partition B [15]:

$$K(A|B) = K_B \times (\log K_A + \log K_B) + |\xi_C| \times (\log n + \log K_A) \quad (4)$$

with  $n$  the total number of points,  $K_A$  the number of clusters in partition  $A$ ,  $K_B$  the number of clusters in partition  $B$  and  $\xi_C$  the set of exceptions between partitions  $A$  and  $B$  as defined in Eq. (2).

---

**Algorithm 1:** Main procedure for building the consensus partition.

---

**Input:** A set  $\mathcal{P}$  of  $m$  partitions over the data  $\mathcal{X}$ .

**Output:** A consensus partition.

```

1  $\mathcal{Q} \leftarrow \square$  /* exceptions after each merge operation */
2 while  $|\mathcal{P}| > 1$  do
3    $A, B \leftarrow \operatorname{argmin}_{A^*, B^* \in \mathcal{P}} K(A^*|B^*) + K(B^*|A^*)$ 
4    $C \leftarrow \operatorname{merge}(A, B, \mathcal{Q}, W)$ 
5   add  $C$  into  $\mathcal{P}$ 
6   remove  $A, B$  from  $\mathcal{P}$ 
/* Solving points marked in last item from  $\mathcal{Q}$  */
7  $\xi_D \leftarrow$  last partition's exceptions added to  $\mathcal{Q}$ 
8 foreach  $\mathbf{p} \in \xi_D$  do
9    $\mathcal{N}_{\mathbf{p}}^D \leftarrow \operatorname{argmax}_{\mathbf{c} \in D} W_D(\mathbf{p}, \mathbf{c})$ 
10 return  $D$ 

```

---

---

**Algorithm 2: Merge** procedure that fuses two partitions into a new one identifying also troublesome points as exceptions.

---

**Input:** Partitions  $A, B \in \mathcal{P}$  s.t.  $|A| > |B|$   
**1** , list with previous merge exceptions  $\mathcal{Q}$  and weight function for previously created partitions  $W$   
**Output:** New partition  $C$  and a set of marked points along with their scores  $\forall c \in C$ .

```

2  $\mathcal{M} \leftarrow \emptyset$ 
3 foreach  $\mathbf{a} \in A$  do
4    $\lfloor$  add  $\Phi_B(\mathbf{a})$  into  $\mathcal{M}[\mathbf{a}]$ 
5 foreach  $\mathbf{b} \in B$  do
6    $\lfloor$  add  $\mathbf{b}$  into  $\mathcal{M}[\Phi_A(\mathbf{b})]$ 
7  $C \leftarrow \emptyset$  /* The new partition to be returned */
8 foreach  $\mathbf{a} \in A$  do
9    $\mathbf{c} \leftarrow \emptyset$ 
10   foreach  $\mathbf{b} \in \mathcal{M}[\mathbf{a}]$  do
11      $\mathbf{c} \leftarrow \mathbf{c} \cup (\mathbf{a} \cap \mathbf{b})$ 
12      $\mathbf{a}' \leftarrow \mathbf{a}$ 
13      $\mathbf{a} \leftarrow \mathbf{a} - \mathbf{b}$  /* updates cluster a */
14      $\mathbf{b} \leftarrow \mathbf{b} - \mathbf{a}'$  /* updates cluster b */
15    $\lfloor$  add  $\mathbf{c}$  into  $C$ 
/* generating the list of marked points by the current fusion */
16  $\xi_C \leftarrow \emptyset$  foreach  $\mathbf{a} \in A$  do
17   if  $|\mathbf{a}| > 0$  then
18      $\lfloor$  add each  $\mathbf{p} \in \mathbf{a}$  into  $\xi_C$ 
19     foreach  $\mathbf{b} \in \mathcal{M}[\mathbf{a}]$  s.t.  $|\mathbf{b}| > 0$  do
20        $\lfloor$  add each  $\mathbf{p} \in \mathbf{b}$  into  $\xi_C$ 
21 add  $\xi_C$  into  $\mathcal{Q}$ 
22 foreach  $\mathbf{p} \in \xi_C$  and  $\mathbf{c} \in C$  do
23    $\lfloor$   $W_C(\mathbf{p}, \mathbf{c}) = \frac{\omega_A(\mathbf{p}, \mathbf{c})}{2} + \frac{\omega_B(\mathbf{p}, \mathbf{c})}{2}$ 
24 return  $C$ 

```

---

## 4 Experimental Results

### 4.1 Experimental Settings

Since external class labels are available for each data set, let us consider the true clustering  $T$  and the final partition obtained by the clustering algorithm as  $C$ . Then, two measures are employed to assess the quality of a clustering solution, namely Entropy and Purity. Entropy is defined in two parts: the former allows to measure the Entropy for a single partition and it is characterized for any partition  $\mathbf{c} \in C$  in Eq. (5). The latter is just a weighted average of the entropy computed for all the partitions in the final solution and it is defined in Eq. (6).

Purity is defined in a similar way, that is first the Purity of a single partition is defined in Eq. (7) and then, the overall Purity of the partition is denoted as Eq. (8).

$$E(\mathbf{c}) = -\frac{1}{\log |T|} \sum_{\mathbf{t} \in T} \frac{|\mathbf{c} \cap \mathbf{t}|}{|\mathbf{c}|} \log \frac{|\mathbf{c} \cap \mathbf{t}|}{|\mathbf{c}|} \quad (5)$$

$$\text{Entropy}(C) = \sum_{\mathbf{c} \in C} \frac{|\mathbf{c}|}{n} E(\mathbf{c}) \quad (6)$$

$$P(\mathbf{c}) = \frac{1}{|\mathbf{c}|} \max_{\mathbf{t} \in T} |\mathbf{c} \cap \mathbf{t}| \quad (7)$$

$$\text{Purity}(C) = \sum_{\mathbf{c} \in C} \frac{|\mathbf{c}|}{n} P(\mathbf{c}) \quad (8)$$

Entropy measures the degree in which the true classes are dispersed within each cluster. A good solution is the one that does not break the true clusters into too many parts. Purity is targeted to measure the extent to which each cluster contains documents from mostly a single true class. Thus, a good solution should present homogeneous clusters in terms of the true classes of the contained documents.

Since the quality of the overall solution depends on the initial source  $k$ -Means clusterings, which in turn have a random nature, we follow the scheme presented in [29] to eliminate some of this sensitivity in the performance assessment. This is, we use several values for  $k$  and for each specific value, the overall clustering procedure is repeated 10 times and the best performance solution is kept. Additionally, since partition quality improves as the number of clusters increases, relative performances are reported for each clustering solution. To compute the relative entropy, we divide the entropy attained by a particular solution by the smallest entropy for that particular data set and value of  $k$ . In case of relative purity and in order to allow the same interpretation of the relative entropy, we divide the best Purity attained for that particular data set and value of  $k$  by the entropy value obtained by the clustering solution under evaluation. Since these two ratios represent the extent to which a specific algorithm performed worse than the best algorithm, for each dataset better solutions are closer to 1.0 and they are worse as they become greater than 1.0. Finally, as a performance summary for each solution the average relative performance across all data sets are reported for each clustering solution.

## 4.2 Results and Interpretations

The result Tables 1, 2, 3 and 4 show the relative performances attained by the proposal, each source clustering and another ensemble method recently proposed in [3].

As we can see from Tables 1 and 2, the results on the relative entropy show that our proposed method achieves significantly better results than the method of Fraj et al. [3] on the same data sets.



**Table 1.** Average relative entropy

| k  | LDA   | skipgram | tfidf        | fracj | proposal     |
|----|-------|----------|--------------|-------|--------------|
| 5  | 1.737 | 1.509    | <b>1.031</b> | 2.271 | <b>1.075</b> |
| 10 | 1.565 | 1.447    | <b>1.019</b> | 2.045 | <b>1.074</b> |
| 15 | 1.727 | 1.453    | <b>1.021</b> | 2.239 | <b>1.057</b> |

**Table 2.** Relative entropy

| k  | dataset     | lda          | skipgram | tfidf        | fracj | proposal     |
|----|-------------|--------------|----------|--------------|-------|--------------|
| 5  | WebKB       | <b>1.000</b> | 1.111    | 1.008        | 1.289 | <b>1.012</b> |
| 5  | 20Newsgroup | 1.306        | 1.256    | <b>1.000</b> | 1.813 | <b>1.116</b> |
| 5  | BBCSport    | 3.641        | 1.948    | <b>1.000</b> | 4.397 | <b>1.144</b> |
| 5  | Reuters-R8  | <b>1.000</b> | 1.722    | 1.116        | 1.586 | <b>1.028</b> |
| 10 | WebKB       | <b>1.000</b> | 1.216    | 1.074        | 1.256 | <b>1.021</b> |
| 10 | 20Newsgroup | 1.147        | 1.310    | <b>1.000</b> | 1.653 | <b>1.023</b> |
| 10 | BBCSport    | 3.105        | 1.628    | <b>1.000</b> | 3.541 | <b>1.191</b> |
| 10 | Reuters-R8  | 1.010        | 1.632    | <b>1.000</b> | 1.727 | <b>1.059</b> |
| 15 | WebKB       | <b>1.000</b> | 1.246    | 1.082        | 1.282 | <b>1.012</b> |
| 15 | 20Newsgroup | 1.210        | 1.375    | <b>1.000</b> | 1.784 | <b>1.205</b> |
| 15 | BBCSport    | 3.689        | 1.642    | <u>1.000</u> | 4.185 | <u>1.000</u> |
| 15 | Reuters-R8  | 1.009        | 1.548    | <b>1.000</b> | 1.707 | <b>1.010</b> |

Going more into details, from Table 2 we can see that overall the TFIDF first and the LDA view second have the best results in term en entropy and are used as baseline for the relative entropy. We can see that for many data set our proposed method not only is close from the best entropy result, but that it achieves better results on average than the 3 original lda, skipgram and tfidf views, and always better results than the method from Fraj et al.

Since each view may hold its own truth, it is only logical that we rarely achieve fusion results that are better than all original view. This is a common problem in multi-view clustering [30] and should be considered as normal. Regardless, it is worth mentioning that our proposed method still achieves the

**Table 3.** Average relative purity

| k  | lda   | skipgram | tfidf        | fracj | proposal     |
|----|-------|----------|--------------|-------|--------------|
| 5  | 1.129 | 1.083    | <b>1.001</b> | 1.446 | <b>1.020</b> |
| 10 | 1.112 | 1.094    | <b>1.006</b> | 1.283 | <b>1.010</b> |
| 15 | 1.119 | 1.096    | <b>1.000</b> | 1.263 | <b>1.020</b> |

**Table 4.** Relative purity

| k  | dataset     | lda          | skipgram | tfidf        | fracj | proposal     |
|----|-------------|--------------|----------|--------------|-------|--------------|
| 5  | WebKB       | 1.006        | 1.037    | <b>1.000</b> | 1.344 | <b>1.011</b> |
| 5  | 20Newsgroup | 1.087        | 1.024    | <b>1.000</b> | 1.513 | <b>1.045</b> |
| 5  | BBCSport    | 1.422        | 1.139    | <b>1.000</b> | 1.822 | <b>1.017</b> |
| 5  | Reuters-R8  | <b>1.000</b> | 1.131    | 1.005        | 1.106 | <b>1.005</b> |
| 10 | WebKB       | 1.020        | 1.110    | <b>1.020</b> | 1.200 | <b>1.000</b> |
| 10 | 20Newsgroup | 1.049        | 1.121    | <b>1.000</b> | 1.226 | <b>1.026</b> |
| 10 | BBCSport    | 1.353        | 1.066    | <b>1.000</b> | 1.586 | <b>1.013</b> |
| 10 | Reuters-R8  | 1.027        | 1.077    | <b>1.005</b> | 1.119 | <b>1.000</b> |
| 15 | WebKB       | 1.013        | 1.132    | <b>1.000</b> | 1.209 | <b>1.027</b> |
| 15 | 20Newsgroup | 1.053        | 1.098    | <b>1.000</b> | 1.279 | <b>1.054</b> |
| 15 | BBCSport    | 1.368        | 1.045    | <b>1.000</b> | 1.436 | <b>1.000</b> |
| 15 | Reuters-R8  | 1.042        | 1.109    | <b>1.000</b> | 1.127 | <b>1.001</b> |

best results in the case of the BBCSport data set with 15 clusters in terms of relative entropy.

From Tables 3 and 4, we can see that the results in term of purity are the same than the one we had with entropy, thus enabling us to affirm that our proposed method proved superior than the one of Fraj et al. on all data sets regardless of the number of clusters.

Like for entropy, we can see that we rarely achieve the best results among views, but that we still do better than the average of the 3 original views, and from Table 3 we can see that our algorithm remains very competitive even when compare to the best view.

The best performances of our proposed algorithm for relative purity are for the BBCSport data set with 15 clusters, Reuters-R8 with 15 clusters and WebKB with 10 clusters. For all 3 cases, we not only get better results than other methods in the literature, but we also do better than the best views in term of relative purity.

## 5 Conclusion and Future Works

We have presented a new clustering fusion method applied to the case of multi-view text corpus clustering. Our method was applied to 4 data sets that are very common in the literature (20Newsgroup, Reuters-R8, WebKB and BBCSport) and has proved to be competitive with state of the art methods. Unlike previously proposed methods, our algorithm relies on the notion of Kolmogorov complexity and information compression thus giving it a solid theoretical background on how to best fusion the clustering partitions.

In our future works, we plan on coupling our proposed method with existing collaborative method so that we could have a collaborative step first, and a

merging step then. We hope that doing so may help to detect incompatible or noisy views, but could also ease the merging process by creating closer partition with collaborative clustering before hand. Other possible extensions of this work include applications on merging multi-view clustering partitions in fields other than text mining and natural language processing.

## References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**(Jan), 993–1022 (2003)
2. Cornuéjols, A., Wemmert, C., Gançarski, P., Bennani, Y.: Collaborative clustering: why, when, what and how. *Inf. Fusion* **39**, 81–95 (2018)
3. Fraj, M., HajKacem, M.A.B., Essoussi, N.: Ensemble method for multi-view text clustering. In: *Computational Collective Intelligence - 11th International Conference, ICCCI 2019, Hendaye, France, 4–6 September 2019, Proceedings, Part I*, pp. 219–231 (2019). [https://doi.org/10.1007/978-3-030-28377-3\\_18](https://doi.org/10.1007/978-3-030-28377-3_18)
4. Fred, A.L., Jain, A.K.: Combining multiple clusterings using evidence accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(6), 835–850 (2005)
5. Ghosh, J., Acharya, A.: Cluster ensembles. *Wiley Interdisc. Rev.: Data Min. Knowl. Discov.* **1**(4), 305–315 (2011)
6. Greene, D., Cunningham, P.: A matrix factorization approach for integrating multiple data views. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) *ECML PKDD 2009. LNCS (LNAI)*, vol. 5781, pp. 423–438. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-04180-8\\_45](https://doi.org/10.1007/978-3-642-04180-8_45)
7. Hussain, S.F., Mushtaq, M., Halim, Z.: Multi-view document clustering via ensemble method. *J. Intell. Inf. Syst.* **43**(1), 81–99 (2014). <https://doi.org/10.1007/s10844-014-0307-6>
8. Janssens, F., Glänzel, W., De Moor, B.: Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 360–369. ACM (2007)
9. Li, T., Ogihara, M., Ma, S.: On combining multiple clusterings. In: *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pp. 294–303. ACM (2004)
10. Liu, H., Zhao, R., Fang, H., Cheng, F., Fu, Y., Liu, Y.Y.: Entropy-based consensus clustering for patient stratification. *Bioinformatics* **33**(17), 2691–2698 (2017)
11. Liu, X., Glänzel, W., De Moor, B.: Hybrid clustering of multi-view data via Tucker-2 model and its application. *Scientometrics* **88**(3), 819–839 (2011). <https://doi.org/10.1007/s11192-011-0348-3>
12. Liu, X., Ji, S., Glänzel, W., De Moor, B.: Multiview partitioning via tensor methods. *IEEE Trans. Knowl. Data Eng.* **25**(5), 1056–1069 (2012)
13. Liu, X., Yu, S., Moreau, Y., De Moor, B., Glänzel, W., Janssens, F.: Hybrid clustering of text mining and bibliometrics applied to journal sets. In: *Proceedings of the 2009 SIAM International Conference on Data Mining*, pp. 49–60. SIAM (2009)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013)
15. Murena, P., Sublime, J., Matei, B., Cornuéjols, A.: An information theory based approach to multisource clustering. In: *IJCAI*, pp. 2581–2587. *ijcai.org* (2018)

16. Rashidi, F., Nejatian, S., Parvin, H., Rezaie, V.: Diversity based cluster weighting in cluster ensemble: an information theory approach. *Artif. Intell. Rev.* **52**, 1341–1368 (2019)
17. Rissanen, J.: Modeling by shortest data description. *Automatica* **14**(5), 465–471 (1978)
18. Romeo, S., Tagarelli, A., Ienco, D.: Semantic-based multilingual document clustering via tensor modeling (2014)
19. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **24**(5), 513–523 (1988)
20. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. *Commun. ACM* **18**(11), 613–620 (1975). The paper where vector space model for IR was introduced
21. Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**(Dec), 583–617 (2002)
22. Topchy, A., Jain, A.K., Punch, W.: Clustering ensembles: models of consensus and weak partitions. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(12), 1866–1881 (2005)
23. Wallace, C.S., Boulton, D.M.: An information measure for classification. *Comput. J.* **11**(2), 185–194 (1968). <https://doi.org/10.1093/comjnl/11.2.185>
24. Wu, J., Liu, H., Xiong, H., Cao, J., Chen, J.: K-means-based consensus clustering: a unified view. *IEEE Trans. Knowl. Data Eng.* **27**(1), 155–169 (2014)
25. Xie, X., Sun, S.: Multi-view clustering ensembles. In: International Conference on Machine Learning and Cybernetics, ICMLC 2013, Tianjin, China, 14–17 July 2013, pp. 51–56 (2013). <https://doi.org/10.1109/ICMLC.2013.6890443>
26. Yi, J., Yang, T., Jin, R., Jain, A.K., Mahdavi, M.: Robust ensemble clustering by matrix completion. In: 2012 IEEE 12th International Conference on Data Mining, pp. 1176–1181. IEEE (2012)
27. Yu, S., Moor, B., Moreau, Y.: Clustering by heterogeneous data fusion: framework and applications. In: NIPS Workshop (2009)
28. Zamora, J., Allende-Cid, H., Mendoza, M.: Distributed clustering of text collections. *IEEE Access* **7**, 155671–155685 (2019)
29. Zhao, Y., Karypis, G.: Criterion functions for document clustering: experiments and analysis. Department of Computer Science, University of Minnesota, Technical Report TR 01-40 (2001)
30. Zimek, A., Vreeken, J.: The blind men and the elephant: on meeting the problem of multiple truths in data from clustering and pattern mining perspectives. *Mach. Learn.* **98**(1–2), 121–155 (2015). <https://doi.org/10.1007/s10994-013-5334-y>



# Application of Visual Saliency in the Background Image Cutting for Layout Design

Liyu Zhu<sup>1</sup>, Xueni Cao<sup>1</sup>, Ying Fang<sup>1</sup>, Liqun Zhang<sup>1</sup>(✉), and Xiaodong Li<sup>2</sup>(✉)

<sup>1</sup> School of Design, Shanghai Jiao Tong University, Shanghai, China  
zhanglliqun@gmail.com

<sup>2</sup> China National Gold Group Gold Jewellery Co., Ltd., Beijing, China  
lixiaodong@chnau99999.com

**Abstract.** In many poster designs, an image usually will be used as a back-ground image, and text and picture will be carried out on the background image later. For intelligent layout design, cropping a suitable background image should be the first problem to be solved. In this paper, through eye movement experiments, ground truth saliency maps of the posters are obtained. Then, the characteristics of the saliency maps of background images are summarized. The characteristics are mainly the rules of the location and size of the salient areas in the background image. The research found that the salient areas of the poster background images are more concentrated in the upper and middle of the poster image, and they are distributed in an inverted triangle. These rules can cut a more suitable background image for typesetting.

**Keywords:** Layout design · Background image · Image cutting · Saliency map

## 1 Background

With the rapid development of deep learning in the field of images processing, intelligent design has played an increasingly important role in contemporary design activities. Computers can replace manual work to complete more complicated work, thereby liberating designers and enabling designers to do more creative work. For example, the Luban developed by Alibaba's Intelligent Design Lab has changed the traditional design mode, and automatically generates visual image designs that meet requirements and standards through user input of styles and sizes.

Many existing automatic layout generation technologies use photographic images as the background in their typesetting process. However, their technologies mainly focus on how to select the suitable font size of the text, the reasonable position of the text, and so on. Few scholars have studied how to better cut images with a specific role, for example, as a background in typographic design. For the existing automatic layout process, the cutting of the background image is the first step of the entire process, which is inconspicuous but important. Although there are mature techniques for image cropping, the image cropping seems to be the last step in related research. The main purpose of these studies is to cut the content of an image intact, or to cut it to have an aesthetic sense.

However, from the view of a complete design activity, there may be other tasks after cutting the image. This is not considered in existing research. These methods are often applied to the cropping of photographic photos or thumbnails, but for layout design, images cropped with existing cropping techniques are not necessarily applicable.

When people watch an image or a real scene, they will identify areas of interest to their own, so that their brains will ignore areas they are not interested in during further advanced visual processing, reducing the complexity. This is the attention mechanism of the human. In the computer field, computers mimic human attention mechanisms to better detect visual saliency. The study of visual saliency is the basis of other computer vision problems. Saliency detection has a wide range of applications. After successfully detecting saliency, the computer can further identify the content in the image or scene, and then complete some more intelligent tasks, such as image segmentation, text detection, face recognition, image cropping and so on.

As a background image, what should be its visual saliency? How do people pay attention to background images in typography? This is the main problem of the research in this article. Secondly, this article also hopes to summarize some rules. By applying these rules to the cutting of background images in typography, the images that are more suitable for typography can be cut out. Computers can serve more specific designs in the future. By researching more suitable cutting methods, the automatically generated layout design will produce better results.

The research of visual saliency and the application of visual saliency in cutting images are introduced in Sect. 2. Section 3 describes the specific experimental process in detail. Section 4 analysis the data and certain rules are obtained. The conclusion is summarized in Sect. 5. Finally, the future plan is discussed in Sect. 6.

## 2 Related Works

### 2.1 Development and Application of Visual Saliency

Visual saliency is the ability of the visual system (whether human or machine) to select a subset of visual information for further processing. This mechanism serves as a filter to select interesting information related to the current behavior or task and ignore extraneous information [1].

Human visual attention models can be divided into “bottom-up” and “top-down” models. The “bottom-up” saliency model is data-driven and affected by the contrast of image elements such as pixels and blocks with neighboring areas. Also, a “bottom-up” saliency mode considers the uniqueness of each block in overall image [2]. While “top-down” saliency models are task-driven and require the prior knowledge. And human complex inferential cognitive processes, psychological activities and subjective emotions are all needed. Many researches in related fields of psychology study “top-down” saliency models.

The earliest saliency models can be traced back to the work of Itti et al. [3]. Their models combined the cognitive theory of psychology with the early computational models, which triggered the first wave of research on visual saliency. Later, more scholars began to study how to build saliency models to predict gaze points in order to better understand the human visual attention mechanism. The second boom originated from

the research by Liu et al. [4] and Achanta et al. [5], who considered the saliency detection as a binary segmentation problem with foreground pixels 1 and background pixels 0, and thus the saliency detection was opened the boundary with computer vision research [6].

In fact, visual saliency has already been applied in the design field. Bylinskii et al. [7] proposed a saliency calculation model for image design and data visualization through research, and analyzed the application possibility of the saliency model, including applying the saliency model to interactive design applications that can be fed back in real time, to generating thumbnails and to the redesign of charts and so on. Jahanian et al. [8] analyzed the saliency of the cover image and adjusted the layout of the cover text to better achieve the visual balance. Ali et al. [9] calculated the features affecting visual balance by collecting tens of thousands of high-quality aesthetic photos.

## 2.2 Application of Visual Saliency in Cutting

The existing methods for automatically cutting images can be divided into two categories, one is based on attention, and the other is based on aesthetics. The aesthetic-based method is more in line with the photographer's composition principle. Therefore, the current mainstream method is aesthetic-based cutting.

The main idea of attention-based method is to keep the most salient areas in the image, that is, the most relevant parts in the photo, and crop out other unrelated parts. The importance of each pixel in the image is determined by the saliency. The saliency of the image mainly comes from the saliency distribution map, human eye tracking. Chen et al. [10] studied the problem of image cropping earlier and proposed an image adaptive method based on user's attention to facilitate users to view images on different displays. This model was based on the three attributes of region of interest, attention value, and minimal perceptible size. They used the Itti model to calculate the pixel saliency value combined with the face and text detection and finally generates a salient map. Suh et al. [11] later used the sum of saliency values within the clipping rectangle to determine the optimal clipping position and generated a thumbnail. Santella [12] obtained saliency maps by acquiring user fixation data, combining image segmentation results, identifying important image content, and calculating the optimal cutting amount. Marchesotti et al. [13] performed image cropping by training a classifier using a labeled image saliency database. Chen et al. [14] explored different search algorithms for optimal image cropping based on saliency frames.

Aesthetic-based image cropping method improve the aesthetic value of images by cutting method. The most important point of this type of method is to train a classifier to determine the score of the cropped image or the label of "beauty" or "not beautiful" by extracting the features of the image. Cutting is the first step of this type of method. After generating candidate regions, a classifier or search algorithm is used to find the optimal cropping region.

However, for cutting background image of the layout design, its aesthetic degree is not the main purpose of the cut. Due to the requirements that the background image need to be typeset, the image as a background may have its own characteristics. Existing image cutting techniques uses the composition rules and aesthetic principles of traditional photography, but the results are often independent and complete, which do not may meet

the requirements of the background. Therefore, in this paper, the salient characteristics of the background image are mainly discussed. The characteristics obtained from the study then will be used as the main basis for cutting the background image.

### 2.3 “Figure-Ground” Relationship

The “figure-ground” relationship is a basic visual phenomenon that comes from Gestalt perception theory. Edgar Rubin started the study of the relationship from the perspective of psychology. In his research, he pointed out that people tend to highlight a part of the observed things as a same object as the “figure” and the rest as the ground. Rudolf Arnheim then systematically applied it to the visual arts. He believed that the “figure” often appears as a closed, smaller area in “Art and Visual Perception”. British scholar E.H. Gombrich in his book “Art and Illusion” proposed that people pay attention to certain graphics due to their own experience. The “figure” in image often has the characteristics of “representation”, “complete”, “small area”, “clear outline” and so on. The “ground” in a picture often has the characteristics of “neglected” and “fuzzy appearance”.

The relationship of the “figure-ground” in the traditional sense and in this study may not be exactly the same. In this article, this concept is used to distinguish the background image in the poster from some elements arranged on the background image better. However, the above-explained “figure-ground” relationship is more described from the perspective of human visual cognition and psychology. In this article, the “figure-ground” relationship is mainly understood from the operational level of typographic design activity.

## 3 Experimental

Many scholars have done experiments related to visual saliency. Borji et al. [15] asked the subjects to manually select and segment more significant objects to finally obtain a ground true saliency map. Xu et al. [16] labeled the salient objects in the picture, and then analyzed the impact of high-level information (objects, semantics) on attention. Koehler et al. [17] divided the experiment into three parts, allowing the subjects to freely watch, search for significance, and prompt the search of objects.

The experiment is designed to obtain the ground true salient map of the poster. In this study, posters are collected from various fields as test samples for the layout design background map of this study. Its content includes electronics, food, cosmetics, stationery, daily chemical, etc. to exclude the top-down interference caused by the participants due to different experiences and habits. In addition, based on the life experience of the participants, the text of the control test posters is all in Chinese.

Then, the posters more match the study were selected. The screening requirements are: 1. Posters should have obvious “figure-ground” relationships, that is, posters with complex effects or inconspicuous background images should be discarded. 2. The background of the poster should not be a solid color or an inconspicuous gradient or a regular shading. 3. The posters should be fully designed, that is, posters with logos in the corners of the layout, which only reflect the image content, should be discarded. Because the



cutting of such poster background images is the same as the cutting of pure images. After screening, there are 185 qualified posters.

Then the eye-movement experiments were performed on the screened posters to obtain the areas of interest and fixation points. Each picture was tested by 15 participants, and the test poster was divided into 4 groups. A total of 60 participants were tested. In the course of the experiment, the subjects browsed at will without any action, and the test poster was automatically played on the screen. Participants were 18–25 years old. During the experiment, adjust the resolution of all test samples to 1280 \* 1024, and fill the blank area with a dark gray bottom. In this study, an eye tracker model Tobii T60 was used to perform eye movement experiments in a bright room 60 cm from the screen. Each test sample was stared at by the subject for 3 s.

## 4 Data Analysis

All the 185 test samples can be divided into 2 categories according to the function of the background image. The type 1 of background image contains the main information of the poster, such as the main product, publicity portrait, etc., the area containing the content is more important, and generally cannot be blocked. The type 2 of background image mainly plays the role of accentuating the atmosphere. There is no main content in the background image. Most of them are landscape scenes. The type 2 of poster is more in line with the concept of “ground” in traditional understanding. Figure 1 shows these two different categories.

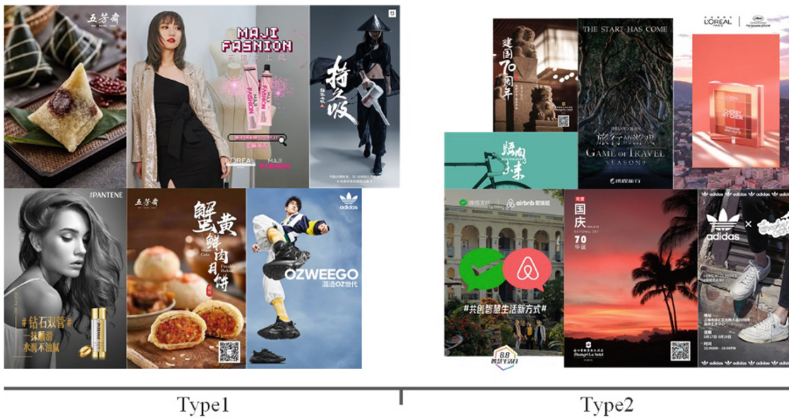


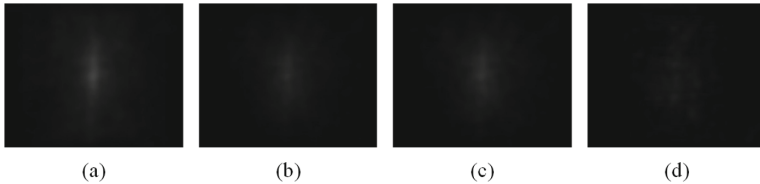
Fig. 1. Two different categories of posters.

### 4.1 The Saliency Map of Background Image

From the “figure-ground” relationship, the saliency on the figure is always stronger than on the ground. In a similar way, in the layout design, the saliency on the text and icons is

always stronger than the on the background image. By calculating the average value of RGB of each pixel of each poster, the average saliency map of the poster shown can be obtained in Fig. 2. Comparing Fig. 2(a)(b), it can be found that the saliency area in (a) is larger and stronger. A vertical white line can be found in (a). This means that the posters have consecutive saliency in the vertical direction. This may be due to the fact that most of the posters in the experiment are in vertical composition, and many layouts have obvious text arrangements in the vertical direction, which draws more attention from the participants. In general, the saliency of the “figure” in a poster is indeed stronger than that of the “ground”. According to the above classification of the poster, it can be seen that the saliency of type 2 posters is much weaker than that of type 1 posters. All saliency maps in Fig. 2 show that when people watching a poster, their attention is always concentrated in the center of the image. There may be two reasons to explain this phenomenon. One is that when designers design posters, they often place important content in the center of the image, whether it is a “figure” or the important content on the background image (for type1). Second is the visual characteristics produced by human experience. Human vision is often more sensitive to the center of the image.

In order to further study how to use the saliency of the image to reasonably cut the background image, the following research will focus on the two points of saliency on the background image. 1. Study the location characteristics of salient areas of the background image. 2. Study the salient area of the background image.



**Fig. 2.** (a) The average saliency map of all complete posters. (b) The average saliency map of all posters without the “figures”. (c) The average saliency map of all posters without the “figures” for type 1. (d) The average saliency map of all posters without the “figures” for type 2.

## 4.2 Position of the Center of the Salient Region

The study uses Tobii Studio analysis software to analyze the eye movement data. The software can automatically form the subject’s interest area according to the subject’s fixation point, and form salient regions of different sizes on the test poster, as shown in Fig. 3(b). In order to study the influence of the “figure” on the saliency map of the background images, the positions and areas of the “figures” in each poster are also drawn in the study. The region of the “figure” manually divided by orange wireframe are shown in Fig. 3(c). The barycenter coordinates of each multi-deformation are calculated here to represent the position of each salient region in further study. There are three kinds of regions should be calculated. As shown in Fig. 3(c), the first kind of region is divided by orange line, which indicates the “figures” on the poster, including text, icons, product pictures, photos, etc. (It is represented by SS in the following table.) The second kind of

the region, divided by the green line, indicates the salient region formed on the “figure” of the poster. (It is represented by S-SS in the following table.) And the third kind of the region, as shown in Fig. 3(c), indicates the salient region formed on the background with orange line. (It is represented by S-BG in the following table.)

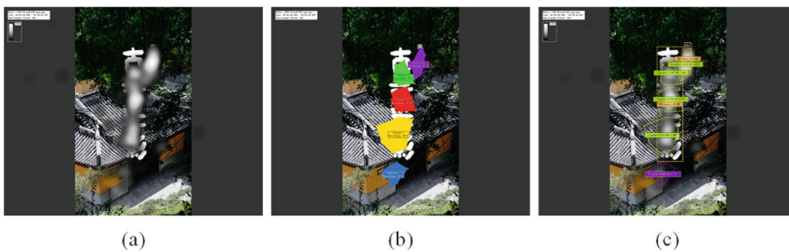
$$C_x = \frac{\sum C_{ix}A_i}{\sum A_i}, C_y = \frac{\sum C_{iy}A_i}{\sum A_i} \tag{1}$$

$$x = \frac{x_1 + x_2 + x_3}{3}, y = \frac{y_1 + y_2 + y_3}{3} \tag{2}$$

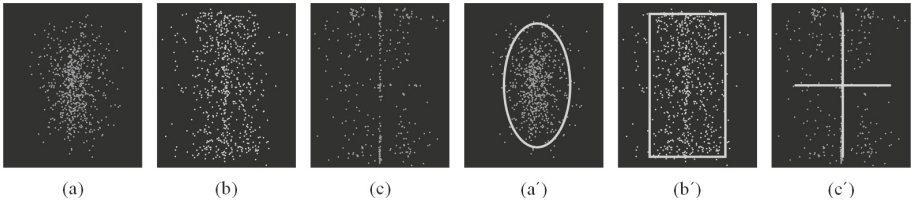
$$A = \frac{1}{2}|x_1y_2 - x_2y_1 + x_2y_3 - x_3y_2 + x_3y_1 - x_1y_3| \tag{3}$$

The formula for calculating the barycentric coordinates of a polygon is shown in (1). The polygon X can be divided into n finite simple figures (the simple figures here use triangles to calculate) X1, X2, . . . Xn. Each X has its barycentric coordinates Cn and its area An. The coordinates of the center of gravity (Cx, Cy) of this polygon can be obtained from these two quantities. If the vertices of each polygon are n, the number of triangles is n - 2. If the three vertices of each triangle are (X1, Y1), (X2, Y2), (X3, Y3), according to the formula (2), the barycentric coordinates (X, Y) of each triangle can be calculated. And the area of each triangle also can get by formula (3). Finally, the barycentric coordinates of each divided region can be obtained.

After calculation, the barycentric coordinates of the three kinds of regions are shown in Fig. 4. The image sizes in Fig. 4 have selected the average size of all posters, which is 860 \* 1010 pixels. It can be seen from Fig. 4(a) that the salient region on the background image is concentrated on the center of the poster and is oval in shape. Figure 4(b) shows the position of barycentric coordinates of the salient region on the “figures”. It is rectangular in shape. Figure 4(c) shows the barycentric coordinates of the “figures”. The position has a clear cross shape, which reflects a certain layout alignment rule. What should be worth noting is that the barycentric coordinates of all the salient regions marked here, is no difference in saliency strength.



**Fig. 3.** (a) Heat map formed based on eye movement data. (b) Salient regions automatically formed by Tobii Studio. (c) Three kinds of regions divided with different color line. (Color figure online)



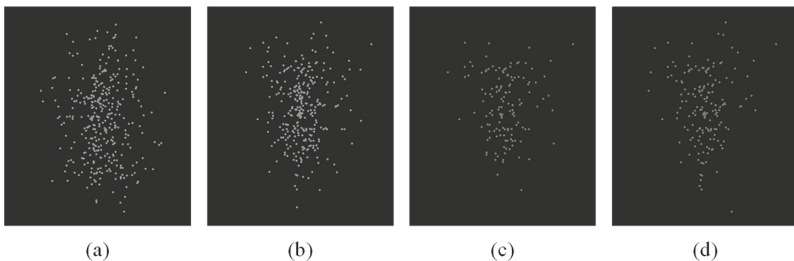
**Fig. 4.** (a) The barycentric coordinates of the salient regions on background image. (b) The barycentric coordinates of the salient regions on “figures”. (c) The barycentric coordinates of the “figures” on the posters.

### 4.3 Position of the Salient Region on the Background Image

The study of the positional characteristics of salient regions on the background image better helps to place the cropping frame in an appropriate position.

#### More Salient Region on the Background Image

In the experiment, the eye tracker records the fixation points of each participant on the poster. The software automatically generates the area of interest, that is, the salient region on the poster, based on all participants’ fixation points. The region being watched by more participants is more salient. In the software, 0–100 is used to indicate the saliency of each region. Compare all the posters to get the absolute saliency values, as shown in Fig. 5(a) (b) (c) show the positions of saliency values  $\leq 50$ ,  $> 50$ ,  $> 90$  respectively. An inverted triangle can be found from the Fig. 5(a) to (c) gradually. (d) shows the location of the most salient region on the background image in each poster, which can better represent the characteristic of the salient position on each poster. And it can be observed that the distribution of the position coordinates is a clear inverted triangle. The position distribution of the salient regions with weaker saliency is usually closer to the center and downward, and the diffusion to the surroundings is more uniform. while the position distribution of the salient regions with stronger saliency is closer to the picture. Comparing the horizontal axis of the picture, it can be found that the position of the region with stronger saliency is more concentrated in the center.



**Fig. 5.** (a) The distribution of the barycentric coordinates of the saliency regions with saliency value  $\leq 50$ . (b) The saliency value  $> 50$ . (c) The saliency value  $> 90$ . (d) The distribution of the barycentric coordinates of the most saliency regions on each poster.

In general, in posters, the more salient areas in the background image are often concentrated in the center of the poster. The closer the position of the saliency position distributes on middle of the horizontal axis, the larger it spread over on the vertical axis. And the most fundamental is that the closer the position of the saliency position to the center of the poster, the higher the probability is. The most salient region location coordinates in each poster will be used to do more analysis.

### **Distribution Characteristics of the Salient Region on the Background**

To study the characteristics of the salient region of the background image in the entire poster layout, four variables can be proposed here for research. 1. The degree of deviation of the most salient region on the background from the center of the poster. 2. The degree of the most salient region on the background from the “figures”. 3. The position of barycentric coordinates on the horizontal axis of the most salient region on the background image. 4. The position of barycentric coordinates on the vertical axis of the most salient region on the background image. After deleting the samples that did not form a saliency region on the background image, there are 165 samples. The analysis shows that these variables all conform to the normal distribution.

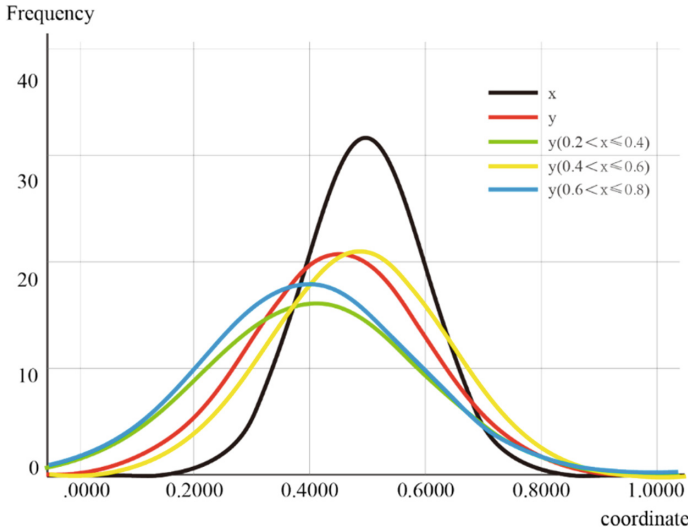
The degree of deviation of the most salient region on the background from the center of the poster calculates the distance from the most salient region to the center, the degree of the most salient region on the background from the “figures” calculates the average distance from the most salient region to the “figures”. Comparing these two variables shows that the latter is much longer than the former.

Although the distance from the salient region on the background image to the center of poster and the “figures” well describes the remote situation of the salient region on the background image, it cannot be further used in actual operations. Therefore, the position of barycentric coordinates on the horizontal axis and the vertical axis of the most salient region on the background image should be analyzed. All x and y coordinate are normalized. According to the above qualitative analysis, it can be found that when the x coordinate is determined in different region segments, the distribution of the y coordinate is different. So here the x-axis is divided into 5 segments, which are 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, 0.8–1.0. The distribution curve of the y-coordinates in the intervals of 0.2–0.4, 0.4–0.6, and 0.6–0.8 is shown in Fig. 6 (x has almost no distribution in the intervals of 0–0.2, 0.8–1.0, so it is not shown in the figure) It can be seen from Fig. 6 that when the x coordinate is determined in a certain interval, the distribution probability of the y coordinate is significantly different. Especially the skewness value, when x is in the interval of 0.2–0.4 and 0.6–0.8, the distribution of y coordinate shifts to the left obviously. The skewness of the x and y coordinates are slightly greater than 0, that is, the entire background salient region has a tendency to shift to the upper left.

### **Typographical Factors that Influence the Location of Salient Region on Background**

In a layout design, there are many factors that affect the location of the salient regions on the background image. The following six factors are listed here:

- The length of the poster.
- The width of the poster.
- The number of “figures” on the poster.



**Fig. 6.** Shows the details of the normal distribution curve of the x-coordinate and y-coordinate.

- The area ratio of the “figures” on the poster.
- The layout category.
- The role of the back-ground images.

According to the role of the background image, the posters can be divided into 2 categories, as shown in Fig. 1. Here raises another classification of the poster according to the typographic category to consider the influencing factors more comprehensively. The layout category has something to do with the text in the poster. It categorizes posters based on the number of texts. As shown in Fig. 7, there is almost no text logos or other elements in the type 1, only one line of text or one text segment in the type 2, and no less than two text segments in the type 3.

After analysis of variance, it is found that different layout categories have no effect on the position of salient region on background. However, different role of different background images in the poster will have different effects on the position of salient region on the background image. According to the one-way analysis of variance, it can be seen that the average distance (211.47) from the “figures” of the type 1 is significantly lower than the average distance (408.47) from the “figures” of the type 2. In addition, under this category, the average distance between the salient region on the background image and the center of the poster in type 1 is lower too. The comparison can be seen in Fig. 8(a).

Thereafter, Pearson correlation analysis was performed on the above six layout factors and the positional variables of the background image salient region. It can be found that there are two factors that have a greater impact on the position of the salient region on background image. One is the “figures” area in the poster. Another is the different role of background image in the poster. The study found that only the area ratio of the “figures” in the posters and the different role of background image effects on all the

positional variables of the background image salient region. In addition, the area ratio of the “figures” has a strong negative correlation with the x-coordinate of salient region on the background. (The correlation coefficient is  $-0.646$ ).

In general, when there is main content in the background image, such as main promotional products, portraits, the salient region on the background image is farther from the center of the poster or the “figures”. Because the “figures” in the poster has strongly saliency. While the content the background image also attracted significant attention of the participants, so they need to stay away from each other and not interfere

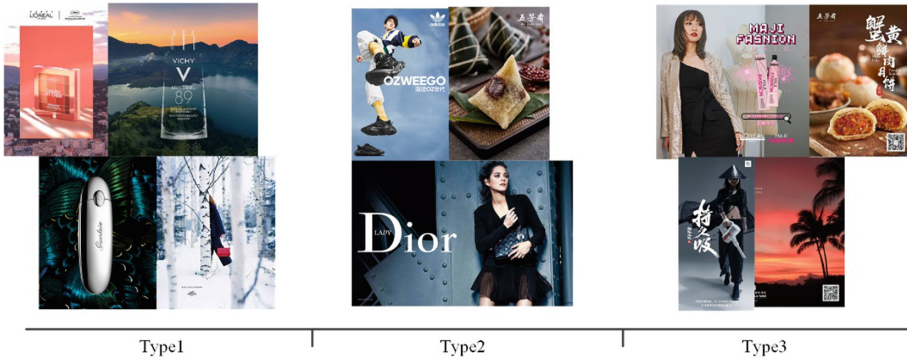


Fig. 7. The posters are divided into three categories based on the number of texts.

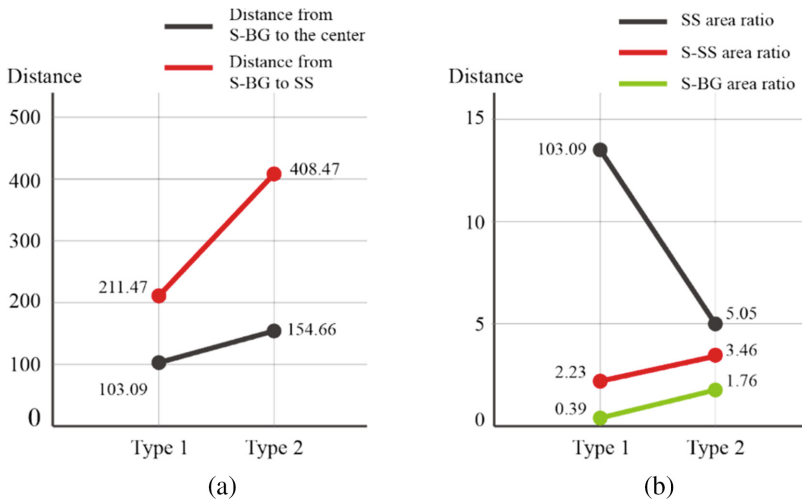


Fig. 8. (a) Comparison of the average distance between the salient region on the background from the center of the poster and the “figures” under different role of the background image. The distance here is relative to the  $860 * 1010$  poster size. (b) Comparison of the area ratio of the “figures” in the poster (%), the area ratio of the salient region on the “figures” (%), and the area ratio of the salient region on the back-ground image (%) under different role of the background image.

with each other. When the background image in the poster mainly serves as a foil, the salient regions on the background image will be closer to the “figure” and the center of the poster. The area of the “figure” in the poster has a strong negative correlation with the location of the salient region on the background image. That is, when the area occupied by the “figures” in the poster is larger, the salient regions on the background is often placed on the left side of the poster.

#### 4.4 Area of the Salient Region on the Background Image

The area characteristics of the salient region on the background image are studied in order to better scale the image. The salient region on the image should be scaled to a reasonable size before cropping.

##### **The Area Characteristics of the Most Salient Region in the Background Image**

Three types of regions have been divided in the study, as shown in Fig. 3(c), which are the region division of the “figure” in the poster, the salient region division on the “figure” in the poster, and the salient region division on the background image in the poster. Tobii Studio analysis software can be used to obtain the area ratio of these regions in the respective poster. It can be found that the area ratio that attracted the attention of the subjects on the “figure” is almost similar to the area ratio of the salient region formed on the background image. It shows that the area formed by the fixation point is almost uniform on the “figure” or the “ground”. If others want to study the relationship between “figure” and “ground” in the poster through eye movement data in an experiment, they can choose to look at the time or the order of the fixation points.

##### **Layout Factors that Affect the Area of the Background Salient Region**

Pearson correlation analysis was performed using the 6 layout factors mentioned above and the area of background salient region. It is found that the area of the “figure” and the number of “figures” in the poster have a strong negative correlation with the area of the salient region on the background image. In a poster, the larger the area of the “figure”, the smaller the area of salient region on the background image in the poster (the correlation coefficient is  $-0.696$ ). And to some extent, the layout category is also classified based on the number of the text in the poster, so the layout category also has a certain negative correlation with the area of the salient region on the background image.

The different role of background images in the poster have different effects on the area of salient region on the background image too. See Fig. 8(b). The average area ratio of the salient region on the background image in type1 (0.64) is significantly lower than the average area ratio in type 2(1.80). The samples in different function of background all show consistency in the salient region on the “figure”.

#### 4.5 Cut the Background Image with the Saliency Characteristics

After research, it is found that in the poster, the position and coordinate distribution of the salient region on the background image has a negative correlation with the area of the “figure” in the poster, and the negative correlation with the x coordinate is stronger. And the area of the salient region on the background image also have a negative correlation

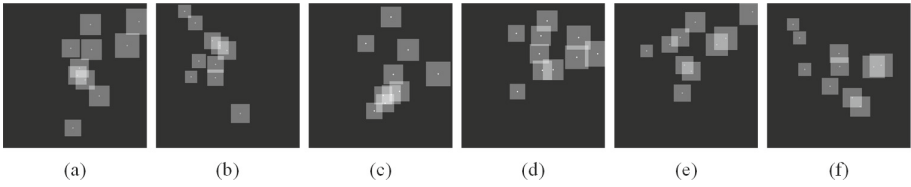


with the area of the “figure” in the poster. In addition, when the background image plays a different role in the poster, it will also have different effects on the position and area of the salient region on the background image. Although many influencing factors have appeared, a simple linear model is not enough to complete the coordinates of the salient region on the background image (the area of the “figure” in the poster can only explain about 23% of the coordinates of the salient region on the background image), its internal relationship still needs further study. Therefore, in this paper, the x-coordinate distribution characteristics of the salient region on the background are used to generate random x-coordinate, and according to the x-coordinates, the y-coordinates that meet the distribution characteristics are randomly generated later. Next, the size of the salient area in the background image is decided and finally uses these saliency characteristics to cut the image.

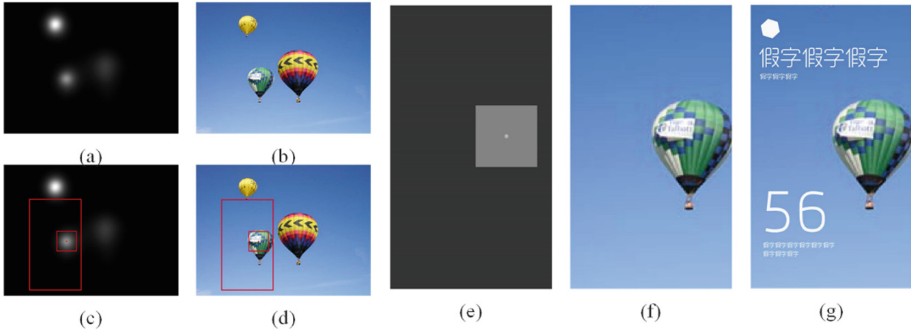
This research uses the Montecarlo algorithm to randomly generate x-coordinates that conform to the distribution rule. This method first needs to generate 2 random numbers and the probability that these 2 random numbers will be selected. If the probability of the first random number is greater than the probability of the second random number, the first random number can be used. The distribution characteristics of the x coordinate are shown in Fig. 6. When the coordinate value of x is determined, the y coordinate is randomly generated according to the distribution characteristics of y corresponding to x in different intervals. Both the x and y coordinates are related to the area of the salient region on the background image. The x and y coordinates are positively correlated with the area of the significant region. It is learned from the correlation that in a poster, from the upper left corner to the lower right corner, the area of the background salient region is gradually increase.

Here, the x-coordinate ratio and y-coordinate ratio are used as independent variables, and the S-BG area is used as the dependent variable for linear regression analysis to obtain a simple linear model. The R<sup>2</sup> value of this model is 0.239. The model formula is, the area ratio of the salient region on the background image =  $0.194 + 3.309 * x \text{ coordinate} - 0.549 * y \text{ coordinate}$  (the x coordinate and y coordinate have been standardized here). And the DW value was near the number 2, which indicated that the model did not have autocorrelation. There is no correlation between them, and the model is better. Figure 9 shows some sets of generated coordinate points and areas on the pictures with the size of 1000 \* 1000 pixels. Ten sets of data were generated in each picture.

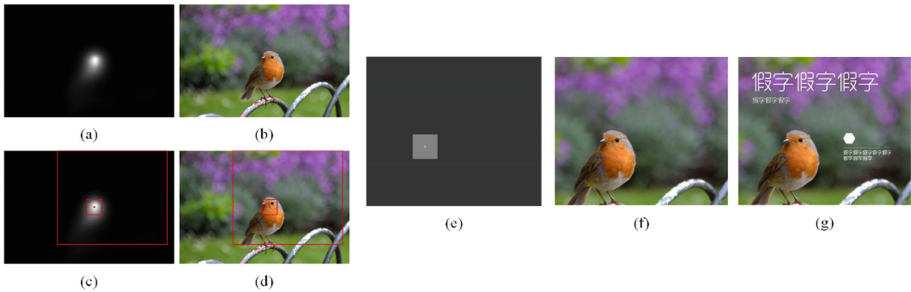
The randomly generated picture can simply become a cutting suggestion map. Based on the salient map of each image, it displays several cropping suggestions for the image, including where the image’s salient position should be and how the image should be scaled and cut. It is worth noting that the generated values are all ratios. In actual operation, the actual poster length and width required can be brought into the actual coordinate values and the area size of the salient region. As shown in Figs. 10(e) and 11(e) shows the effect of the cutting recommendation map under different specific sizes. Figures 10 and 11 completely show the process of cutting an image by using the generated cutting suggestion map, and then making a layout design. The cropped images and their saliency map used here comes from a model to predict human eye gaze points constructed by M Cornia et al. [18].



**Fig. 9.** Shows 6 groups of cutting suggestions randomly generated by the program, each group has 10 randomly generated x and y coordinates and the corresponding area of the salient region on background image.



**Fig. 10.** Use the generated cutting suggestion image to cut the image, and then do layout design. Here is a portrait composition.



**Fig. 11.** Use the generated cutting suggestion image to cut the image, and then do layout design. Here is a horizontal composition.

## 5 Conclusion

By studying the salient characteristic of the background image in the poster, this article aims to confirm that using the saliency of an image can reasonably cut it as a background image for typographic design. In this paper's research on the saliency of poster background images, it is explained that the saliency of background images in poster design has its own characteristics in terms of its location distribution and area of salient region. As a background image in a poster, the location distribution of its salient area

will show an inverted triangle, and the probability density at the center of the poster is higher. A simple linear model is also established in this article. The area of the salient region is roughly calculated from the known coordinates of the salient region. Analyzing the correlation, we know that the area of the salient region is gradually increase from the upper left corner to the lower right corner of the poster. Using the analyzed saliency rules in the background image, an image can be cut to make the cropped picture more consistent with the image as a background image in the layout design.

In the future, intelligent design will better serve the society, and it will also become a powerful assistant for designers. These studies will provide the basis for future automatic layout design. When the machine uses an image as the background, cutting the image will be an important step. Cutting an image that is more suitable for typesetting has also become a first step for the research. Although many previous studies have used the saliency to cut the images, the focus of these research is how to quickly find accurate regions of significant objects and the accuracy of border areas to ensure that important areas are retained. The use after cutting is not taken into account. The research in this article focuses on the saliency of special images as background images. By studying the saliency characteristics of this kind of image, they are used to cut out images suitable for a certain type of use. This provides the basis for the general direction of future intelligent typesetting.

## 6 Discussion and Future Work

This article mainly studies the characteristics of saliency of the background image, however one thing to note is that most of the posters in this study are vertical. This may have a bearing on the outcome. The characteristics in the horizontal composition posters may not be able to summarize well.

Many factors that affect the saliency of the background image have been proposed in this paper. The role of the background image and the area of the “figure” in the poster are the more influential factors in the existing research. The relationship between these influencing factors and saliency has not been further studied in this article. Some simple linear models may not explain these relationships. Among them, there may be more complicated non-linear relationships that are worthy of further in-depth study.

## References

1. Borji, A., Cheng, M.-M., Jiang, H., et al.: Salient object detection: a benchmark. *IEEE Trans. Image Process.* **24**, 5706–5722 (2015). <https://doi.org/10.1109/TIP.2015.2487833>
2. Zhao, R., Ouyang, W., Li, H., et al.: Saliency detection by multi-context deep learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1265–1274 (2015). <https://doi.org/10.1109/cvpr.2015.7298731>
3. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1254–1259 (1998). <https://doi.org/10.1109/34.730558>
4. Liu, T., Yuan, Z., Sun, J., et al.: Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 353–367 (2010). <https://doi.org/10.1109/CVPR.2007.383047>

5. Achanta, R., Hemami, S., Estrada, F., et al.: Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597–1604. IEEE (2009). <https://doi.org/10.1109/cvpr.2009.5206596>
6. Borji, A.: What is a salient object? A dataset and a baseline model for salient object detection. *IEEE Trans. Image Process.* **24**, 742–756 (2014). <https://doi.org/10.1109/TIP.2014.2383320>
7. Bylinskii, Z., Kim, N.W., O'Donovan, P., et al.: Learning visual importance for graphic designs and data visualizations. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, pp. 57–69 (2017). <https://doi.org/10.1145/3126594.3126653>
8. Jahanian, A., Liu, J., Tretter, D.R., et al.: Automatic design of magazine covers (2012). <https://doi.org/10.1117/12.914596>
9. Jahanian, A., Vishwanathan, S.V.N., Allebach, J.P.: Learning visual balance from large-scale datasets of aesthetically highly rated images. In: Proceedings of SPIE - The International Society for Optical Engineering, vol. 9394 (2015). <https://doi.org/10.1117/12.2084548>
10. Chen, L.-Q., Xie, X., Fan, X., et al.: A visual attention model for adapting images on small displays. *Multimed. Syst.* **9**, 353–364 (2003). <https://doi.org/10.1007/s00530-003-0105-4>
11. Suh, B., Ling, H., Bederson, B.B., et al.: Automatic thumbnail cropping and its effectiveness. In: Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology, pp. 95–104 (2003). <https://doi.org/10.1145/964696.964707>
12. Santella, A., Agrawala, M., DeCarlo, D., et al.: Gaze-based interaction for semi-automatic photo cropping. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 771–780 (2006). <https://doi.org/10.1145/1124772.1124886>
13. Marchesotti, L., Cifarelli, C., Csurka, G.: A framework for visual saliency detection with applications to image thumbnailing. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2232–2239. IEEE (2009). <https://doi.org/10.1109/icc.2009.5459467>
14. Chen, J., Bai, G., Liang, S., et al.: Automatic image cropping: a computational complexity study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 507–515 (2016). <https://doi.org/10.1109/cvpr.2016.61>
15. Borji, A., Sihite, D.N., Itti, L.: What stands out in a scene? A study of human explicit saliency judgment. *Vis. Res.* **91**, 62–77 (2013). <https://doi.org/10.1016/j.visres.2013.07.016>
16. Xu, J., Jiang, M., Wang, S., et al.: Predicting human gaze beyond pixels. *J. Vis.* **14**, 28 (2014). <https://doi.org/10.1167/14.1.28>
17. Koehler, K., Guo, F., Zhang, S., et al.: What do saliency models predict? *J. Vis.* **14**, 14 (2014). <https://doi.org/10.1167/14.3.14>
18. Cornia, M., Baraldi, L., Serra, G., et al.: Predicting human eye fixations via an LSTM-based saliency attentive model. *IEEE Trans. Image Process.* **27**, 5142–5154 (2018). <https://doi.org/10.1109/TIP.2018.2851672>



# Gamification Elements on Social Live Streaming Service Mobile Applications

Franziska Zimmer<sup>1</sup>(✉), Katrin Scheibe<sup>1</sup>, and Hantian Zhang<sup>2</sup>

- <sup>1</sup> Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany  
{franziska.zimmer, katrin.scheibe}@hhu.de
- <sup>2</sup> Sheffield Hallam University, Howard Street, Sheffield S1 1WB, UK  
hantian.zhang@shu.ac.uk

**Abstract.** Social live streaming services (SLSSs), a kind of synchronous social networking service, are slowly but surely becoming a part of people’s daily lives. To keep users interested, a wide range of gamification elements are implemented on these services, increasing the user engagement and changing their behavior. This study examined 20 different SLSS mobile applications and the applied gamification elements. A literature review as well as a content analysis were used to find appropriate SLSS apps and game elements. What kind of mechanics can be found on SLSS mobile apps and how many are implemented on each system? On three of the observed apps we could identify all game elements. Chinese SLSS apps are the most gamified ones. On Ustream, no game element is implemented. The game mechanics *following others* as well as *customization* are the most often applied; *capturing a moment* of a stream is the least often implemented.

**Keywords:** Social live streaming service · Gamification · Content analysis · Mobile application

## 1 Introduction

Gamification is a promising and tactical strategy often used in education or online applications [5]. One definition of gamification was coined by Deterding, Nacke and Dixon: “gamification [is] the use of game design elements in non-game contexts” [6, p. 1]. Generally speaking, gamification can be seen to be comprised of three main elements as proposed by Hamari, Koivisto, and Sarsa [14]: the elements or mechanics that are used within a system are aimed at making the user undergo a gameful experience. This in turn leads to psychological outcomes, which can be, for example, encountering a feeling of competence when solving a task or quest, but also enjoyment which is one of the main ambitions of gamification. Based on these positive experiences, it is estimated that the user will change their behavior. These behavioral outcomes can be seen in better results if learning and language apps are concerned, or health applications and physical fitness. Gamification is estimated to be implemented by many companies in the near future, proving it to be an important mechanic in business contexts [19].

The potentials of gamification are also seen for other areas, for example government services and public engagement, crowdsourcing, commerce, exercise [20], marketing and advertising, environmental behavior, and information systems [24].

Social networking services (SNSs), a special type of information system, are rather recommended by users if they are gamified, also, the intention to use the service increases [13]. Recently, SNSs such as Facebook and the video sharing platform YouTube implemented the live streaming feature, making them an embedded social live streaming service (SLSS), e.g. YouTube Live. There are two other types of SLSSs - general live streaming services where no specific focus or subject is prevalent (e.g. Periscope), and topic-specific live streaming services attracting one special interest group concerned with a certain kind of content, e.g. Twitch for eSports. This kind of SNS seems to be especially attractive in China, as there are already over 200 different SLSSs [27]. The implementation of gamification elements makes users of SLSSs feel rewarded and motivated through the interaction with the game mechanics [35].

SLSSs are mainly used out of boredom, for socializing, communication, and entertainment [3, 9, 10, 16]. In this context, the Uses and Gratifications Theory (UG&T) by Blumler and Katz [1] needs to be mentioned. If one applies media, it is usually goal-oriented and underlines a kind of expectation [22]. McQuail [30] states four main goals or motivations to use media: entertainment, information, personal identity and social interaction. In the context of SLSSs, the aspect of personal identity shall be redefined as self-presentation [45].

The Self-Determination Theory (SDT) proposed by Ryan and Deci [32] also concerns human needs and user motivation. They describe motivation as an action that drives people which is influenced by external and internal factors. Intrinsic motivation makes people engage in activities they find interesting but also challenging, making it an internal driving factor. Extrinsic motivation is an influence from outside, for example monetary rewards or fame. Hamari et al. [14] state that users of an information system are intrinsically motivated by game design elements.

All in all, gamification is applied to motivate the user and for repetitive information system usage [4].

## 2 Related Work

There are already studies on the impact of gamification in context with live streaming behavior. First of all, the game mechanics are seen as a motivating factor, making users want to keep using a service [12, 15, 23, 39].

A live streaming application was developed in three different versions to test the effect of gamification elements on the broadcasting behavior of SLSSs users [42]. The first version did not contain any game mechanics, the second version contained levels, and the third version added challenges and badges. The results indicate that the more gamification elements are implemented and used, the longer the streaming time.

The impact of gamification was investigated using YouNow as a case study [35]. Three different user groups (producers of streams or streamers, participants, and consumers) were analyzed, all seem to feel rewarded through different gamification elements. Most motivated by gamification mechanics are the producers of streams. Overall,

all elements are at least perceived as being neutral but most often as highly rewarding and motivating.

YouNow was also the focus of a study showcasing if a difference between giving and receiving gratifications in a gamified SLSS could be observed. All game mechanics are perceived as being fun, useful, rewarding, and motivating. In general, the users seem to rather want to receive gratifications from others than taking the action to give gratifications to the streamer [36].

Another service, Twitch, is an ideal platform to investigate as the activities of streaming and watching streams are highly gamified. For example, streams can be individualized through customization to keep the viewers entertained [38]. But, the study also points out that not all features will suit all streamers and their streams. A streamer who garners a lot of viewers is not able to read all commands and chat comments, therefore some actions are getting lost in the chat history.

Twitch was the case study of another study concerned with a web-based leaderboard tool developed to amplify the gamification effect of word-of-mouth referrals which is intended to help the streamer grow his audience [2]. Since word-of-mouth programs, which give the customers incentives to share with their friends and families, for example a referral code, are successfully employed by many companies on social media, the impact could likely be as effective for streamers. As the study points out, the tool increased the number of new viewers and is also appreciated by the Twitch community.

Lu, Xia, Heo, and Wigdor [27] mention the engaging role of the gifting function and fan groups in Chinese SLSSs. As gift-sending viewers are sometimes treated more special by the streamer, gifting seems to be a popular option in streamer and viewer interaction. Gifting serves as a more meaningful and expressive way of communication than texts. China is the country with the most SLSSs as of now, applying various game mechanics and elements to keep the streamers and viewers engaged [37]. The study also investigated the amount of game mechanics found on the most popular SLSSs websites in the world. The features that were implemented predominantly are following others, leaderboards, and, ranking third together with currency, badges, and gifts.

In summary, gamification is motivating for streamers and viewers of SLSSs, keeping them engaged and wanting them to keep on using a service. But, to our knowledge, a study on SLSS apps (Fig. 1) and the most applied game design elements on them has not been conducted yet. Therefore, we are going to close this research gap. Based on these aspects, we arrive at the following research question:

**RQ1.** Which gamification elements are implemented on social live streaming service mobile apps?

### 3 Method

For this study, the aim is to get an overview of the implemented game mechanics and game elements on different SLSS mobile applications. It is possible to add gamification elements to the layout of the stream via bots (e.g. a ranking that lists top gifting viewers). This kind of game mechanic was not considered in this study. The focus lies on the game mechanics prepared and applied by the system and the apps itself.

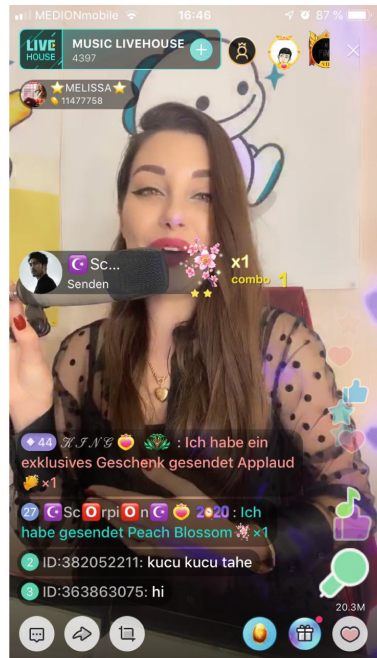
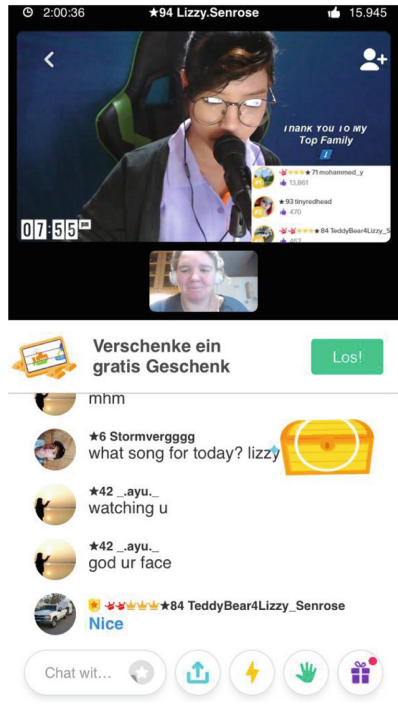
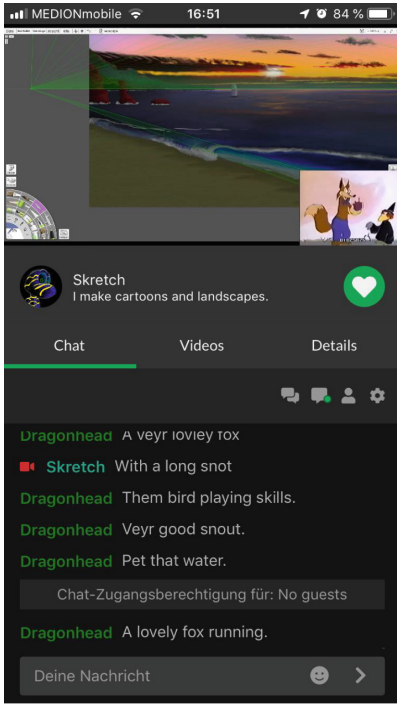


Fig. 1. Screenshots of (left to right) Picarto, YouNow, niconico, Bigo Live



**Table 1.** SLSS apps and their global and country-specific ranking

| SLSS                    | Global rank | Rank in country |
|-------------------------|-------------|-----------------|
| YouTube.com             | 2           | USA: 2          |
| Facebook.com            | 5           | USA: 4          |
| live.qq.com             | 6           | China: 3        |
| Twitch.tv               | 40          | USA: 18         |
| yy.com                  | 96          | China: 29       |
| Nicovideo.jp (niconico) | 222         | Japan: 19       |
| Mixer.com               | 1,370       | USA: 639        |
| Huya.com                | 1,750       | China: 244      |
| Pscp.tv                 | 10,710      | USA: 6,933      |
| Kuaishou.com            | 11,518      | China: 1,116    |
| Bigo.tv                 | 12,610      | China: 6,777    |
| Younow.com              | 14,061      | USA: 18,991     |
| Longzhu.com             | 17,145      | China: 1,359    |
| Chushou.tv              | 21,036      | China: 1,365    |
| Ustream.tv              | 23,025      | USA: 34,269     |
| Picarto.tv              | 29,482      | USA: 11,095     |
| Huajiao.com             | 32,190      | China: 2,119    |
| Lai Feng.com            | 37,746      | China: 4,669    |
| Qiuxiu (x.pps.tv)       | 59,691      | China: 7,013    |
| Yizhibo.com             | 88,764      | China: 6,523    |

Data source: Alexa (as of December 29th, 2019)

Furthermore, SLSS websites of the applications were not included in this study, as research on the same subject was already conducted [37]. Also, not every implemented game mechanic of a system may be used by each user group (producer, participant, or consumer). The systems were examined from each user group's perspective, but because of only a few differences we showed no differentiation in the results section. As our investigative method, a total of 20 SLSS mobile applications were examined and evaluated for a defined set of gamification elements. For this study, a content analysis was conducted with the conventional and deductive approach applying literature review [8, 18]. Via the directed approach [8], we examined SLSSs for different game mechanics and categorized them.

### 3.1 Appropriate SLSSs

A first analysis of SLSS websites and the applied game elements was conducted [37]. Based on this, as a comparative measure, our study focuses on the game mechanics of

the corresponding apps. Primarily, the SLSSs websites were selected through literature research [e.g. 21, 27, 31, 46] as well as online research. We consulted the homepage of the Nanjing Marketing Group, a website specialized on Chinese markets, as China has a big user group of SLSS websites [28]. The 11 highest ranking websites in China were selected from Alexa.com and the corresponding apps investigated. Furthermore, the phrase “live stream” or “#livestream” were searched for on various social media sites (e.g. Instagram, YouTube, Facebook, and Twitter) to gather the Western SLSS websites. After gathering the SLSSs, we checked their Alexa rankings compared to other websites of the world as well as their positions in the country with the most users. From this, the websites were chosen. The Table 1 displays all relevant SLSSs which were examined for the implemented game mechanics.

**Table 2.** Game mechanics found on SLSSs (modified from [37])

| Game mechanics         | Description   | Literature     |
|------------------------|---|----------------|
| Badges                 | Visual elements that are awarded for fulfilling tasks   | e.g., [11]     |
| Capturing moments      | Recording a short clip of a live stream   | e.g., [35]     |
| Collaboration and team | Broadcasting; via split screen of two or more users   | e.g., [35]     |
| Collecting             | Collection of different things, e.g. awards or gifts  | e.g., [26]     |
| Currency               | Bought with real money or earned through tasks to buy gifts                                   | e.g., [7]      |
| Points                 | Earned through different tasks or site activities   | e.g., [33]     |
| Customization          | Changing features of the channel, profile website, or chat                                    | e.g., [40]     |
| Following others       | Users stay up to date through following, becoming a fan, subscribing, or befriending function | e.g., [36]     |
| Gifts                  | Viewers can show their appreciation with gifts  | e.g., [27]     |
| Challenges and goals   | Users can achieve goals and solve tasks that are predefined by each platform                  | e.g., [41]     |
| Leaderboards           | Statistics of the best streamers or viewers according to different criteria                   | e.g., [33]     |
| Progress bar           | Overview of current status until reaching a next step (e.g. level)                            | e.g., [34]     |
| Likes                  | A kind of social feedback from viewers towards streamers                                      | e.g., [29]     |
| Levels                 | Displaying the users' experience in a system  | e.g., [44, 46] |

### 3.2 Game Mechanics

The game mechanics were selected by applying different methods. Literature reviews on gamification (e.g. [14]) and research on different game mechanics (e.g. [29, 41, 43])

were especially considered. All in all, a list consisting of over 20 gamification elements was created. Following, the conventional approach via observing SLSS websites was applied to get an impression on what game mechanics are implemented on SLSSs. The game elements we could not identify on SLSSs websites were not included in the list. The 14 game design elements and a short definition of each one are listed in Table 2.

### 3.3 The Examination

Each SLSS mobile app was examined by a pair of two researchers [25]. Each game mechanic presented in the app was discussed. The coders always arrived at the same conclusion on which category was appropriate for the game mechanic that was observed. For example, if some form of money exchange could be recognized on the SLSS, it was classified as the ‘currency’ category. Since the three researchers did not have the appropriate language skills for Japanese, a fluent speaker was present for the investigation of the Japanese app. All in all, we identified fourteen different game mechanics that are applied by different SLSSs (Table 2).

## 4 Results

Taking a look at Table 3, the number of game mechanics identified on each service are listed. Most Chinese live streaming applications have nearly all game mechanics implemented. At least eight or more gamification mechanics were observed on each Chinese SLSS app. Two of the services from China, Huajiao and Yizhibo, have all 14 game elements, four services have 13 game elements, one service has 12, and three services have 11 game elements. From China, the least elements (eight) were found on Kuaishou.

For the US-American services, only YouNow’s app has all game mechanics implemented. Facebook Live has a number of six, YouTube Live five and one service (Ustream) even has no game elements. For YouTube and Facebook, we have to take into consideration that the services are already established and important SNSs which embedded the function of streaming live broadcasts. The Japanese SLSS mobile application niconico has 11 and the German service Picarto has three game mechanics.

Following, some examples for the detected gamification elements will be mentioned, as the observation table only displays if a certain game mechanic was implemented or not and no details are included (Table 4). Every checked platform, except for Ustream, has the function *following others* and *customization*, as Ustream has no game element.

On Bigo Live, streamers can add stickers to their stream as a kind of *customization*. *Collecting* is implemented on, for instance, Quixiu. Viewers can open chests when watching the stream to earn random awards. Streamers collect gifts and exchange them for income.

With *leaderboards* users are able to compare their performance and accomplishment with other users, YouNow has leaderboards for top broadcasters, top fans, and top moment makers. On the SLSS mobile application of YY, Y coins and red diamonds are implemented as *currency* to be able to buy gifts. *Gifts* on SLSSs serve as a reward for

**Table 3.** Number of game mechanics per SLSS mobile application (N = 14).

| SLSS mobile application ( <i>country of origin</i> ) | Number of game mechanics |
|--|--------------------------|
| <i>China</i>   |                          |
| Huajiao  | 14                       |
| Yizhibo  | 14                       |
| Bigo Live  | 13                       |
| Laifeng  | 13                       |
| Long Zhu   | 13                       |
| Qiuxiu   | 13                       |
| YY   | 12                       |
| Chushou  | 11                       |
| Huya   | 11                       |
| QQ Live  | 11                       |
| Kuaishou   | 8                        |
| <i>USA</i>   |                          |
| YouNow   | 14                       |
| Mixer  | 11                       |
| Twitch   | 9                        |
| Periscope  | 7                        |
| Facebook Live  | 6                        |
| YouTube Live   | 5                        |
| Ustream  | 0                        |
| <i>Japan</i>   |                          |
| niconico   | 11                       |
| <i>Germany</i>                                       |                          |
| Picarto  | 3                        |

the streamer. They are implemented on 18 of the 20 observed services, whereby they are built-in on Picarto and Ustream.

Chushou has *challenges and goals* that are called “missions,” thereby users are able to get badges and to earn “active coins.” The Facebook Live function in the mobile application offers the opportunity to invite friends to the stream. If the friend accepts the invitation, they are able to *collaborate* and stream via split screen.

On Huajiao and Huya experience *points* can be earned by e.g. sending gifts or watching streams. Experience points are for leveling up. On most platforms, a *progress bar* displays the progress to a next level. Whereby Mixer has a progress bar for streamer progression which tracks a streamers growth. Levels display the experience of users, we could identify them on 12 apps, e.g. QQ Live, NicoVideo, Laifeng, and Long Zhu.



*Badges* on Twitch will be earned when fulfilling specific things, such as purchasing bits or giving gifts. Also, subscribers can get a so called “Subscriber Badge.” Via *likes* users can show that they like a streamers live show. Periscope provides the opportunity to send likes with colorful hearts which are shown in the live stream. The least often applied function was *capturing a moment*, we could identify it on Yizhibo, Huajiao, YouNow, and Twitch. Thereby, a user is able to record a certain period of a stream.

**Table 5.** Number of SLSS mobile application having game mechanics (N = 20).

| Game mechanic        | Number of SLSS mobile apps |
|----------------------|----------------------------|
| Customization        | 19                         |
| Following others     | 19                         |
| Gifts                | 18                         |
| Collecting           | 18                         |
| Currency             | 16                         |
| Points               | 16                         |
| Leaderboards         | 13                         |
| Badges               | 13                         |
| Challenges & goals   | 13                         |
| Levels               | 13                         |
| Progress bar         | 13                         |
| Likes                | 12                         |
| Collaboration & team | 11                         |
| Capturing a moment   | 4                          |

The most often detected game mechanics on the observed SLSS mobile applications were customization and *following others*. 19 of the 20 platforms had these functions implemented (Table 5). As customization allows users to be individual, it is a very popular function. With following others, users stay up-to-date about the users’ activities. The functions *gifts* and *collecting*, with 18 each, as well as *currency and points*, with 16 each, were also found often on SLSS mobile apps.

A number of 13 of the 20 observed apps had *leaderboards, badges, challenges and goals, levels, and a progress bar*. *Likes* were observed on 12 SLSS mobile applications and *collaboration and team* on eleven apps. The least often implemented game mechanic is *capturing a moment*. It was only found on four observed systems. As some services (e.g. YouTube Live) offer the opportunity to watch the completely recorded former live streams again, it is not necessarily used on each system.

## 5 Discussion

In this study, a content analysis on 20 different SLSS mobile applications was conducted to discover which game mechanics are applied on each service. Thereby, the applications were checked for a number of 14 game elements. Eleven mobile apps were from

China, six from the United States of America, one from Japan, and one from Germany. The results show that Chinese SLSSs mobile applications apply on average the most game elements. Two of the Chinese services have all 14 and four services have 13 game elements implemented. From the United States of America, only one service (YouNow) has all game mechanics implemented and the following one with the second most has 11 game mechanics. The Japanese one has 11 as well and the German has 3 gamification elements. The tendency shows that Chinese SLSSs have more game mechanics implemented than the ones from the USA. A comparison with German or Japanese services is not applicable as only one service of each country was considered.

Why is there such a tendency for gamification on Chinese SLSS? An explanation could be that there is a more intense competition between the various live streaming services (a number of 200) in China. Following Hamari and Koivisto [13], the intention to use a service increases, if gamification is applied. Furthermore, Scheibe and Zimmer [37] explored similar results for SLSSs websites. Here, the authors consulted findings from Hofstede's country comparison [17] where China is presented as a pragmatic culture. The explanation about China's society could be that they are "driven by competition, achievement and success" [17] which are attributes of gamification, providing a possible explanation for the gamification phenomenon in Chinese SLSSs.

In contrast to other social networking services, SLSSs offer a great variety of gamification elements to their users. Since the primary interaction among users on SLSSs follows the one-to-many communication model during a live stream, gamification elements offer an additional way of interaction on SLSSs.

This study shows that there is a great variety of game mechanics which can be used in many different ways (e.g. different kinds of currencies). The most often implemented game mechanics on SLSSs mobile apps are *customization* and *following others* followed by *gifts* and *collecting* as well as *currency* and *points*. *Capturing a moment* of a stream was implemented the least often. On Ustream, no gamification elements are implemented, and we hypothesize that there are no gamification elements needed, as it is provided for the professional and public live streaming.

Comparing our results to the study of Scheibe [35], she found out that streamers are feeling very rewarded and motivated when getting fans or subscribers. Also, gifts have a strong positive effect on the streamer's motivation, but also on the viewer's motivation when giving gifts to streamers. Earning coins is seen as a reward by all users.

When taking a look at the results from Scheibe and Zimmer [37] about the applied gamification elements on SLSSs websites, a few differences can be observed while keeping in mind that one year has passed since the mentioned study was conducted. Additional gamification elements could have been added to the SLSS websites as well. For example, customization is more often implemented on SLSS mobile apps than on the SLSS websites. Further research should concentrate on the differentiation of applied game mechanics to a website and to a mobile app, as the user experience and behavioral effect may vary depending on the distinct interface structures and layouts.

The limitations of this study should be mentioned. First, we have only observed a small amount of SLSS mobile applications. There is an undefined number of services which remain undiscovered, since in China there are over 200 individual live streaming

systems [27]. The number of services that were checked in this study is 20, whereof only eleven are Chinese.

Furthermore, live streaming platforms from other countries were not examined and should be considered in further studies, e.g. African, South American, and other Asian countries. Although our approach followed the four eyes principle, there might be a bias by missing gamification elements while checking the services.

As an outlook, other types of social networking services should be checked and compared to SLSSs. As research points to the implication that SLSSs are mostly applied by generation Z, this could be an important aspect on why gamification works for live streaming; is generation Z more prone to apply gamification elements than for example the baby boomers or generation Y? Also, comparing the acceptance of SLSSs without gamification elements and with gamification elements, like e.g. Wilk, Wulffert, and Effelsberg [42] did, but with public SLSSs would be an interesting investigation.

## References

1. Blumler, J.G., Katz, E.: *The Uses of Mass Communications: Current Perspectives on Gratifications Research*. Sage, Newbury Park (1973)
2. Browne, J.T., Batra, B.: *Twickle: growing Twitch streamer's communities through gamification of word-of-mouth referrals*. In: *Proceedings of TVX 2018*, pp. 149–154. ACM, New York (2018)
3. Chen, C., Lin, Y.: *What drives live-stream usage intention? The perspective of flow, entertainment, social interaction and endorsement*. *Telematics Inform.* **35**, 293–303 (2018)
4. Deterding, S.: *Gamification: designing for motivation*. *Interactions* **19**(4), 14–17 (2012)
5. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: *From game design elements to game-fulness: defining "gamification"*. In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pp. 9–15. ACM, New York (2011)
6. Deterding, S., Nacke, L.E., Dixon, D.: *Gamification: toward a definition*. In: *Proceedings of the 29th CHI Conference on Human Factors in Computing Systems*, pp. 1–4. ACM, New York (2011)
7. Dicheva, D., Dichev, C., Agrev, G., Angelova, G.: *Gamification in education: a systematic mapping study*. *Educ. Technol. Soc.* **18**(3), 74–88 (2015)
8. Elo, S., Kyngäs, H.: *The qualitative content analysis*. *J. Adv. Nurs.* **62**(1), 107–115 (2008)
9. Friedländer, M.B.: *Streamer motives and user-generated content on social live-streaming services*. *J. Inf. Theory Pract.* **5**(1), 65–84 (2017)
10. Gros, D., Wanner, B., Hackenholt, A., Zawadzki, P., Knautz, K.: *World of streaming. Motivation and gratification on Twitch*. In: Meiselwitz, G. (ed.) *SCSM 2017*. LNCS, vol. 10282, pp. 44–57. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-58559-8\\_5](https://doi.org/10.1007/978-3-319-58559-8_5)
11. Hamari, J.: *Do badges increase user activity? A field experiment on effects of gamification*. *Comput. Hum. Behav.* **71**, 469–478 (2017)
12. Hamari, J., Hassan, L., Dias, A.: *Gamification, quantified-self or social networking? Matching users' goals with motivational technology*. *User Model. User-Adap. Inter.* **28**(1), 35–74 (2018). <https://doi.org/10.1007/s11257-018-9200-2>
13. Hamari, J., Koivisto, J.: *Social motivations to use gamification: an empirical study of gamifying exercise*. In: *Proceedings of the 21st European Conference on Information Systems*, pp. 1–13. Association for Information Systems, Atlanta (2013)
14. Hamari, J., Koivisto, J., Sarsa, S.: *Does gamification work? – a literature review of empirical studies on gamification*. In: *Proceedings of the 47th Hawaii International Conference on System Sciences, Big Island*, pp. 3025–3034. IEEE (2014)



15. Hamari, J., Sjöblom, M.: What is eSports and why do people watch it? *Internet Res.* **27**(2), 211–232 (2017)
16. Hilvert-Bruce, Z., Neill, J.T., Sjöblom, M., Hamari, J.: Social motivations of live-streaming viewer engagement on Twitch. *Comput. Hum. Behav.* **84**, 58–67 (2018)
17. Hofstede Insights: Country Comparison of China and USA. <https://www.hofstede-insights.com/country-comparison/china,the-usa/>. Accessed 27 Jan 2020
18. Hsieh, H.-F., Shannon, S.E.: Three approaches to qualitative content analysis. *Qual. Health Res.* **15**(9), 1277–1288 (2005)
19. Corcione, A., Tardo, F.: IEEE via PR Newswire: Everyone’s a gamer – IEEE experts predict gaming will be integrated into more than 85 percent of daily tasks by 2020 (2014). <https://www.prnewswire.com/news-releases/everyones-a-gamer—ieee-experts-predict-gaming-will-be-integrated-into-more-than-85-percent-of-daily-tasks-by-2020-247100431.html>
20. Ilhan, A., Fietkiewicz, K.J.: Learning for a healthier lifestyle through gamification: a case study of fitness tracker applications. In: Buchem, I., Klamma, R., Wild, F. (eds.) *Perspectives on Wearable Enhanced Learning (WELL)*, pp. 333–364. Springer, Cham (2019). [https://doi.org/10.1007/978-3-319-64301-4\\_16](https://doi.org/10.1007/978-3-319-64301-4_16)
21. Izumi, T., Tarumi, H., Kagawa, E., Yaegashi, R.: An experimental live streaming of an ice hockey game with enhancement of mutual awareness. In: *Proceedings of the 6th International Conference on Collaboration Technologies*, pp. 22–25. Information Processing Society, Hokkaido (2012)
22. Katz, E., Blumler, J.G., Gurevitch, M.: Utilization of mass communication by the individual. In: Blumler, J.G., Katz, E. (eds.) *The Uses of Mass Communications: Current Perspectives on Gratification Research*, pp. 19–31. Sage, Thousand Oaks (1974)
23. Koivisto, J., Hamari, J.: Demographic differences in perceived benefits from gamification. *Comput. Hum. Behav.* **35**, 179–188 (2014)
24. Koivisto, J., Hamari, J.: The rise of motivational information systems: a review of gamification literature. *Int. J. Inf. Manage.* **45**, 191–210 (2019)
25. Krippendorff, K.: *Content Analysis: An Introduction to Its Methodology*, 3rd edn. Sage, New York (2012)
26. Kumar, J.: Gamification at work: designing engaging business software. In: Marcus, A. (ed.) *DUXU 2013. LNCS*, vol. 8013, pp. 528–537. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-39241-2\\_58](https://doi.org/10.1007/978-3-642-39241-2_58)
27. Lu, Z., Xia, H., Heo, S., Wigor, D.: You watch, you give, and you engage: a study of live streaming practices in China. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. ACM, New York (2018)
28. Marszalek, W.: Introduction to live-streaming in China (2018). <https://www.nanjingmarketinggroup.com/blog/live-streaming-china>
29. Matallaoui, A., Koivisto, J., Hamari, J., Zarnekow, R.: How effective is ‘exergamification’? A systematic review on the effectiveness of gamification features in exergames. In: *Proceedings of the 50th Hawaii International Conference on System Sciences*, pp. 3316–3325. IEEE Computer Society, Washington (2017)
30. McQuail, D.: *Mass Communication Theory*. Sage, London (1983)
31. Pires, K., Simon, G.: YouTube live and Twitch: a tour of user-generated live streaming systems. In: *Proceedings of the 6th ACM Multimedia Systems Conference*, pp. 225–230. ACM, New York (2015)
32. Ryan, R.M., Deci, E.L.: Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* **55**(1), 68–78 (2000)
33. Sailer, M., Hense, J.U., Mandl, H., Klevers, M.: Psychological perspectives on motivation through gamification. *Interact. Des. Architecture(s) J.* **19**, 28–37 (2013)

34. Sailer, M., Hense, J.U., Mayr, S.K., Mandl, H.: How gamification motivates: an experimental study of the effects of specific game design elements on psychological need satisfaction. *Comput. Hum. Behav.* **69**, 371–380 (2017)
35. Scheibe, K.: The impact of gamification in social live streaming services. In: Meiselwitz, G. (ed.) SCSM 2018. LNCS, vol. 10914, pp. 99–113. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91485-5\\_7](https://doi.org/10.1007/978-3-319-91485-5_7)
36. Scheibe, K., Meschede, C., Göretz, J., Stock, W.G.: Giving and taking gratifications in a gamified social live streaming service. In: Proceedings of the 5th European Conference on Social Media, pp. 264–273. Academic Conferences and Publishing Limited, Reading (2018)
37. Scheibe, K., Zimmer, F.: Game mechanics on social live streaming service websites. In: Proceedings of the 52nd Hawaii International Conference on System Sciences, pp. 1486–1495. HICSS (ScholarSpace), Honolulu (2019)
38. Siuttila, M.: The gamification of gaming streams. In: Proceedings of the 2nd International GamiFIN Conference, pp. 131–140. CEUR-WS (2018)
39. Sjöblom, M., Hamari, J.: Why do people watch others play video games? An empirical study on the motivation of Twitch users. *Comput. Hum. Behav.* **75**, 985–996 (2017)
40. Strmečki, D., Bernik, A., Radošević, D.: Gamification in e-learning: introducing gamified design elements into e-learning systems. *J. Comput. Sci.* **11**(12), 1108–1117 (2015)
41. Thiebes, S., Lins, S., Basten, D.: Gamifying information systems: a synthesis of gamification mechanics and dynamics. In: Proceedings of the 22nd European Conference on Information Systems, pp. 1–17. Association for Information Systems, Atlanta (2014)
42. Wilk, S., Wulfert, D., Effelsberg, W.: On influencing mobile live video broadcasting users. In: Proceedings of the IEEE International Symposium on Multimedia, pp. 403–406. IEEE Computer Society, Washington, D.C. (2015)
43. Wolf, T., Weiger, W.H., Hammerschmidt, M.: Gamified digital services: how gameful experiences drive continued service usage. In: Proceedings of the 51st Hawaii International Conference on System Sciences, pp. 1187–1196. IEEE Society, Washington, D.C. (2018)
44. Zichermann, G., Cunningham, C.: *Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps*. O'Reilly Media, Sebastopol (2011)
45. Zimmer, F., Scheibe, K., Stock, W.G.: A model for information behavior research on social live streaming services (SLSSs). In: Meiselwitz, G. (ed.) SCSM 2018. LNCS, vol. 10914, pp. 429–448. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91485-5\\_33](https://doi.org/10.1007/978-3-319-91485-5_33)
46. Zimmer, F., Fietkiewicz, K.J., Stock, W.G.: Law infringements in social live streaming services. In: Tryfonas, T. (ed.) HAS 2017. LNCS, vol. 10292, pp. 567–585. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-58460-7\\_40](https://doi.org/10.1007/978-3-319-58460-7_40)

# **Ethics and Misinformation in Social Media**



# A Two-Phase Framework for Detecting Manipulation Campaigns in Social Media

Dennis Assenmacher<sup>(✉)</sup>, Lena Clever, Janina Susanne Pohl, Heike Trautmann,  
and Christian Grimme

Information Systems and Statistics, University of Münster, Münster, Germany  
{dennis.assenmacher, lena.clever, janina.pohl,  
heike.trautmann, christian.grimme}@uni-muenster.de

**Abstract.** The identification of coordinated campaigns within Social Media is a complex task that is often hindered by missing labels and large amounts of data that have to be processed. We propose a new two-phase framework that uses unsupervised stream clustering for detecting suspicious trends over time in a first step. Afterwards, traditional offline analyses are applied to distinguish between normal trend evolution and malicious manipulation attempts. We demonstrate the applicability of our framework in the context of the final days of the Brexit in 2019/2020.

**Keywords:** Social campaign detection · Stream clustering · Unsupervised learning

## 1 Introduction

Social media has become an important infrastructure for modern information sharing and networking. In most developed countries, the majority of people are already connected via one or multiple platforms [6]. Even more important, decision makers like politicians or multiplicators like journalists are also an integral part of social media networks. These groups function as bridge between the social media ecosystem and the offline world outside social media. While politicians try to get in touch with the sentiment of public debates about their programs or decisions, journalists try to pick up stories and use the public sphere as additional outlet.

Quite logically, social media has become a central platform for campaigns. Politicians try to reach the public with their ideas, but in contrast to former media types, users can also reach politicians directly. Both can also try to initiate societal debates by placing topics. And when journalists pick up these topics because they seem of critical importance in social media, their reach goes even beyond the boundaries of the social media ecosystem.

As such it is of utmost importance not only for journalists but for the whole society to provide some transparency on campaigns in social media. This shall provide insights into the origins of and motivations behind an observed topic: is a

campaign organic or orchestrated (automatic as well as human-driven), i.e., who is participating in these campaigns? What means are employed when placing a topic?

These questions go beyond the challenge of classifying single accounts as social bots or humans. We have to consider interaction of actors and thus the complete (or a representative sample of the) data stream, which is produced on a social media platform. These analyses do no longer focus on singular accounts or a group of users but on the content produced over time. Clearly, the corpus of data that needs to be analyzed is far too large for human manual inspection. But also classical methods of data analysis are not capable to store all data and process it in real time. Real-time detection of possible campaigns, however, is necessary to not lag behind with analysis, when topics reach critical popularity. At the same time, we still need to verify whether campaigns are organic or artificial. This decision can usually not be made ad-hoc and often needs a deeper, sometimes even forensic analysis of campaign data.

In order to address both challenges at the same time, we propose a two-phase framework which supports both campaign and trend detection and a-posteriori in-depth analysis of respective data. Our idea integrates a stream-based unsupervised detection of critical topics and an independent, offline, and extendable analytics environment. This allows to instantly identify upcoming and important topics and subsequently analyze and verify their artificial character. Note that this approach should be considered as a human-in-the-loop support tool, where no automatic decision on a campaign’s quality is made. In principle, it is designed to enable detection and transparent analysis of current topics in many contexts, either the discovery of new and interesting topics or the fight against manipulation via artificial campaigns.

The rest of this work is structured as follows: the next section will summarize related research in the context of this work and then Sect. 3 will detail the two-step framework’s concept proposed in this paper. Section 4 shows the application of our framework in the context of the Brexit discussion two months before and at the final Brexit date at the end of January 2020. Finally, Sect. 5 summarizes and discusses the results of our work and provides some future perspectives.

## 2 Related Work

Social media has been discussed as environment for disinformation, manipulation, or deception for more than a decade [8] and since the Brexit decision in 2016 as well as the election of Donald Trump for president of the United States, social media is considered an important infrastructure for manipulating societies [3, 22]. Much effort has been put into the (computer-aided) detection of automation in social media. *Social Bots* are considered very potent actors in the distribution of disinformation [10, 11, 14, 18], and consequently, detection techniques for social bots have been (and still are) an important topic of research [9, 10, 13, 20]. While research started with a focus on the classification of single accounts as bots- or human-driven, some recent publications emphasize the importance of detecting collaboration of multiple actors [9, 12]. An exceptionally early proposal was

made by Lee [16] already in 2014 to discriminate campaigns into *organic* and *non-organic* ones. While the first arises from classic human interaction in social media the latter type of campaigns is promoted by artificial or automated mechanisms or purchased and supported by the social platform [16].

Campaign detection started with offline analysis of network data and topologies, the clustering of posted or shared content, and the investigation of topics' temporal development. All applied techniques and extracted features mainly aimed for supporting or enabling machine learning approaches. More recent detection approaches afterwards focused on the application of machine learning in campaign detection in order to identify characteristic patterns of organic and non-organic campaigns [10, 20].

However, there are some major disadvantages of (supervised) machine learning approaches in this context:

1. Models have to be trained using labelled data. Especially for campaigns in social media, this kind of data is usually not sufficiently available. An insufficient data base, however, makes the approaches imprecise.
2. The learned patterns can only capture the characteristics found in available input and learning data. That is, the machine learning approaches may become outdated and inflexible regarding new kinds of orchestrated campaigns.

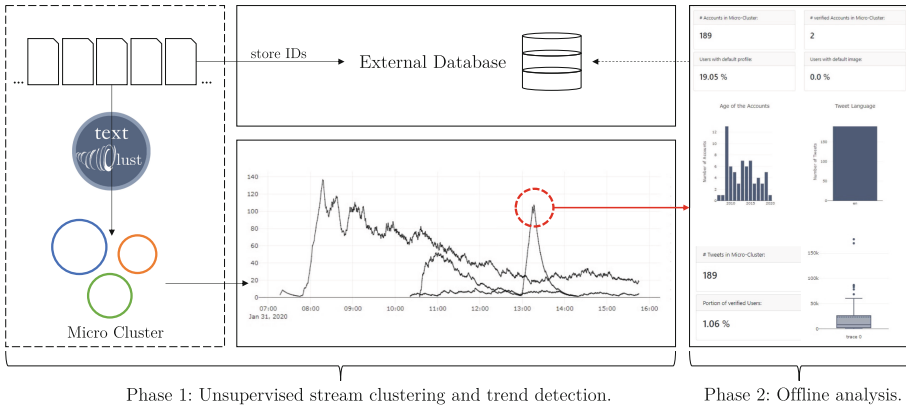
There is some recent work [7, 9, 23] which addresses the application of unsupervised detection methods like clustering and network analysis as solutions to some of the issues. These approaches do not need initial training and can detect unknown characteristics. However, as correctly pointed out in [23], these methods are computationally too complex to handle the observed amount of social media content in real-time.

In this work, we pick up a proposal we recently made, i.e. using stream-clustering approaches for topic detection [2] and apply it as a first step in a two-phase analysis process. We propose the augmentation of the detection of campaign candidates with a subsequent analysis phase. In this second phase, previous mentioned established group- or single account analysis can be applied to verify or reject whether a campaign is malicious or not and to possibly detect responsible actors in this campaign. As such, we consider this work as a step towards an integration of modern classification approaches into campaign detection for fast *and* precise transparency in social media communication.

### 3 The Two-Phase Framework for Detection and Analysis

In the following, we introduce our two-phase framework for automated campaign detection. The framework is depicted in Fig. 1. Within the first phase, the incoming text data stream (e.g. Twitter stream) is processed into *tfidf* vectors and aggregated via the `textClust` algorithm [5]. The algorithm handles text stream data and clusters similar documents together into so-called micro-clusters, which

represent the recently discussed topics of the stream. Additionally, a micro-cluster filtering is applied. By this, topics, which behave suspiciously in terms of their development over time, are extracted. In the second phase, these topics can be further analyzed, via numerous metrics and visual representations of text (meta-) data.



**Fig. 1.** 2-phase framework for analyzing suspicious cluster evolution

### 3.1 Phase 1: Text Stream Clustering

Stream clustering algorithms apply clustering on potentially unbounded data streams in an online fashion. The fact that the stream is potentially unbounded makes it impossible to store the complete data for calculations [4, 19]. Due to this, observations can be processed only once. As the complete range of the data is not known in advance, the stream clustering algorithm needs to be able to adjust clusters online and in real-time.

The stream clustering algorithm can be divided into two phases: In the online phase, micro-clusters are derived directly from the incoming observations. A micro-cluster is an aggregation of observations, which are locally dense. While the concrete observations are discarded after the distance calculations, the clusters are stored as representation of the actual data distribution. In the offline-phase, the respective micro-clusters can be clustered on-demand via traditional clustering techniques. This phase is independent from the online phase and can be scheduled on demand at any point in time. As here only the limited number of micro-clusters, as a representation of the original data is used, the calculations can be done by using the data multiple times.

In contrast to incremental clustering algorithms, stream clustering algorithms must be able to deal with the explicit notion of time. The complete range of data is not known at the beginning and the distribution of the stream data may

change over time (which is known as concept drift). Therefore, micro-clusters need mechanisms to adapt to changes in the data stream. To simulate a temporal drift, micro-clusters are usually weighted. The weight ensures that clusters, which are not updated by new observations for a while, will be decayed slowly. If the weight falls below a threshold, the cluster is removed completely.

**textClust:** The idea of micro-clusters as representation of stream data was originally designed for numeric data. Nevertheless, the idea can be transformed to textual data as well [1].

For our experiments, we use the textClust algorithm [5]. Within the textClust algorithm, the produced micro-clusters  $mc$  are represented as 4-tuples:

$$mc = (w, t, TF, ID)$$

The relative importance of a micro-cluster is reflected by its tokens  $t$  (namely most describing words) and its weight  $w$ . The weight is increased by 1 each time a new observation is allocated to the cluster. To be able to detect concept-drifts and account for temporal changes, the weight is exponentially decayed at each time step by

$$f(w) = w * 2^{-\lambda(t_{now}-t)},$$

where  $\lambda$  denotes the fading factor,  $t_{now}$  the current time and  $t$  the time the specific micro-cluster was last updated. A cleanup procedure is applied every  $t_{gap}$  time steps where all micro-clusters below a predefined threshold are removed from the clustering result. The same applies for all tokens within a respective micro-cluster.

The term frequency of representative cluster words as n-grams is denoted in the  $tf$  vector. Distance calculations between two micro-clusters using the cosine similarity are based on the  $tfidf$  vectors. Note, that the  $tfidf$  representation extends the traditional term frequency by weighting down words that appear in many documents, as they are considered to be less important. For every new observation, first a new micro-cluster is created and second, the distance to all other micro-clusters is calculated. If the new micro-cluster is in small distance (below a certain threshold  $r$ ) to one of the existing micro-clusters, it is merged with the respective cluster. Otherwise, the new micro-cluster remains and is added to the set of all micro-clusters.

The similarity of two  $tfidf$  vectors is calculated via the adjusted cosine-similarity. Within this metric, the average weight of the micro-cluster is taken into account. Therefore, each token (within a certain cluster) is weighted relative to the average weight. Let  $A$  and  $B$  represent two  $tfidf$  vectors from two different micro-clusters. The adjusted cosine similarity between them with their respective means  $\mu_A$  and  $\mu_B$  is then defined as follows:

$$\cos(\alpha) = \frac{\sum_i (A_i - \mu_A)(B_i - \mu_B)}{\sqrt{\sum_i (A_i - \mu_A)^2} \cdot \sqrt{\sum_i (B_i - \mu_B)^2}}$$

The fourth element within the micro-cluster definition  $ID$  captures the post IDs, which relate to the corresponding texts within a cluster. The post ID vector



is irrelevant within the clustering phase, but gets important in the second phase of the framework, when suspicious stream data is analyzed in more detail.

**Micro-cluster Monitoring to Detect Campaigns:** A micro-cluster represents a topic discussed in the text stream. Each cluster consists of tokens, which describe the content, as well as a weight, which represents the importance (number of associated text instances) of the cluster.

Next to the overall topic monitoring of the incoming stream data, we are especially interested in suspicious stream behavior. The identification of rapidly arising and growing clusters might be of interest in the field of trend or campaign detection. Especially, since we are interested in non-organic campaigns, driven by bots or trolls, the temporal evolution of the campaign can be used as an indication for unusual behavior [21]. Since it is not feasible to manually inspect the complete number of micro-clusters over time, an automated filtering step has to be applied. In an earlier work, we already proposed a method that reduces the number of micro-clusters by focusing on micro-clusters that do exhibit a significant change of weights within the last cleanup procedure [2].

In addition to storing only the actual weight  $w$  of the cluster, the weight before the last update  $w_{last}$  is included for calculating the difference  $\Delta_w = w - w_{last}$  within *tgap* cluster updates. Based on this, the average weight change  $\mu_w = \frac{\sum_i \Delta_{w_i}}{k}$  of all micro-clusters  $k$ , as well as the respective standard deviation  $\sigma_w = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (\Delta_{w_i} - \mu)^2}$ , can be computed. The Chebyshev's inequality is used to determine clusters with unusual weight patterns [17]. The inequality states that:

$$P(|X - \mu| \geq t \cdot \sigma) \leq \frac{1}{t^2},$$

where  $X$  is a random variable with expected value  $\mu$ , standard deviation  $\sigma$  and  $t$  any positive number. To ensure a feasible amount of clusters to (manually) analyze in a second step, we chose  $6\sigma$  ( $t = 6$ ) as threshold. The parameter setting can be adjusted depending on the context, as well as the underlying data. With this parameter setting about 3% of the micro-clusters are selected for further analysis, which is (in this case) a suitable amount for further investigations. The set of clusters of further interest  $I$  is thereby defined as:

$$I = \{mc \mid |\Delta_w - \mu_w| \geq 6 \cdot \sigma_w\}$$

### 3.2 Phase 2: Offline Analysis of Suspicious Clusters

Within the first phase of the framework, textual stream content is clustered and suspicious cluster evolution is filtered online and in real-time. In a second offline phase, suspicious clusters can be further examined. Here, all kinds of (computationally) expensive analyses can be applied. On the one hand, the micro-cluster content can be examined by the help of the stored cluster tokens. On the other hand, the user is able to gather meta-data via the *ID* vector of the suspicious

micro-cluster. As the *ID* vector captures all post IDs of the respective cluster, the Twitter REST API can be used to extract post meta data, e.g. the author ID or name. Further, meta data about the author can be gathered simultaneously. With the meta data the user is able to enrich the underlying data enormously. Especially for the detection of non-organic campaigns, further information about the human user is indispensable.

Authors of a micro-cluster can be analyzed regarding the age of their accounts, their post behavior, as well as their number of followers and followees. In the second phase of the micro-cluster analysis, visual representations can help to identify non-normal behavior. A dashboard can extremely help to visualize underlying structures in data and meta data of the post and accounts. Exploring e.g. the number of distinct accounts responsible for a micro-cluster, or checking the average age of the accounts, could help to identify social bots.

Furthermore, established bot detection methods can be applied. A well-known example for a bot detection method, which could be easily applied when the author ID is known, is the Botometer approach [20]. This tool gives an indication, whether an account is presumable steered by a human or a bot, by taking several meta data into account. Applying algorithms like the Botometer in the second phase of the framework can help to give an impression of the origin of the campaign and may help to detect non-organic campaigns.

In this work a first prototype of our dashboard is used for evaluation purposes<sup>1</sup> (see Fig. 2). We only rely on simple offline metrics which can be directly extracted from the tweets gathered during our experiments. Within the dashboard a variety of data and meta data can be visualized. For a first setting, we implemented figures and metrics representing the number of distinct accounts, the age of the accounts, such as the number of followers, and the percentage of verified accounts contributing within the specific topic. Further, we show how many an which posts are contained in this cluster at which point in time. This list is not exhaustive and can be complemented and customized. Up to now, we do not utilize additional supervised methods such as Botometer and leave this open for future research.

## 4 Case Study and Evaluation

In this work we exemplary demonstrate our framework in the context of the Brexit movement. For this purpose, we collected Twitter data by utilizing the platform’s Streaming API. Twitter proclaims that the API provides 1% of the global traffic produced by the platform. Preliminary experiments showed that by filtering specific hashtags (in this case we only filter out tweets containing the term **Brexit**), we are able to obtain almost a complete conversation history [7]. More precisely, we collected data in late 2019, before the Brexit (between 20th and the 27th of November) and on the actual Brexit day on the first of February. We explicitly removed retweets from our analysis since we want to identify trends

<sup>1</sup> A python implementation of `textClust` and the corresponding dashboard can be downloaded here: <https://textclust.com/>.

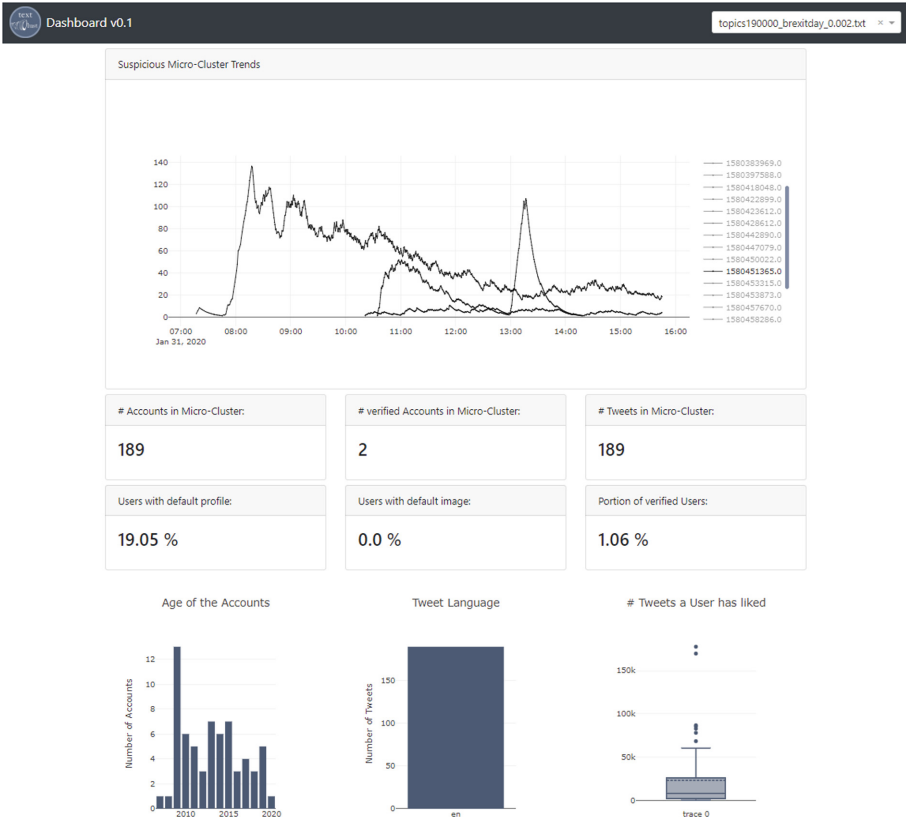
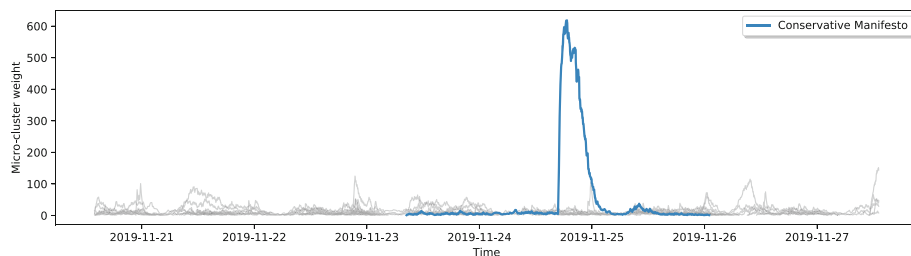


Fig. 2. Dashboard prototype to evaluate micro-cluster trends in the second phase

only based on original content excluding simply exaggerated trends based on retweet cascades [15]. In total we gathered roughly 1.3 million tweets, which were clustered by our `textClust` algorithm.

As specified in Sect. 3.1, the `textClust` algorithm requires some parameters that have to be set in advance. Especially  $\lambda$ ,  $r$  and  $t_{gap}$  do highly influence the final clustering result. The  $\lambda$  parameter affects how fast micro-clusters fade out over time and is thus responsible for the overall lifetime of a topic. While a small value ensures that micro-clusters, which are not frequently updated, are not immediately discarded from the set of all micro-clusters, a larger value dismisses them rigorously. Also,  $t_{gap}$  influences which clusters are discarded since a larger value leaves more time for potential micro-cluster updates (and cleaning). The distance threshold  $r$  affects the granularity of micro-clusters. While a large value merges *tfidf* vectors which are not necessarily very similar to each other (and therefore may represent different topics), a small value only merges sentences which are almost identical. The choice of suitable parameters does highly depend on the underlying data set. Therefore, we cannot rely on best-practice parameter



**Fig. 3.** Large micro-cluster that emerged from promoted Twitter campaign

settings. In context of our data set we systematically tested different parameter combinations. We found that  $\lambda$  also influences the number of identified trends. Since the Brexit day itself was very popular on Twitter with more than one million Tweets only on that day, we set a higher  $\lambda$  in this scenario. Therefore we decided to set  $\lambda$  to 0.001 (November) and 0.002 (Brexit day) respectively. We set  $t_{gap}$  to a fixed value of 100 and specified the distance threshold rather generously as 0.6. For all our experiments we used term-fading (fading according to elapsed time and not number of observations) to compensate variances in the stream throughput due to day/night cycles.

#### 4.1 Identification of Promoted Tweets

A quantitative evaluation of our approach is almost infeasible due to missing ground-truth data. In this proof-of-concept analysis we show that our framework is actually able to detect trending content within the Twitter stream. When we inspected the filtered micro-clusters from the data gathered between the 20th and the 27th of November, we identified one micro-cluster which exhibits a significantly higher cluster weight than all other ones (see Fig. 3). Consequently, we inspected this micro-cluster more in-depth, utilizing our Dashboard prototype. In total 1900 Tweets are assigned to that specific micro-cluster, with 1850 unique users. This implies that this unusual peak cannot be explained by single spamming accounts. However, we found that the message which was tweeted by all these different accounts is always exactly the same, motivating people to vote for the Conservative party to get the Brexit done (see Fig. 4). It has to be again emphasized that we explicitly excluded retweets from our clustering. Therefore, the observed phenomenon is an unusual distribution pattern. Since we have access to the original Tweet IDs, we inspected the Tweet more in detail. Interestingly, each of the Tweets in question consists of an additional button by which people are able to easily share the same content on their profile (via a new original Tweet) with one click. Further investigation revealed that this so-called *call-to-action* button is one feature of Twitter intended for businesses to reach their customers. Surprisingly, this feature also seems to be used in political context and has significant impact on the global conversation stream of that topic.

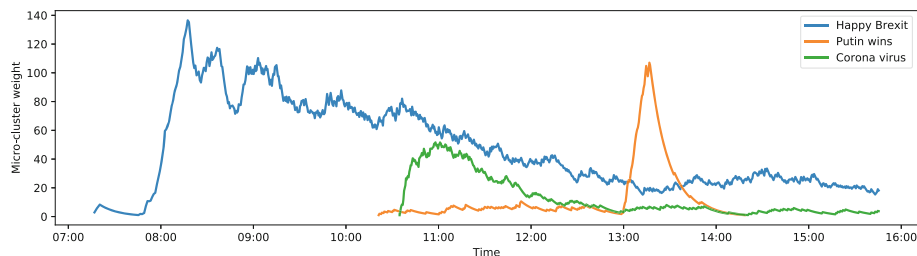


**Fig. 4.** Call-to-action button for promoted tweets

Despite the high cluster weight, the trend lasted only a few hours and completely faded out afterwards.

## 4.2 Organic vs. Non-organic Trends

While our filtering approach during the first phase drastically reduces the number of interesting micro-clusters, it is not guaranteed that all of them do exhibit non-organic trends that should be classified as malicious. In context of the actual Brexit day (first of February 2020), we exemplary show how normal evolving trends can be distinguished from non-organic ones and how the second phase of our framework supports this differentiation. Within Fig. 5, we display three micro-clusters which all represent different topics that were discussed on Twitter that day. The blue trace represents a micro-cluster, containing tweets where users simply wished a happy Brexit day (similar to birthday wishes). As it can be inspected in the Figure, the trend (increase of the micro-cluster weight) started approximately at 8:00 AM with its peak 15 min later. This is not surprising, since it simply reflects that people started posting about the Brexit after they woke up (at nighttime the tweet throughput is significantly smaller than during the day). After the peak of the micro-cluster it slowly fades out until the end of the day, implying that the throughput of newly arriving tweets decreases over time.



**Fig. 5.** Organic and non-organic micro-cluster trends at the Brexit day (Color figure online)

The green trace corresponds to a micro-cluster that summarized tweets about the first two cases of the corona virus in Britain which coincidentally happened at the same day. Again, the micro-cluster was created and immediately increases in its weight. Afterwards, similar to the Happy Brexit micro-cluster, the weight is slowly faded out during the day. The last micro-cluster established at 13PM and captures tweets about Putin which subliminally imply his involvement in the Brexit and that he finally wins. In contrast to the other two clusters, we observe a sharp weight edge with rapid fading after peaking.

While the first cluster is an appropriate example for an organic trend that naturally arose due to the topic relevance, the last two both are not easy to interpret, since they contain controversial content that may originate from targeted opinion manipulation. Again, we utilized our Dashboard prototype to inspect those micro-clusters more in-depth. The corona virus cluster in total consisted of about 300 tweets. All tweets were posted by different authors who mainly originate from the UK. Also, the actual content of the tweets differed from each other. Although the term corona virus was always included in the tweet, the wording was always different. However, most tweets embedded an external URL, which linked to a BBC article which was published one day before<sup>2</sup>. Using these insights, we conclude that the corona virus trend evolved also in an organic manner and was triggered by the newspaper article. Lastly, we inspect the cluster about Putin. Here, we observe completely different meta-data: First, all of the 320 tweets that were assigned to that cluster only originated from 60 accounts. Further inspection of the different users revealed that 124 tweets (almost 40% of the cluster tweets) were produced by one single account. The message which was posted by that account was always the same. The only difference was that each tweet mentioned different political individuals. Hence, we deduce that this micro-cluster resulted from a dedicated spamming attack by one single account. For crossvalidation, we used the Botometer service to check whether this specific account can be classified as a bot (automated program). Although the content score is slightly higher than average, Botometer classifies the account as human. However, as we already stated in preliminary work, the Botometer system can be

<sup>2</sup> The article can be accessed here: <https://www.bbc.com/news/health-51325192>.

fooled and it is furthermore not of the uttermost importance to identify whether an account is automated or not. The overall goal should be the identification of malicious coordinated campaigns, executed by humans or non-humans [12].

## 5 Discussion and Future Work

In this work we proposed a new two-phase framework that is capable of identifying artificially created and organic trends on social media stream data. By utilizing unsupervised stream clustering combined with an additional filtering approach, we can circumvent the problem of missing ground-truth data during the first online phase and simultaneously reduce the amount of unimportant data that has to be inspected manually. Within a second offline phase, we use meta-information that was persisted to secondary memory during clustering to get additional insights into the cluster contents. Within a Dashboard prototype the information is aggregated to valuable KPIs. Our experiments show that our framework is capable of identifying different types of trends. Ranging from simple spammers to coordination via multiple accounts, we revealed organic and non-organic trends that highly affected the overall discussion about the Brexit. We realize that the second offline step is necessary to get reliable insights regarding the type of trend and to verify or reject whether a campaign is malicious or not.

While we currently only employ simple aggregation metrics within the second phase of our framework, there is a lot of room for applying additional, more sophisticated analyses such as the identification of user networks. Upcoming research should also focus on optimal parameter configuration. Ideally, parameters should be automatically adjusted during the online phase. The insights from different cluster evolution can also be used to produce ground-truth data within a semi-supervised setting. Via the cluster filtering method, information of suspicious post development and account meta data is gathered. After validation, this data might serve as ground-truth in supervised campaign detection approaches.

**Acknowledgements.** The research leading to these results received funding by the Federal Ministry of Education and Research, Germany (Project: PropStop, FKZ 16KIS0495K), the federal state of North Rhine-Westphalia and the European Regional Development Fund (EFRE.NRW 2014–2020, Project: MODERAT!, No. CM-2-2-036a), and the Ministry of Culture and Science of the federal state of North Rhine-Westphalia (Project: DemoResil, FKZ 005-1709-0001, EFRE-0801431). All authors appreciate the support of the European Research Center for Information Systems (ERCIS).

## References








1. Aggarwal, C.C.: Mining text and social streams: a review. *SIGKDD Explor. Newsl.* **15**(2), 9–19 (2014). <https://doi.org/10.1145/2641190.2641194>
2. Assenmacher, D., Adam, L., Trautmann, H., Grimme, C.: Semi-automatic campaign detection by means of text stream clustering. In: *Proceedings of the Thirty-Three International Florida Artificial Intelligence Research Society Conference (FLAIRS 2020)*, Florida, USA. AAAI Press (2020). accepted
3. Bessi, A., Ferrara, E.: Social bots distort the 2016 us presidential election online discussion. *First Monday* **21**(11) (2016). <https://doi.org/10.5210/fm.v21i11.7090>
4. Carnein, M., Assenmacher, D., Trautmann, H.: An empirical comparison of stream clustering algorithms. In: *Proceedings of the ACM International Conference on Computing Frontiers (CF 2017)*, pp. 361–365. ACM (2017). <https://doi.org/10.1145/3075564.3078887>
5. Carnein, M., Assenmacher, D., Trautmann, H.: Stream clustering of chat messages with applications to twitch streams. In: de Cesare, S., Frank, U. (eds.) *ER 2017*. LNCS, vol. 10651, pp. 79–88. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-70625-2\\_8](https://doi.org/10.1007/978-3-319-70625-2_8)
6. Chaffey, D.: *Global social media research summary* (2019). <https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/>. Accessed 21 Feb 2020
7. Chen, Z., Subramanian, D.: An unsupervised approach to detect spam campaigns that use botnets on twitter. *CoRR abs/1804.05232* (2018). <http://arxiv.org/abs/1804.05232>
8. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on twitter: human, bot, or cyborg? pp. 21–30 (2010). <https://doi.org/10.1145/1920261.1920265>
9. Cresci, S., Petrocchi, M., Spognardi, A., Tognazzi, S.: On the capability of evolved spambots to evade detection via genetic engineering. *Online Soc. Netw. Media* **9**, 1–16 (2019). <https://doi.org/10.1016/j.osnem.2018.10.005>. <http://www.sciencedirect.com/science/article/pii/S246869641830065X>
10. Ferrara, E., Varol, O., Menczer, F., Flammini, A.: Detection of promoted social media campaigns (2016). <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13034>
11. Fredheim, R.: Putin’s bot army - part one: a bit about bots (2013). <http://quantifyingmemory.blogspot.co.uk/2013/06/putins-bots-part-one-bit-about-bots.html>
12. Grimme, C., Assenmacher, D., Adam, L.: Changing perspectives: is it sufficient to detect social bots? In: Meiselwitz, G. (ed.) *SCSM 2018*. LNCS, vol. 10913, pp. 445–461. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91521-0\\_32](https://doi.org/10.1007/978-3-319-91521-0_32)
13. Grimme, C., Preuss, M., Adam, L., Trautmann, H.: Social bots: human-like by means of human control? *Big Data* **5**(4), 279–293 (2017)
14. Hegelich, S., Janetzko, D.: Are social bots on twitter political actors? empirical evidence from a Ukrainian social botnet. In: *International AAAI Conference on Web and Social Media*, pp. 579–582 (2016). <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13015>
15. Kessling, P., Grimme, C.: Analysis of account engagement in onsetting twitter message cascades. In: Grimme, C., Preuss, M., Takes, F.W., Waldherr, A. (eds.) *MISDOOM 2019*. LNCS, vol. 12021, pp. 115–126. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-39627-5\\_10](https://doi.org/10.1007/978-3-030-39627-5_10)



16. Lee, K., Caverlee, J., Cheng, Z., Sui, D.Z.: Campaign extraction from social media. *ACM Trans. Intell. Syst. Technol.* 5(1) (2014). <https://doi.org/10.1145/2542182.2542191>
17. Mood, A.M., Graybill, F.A., Boes, D.C.: *Introduction to the Theory of Statistics*, 3rd edn. McGraw-Hill, New York (1974)
18. Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., Stieglitz, S.: Are social bots a real threat? an agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *Eur. J. Inf. Syst.* 1–19 (2019). <https://doi.org/10.1080/0960085X.2018.1560920>
19. Silva, J.A., Faria, E.R., Barros, R.C., Hruschka, E.R., Carvalho, A.C.D., Gama, J.: Data stream clustering: a survey. *ACM Comput. Surv.* 46(1), 13:1–13:31 (2013). <https://doi.org/10.1145/2522968.2522981>
20. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: detection, estimation, and characterization. In: *International AAAI Conference on Web and Social Media*, pp. 280–289. AAAI (2017). <https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817>
21. Varol, O., Ferrara, E., Menczer, F., Flammini, A.: Early detection of promoted campaigns on social media. *EPJ Data Sci.* 6(1), 13 (2017). <https://doi.org/10.1140/epjds/s13688-017-0111-y>
22. Woolley, S.: Automating power: social bot interference in global politics. *First Monday* 21(4)(2016)
23. Yang, K.C., Varol, O., Davis, C.A., Ferrara, E., Flammini, A., Menczer, F.: Arming the public with artificial intelligence to counter social bots. *Hum. Behav. Emerg. Technol.* 1(1), 48–61 (2019). <https://doi.org/10.1002/hbe2.115>. <https://onlinelibrary.wiley.com/doi/abs/10.1002/hbe2.115>



# Filter Bubbles and Content Diversity? An Agent-Based Modeling Approach

Poornima Belavadi<sup>(✉)</sup> , Laura Burbach , Patrick Halbach ,  
Johannes Nakayama , Nils Plettenberg , Martina Ziefle ,  
and André Calero Valdez 

RWTH Aachen University, Campus Boulevard 57, 52076 Aachen, Germany  
{belavadi,burbach,halbach,nakayama,plettenberg,ziefle,  
calero-valdez}@comm.rwth-aachen.de

**Abstract.** Personalisation algorithms play an important role in catering the information that is relevant to us. The best results are achieved by the algorithms when they monitor the user activity. Most of the algorithms adapt to the users' personal preferences by filtering out the information that is irrelevant to the user. However, one of the criticisms of this process is that it is leading to informational bubbles called the filter bubbles which is a personal space of content familiar to the user, which would reinforce their confirmational biases or create informational blind spots. This phenomena however is highly debated. In this light, we propose an agent based model study, which tries to verify the implications claimed by the filter bubble theorists and also create an hypothetical environment that does not have a filter bubble and test difference in the information dispersion and opinion formation in both the environments.

**Keywords:** Filter bubbles · Agent based modeling · Personalisation algorithms

## 1 Introduction

The Internet has a lot of information among which some would be relevant, some would be good, and some would be irrelevant to users. Sifting through all the information to find the one relevant to our interests, has become an essential need of users. This is addressed by Internet application providers. These providers rely on personalization algorithms or recommender algorithms to achieve this goal [6]. Social media websites, search engines, and other online applications work towards the goal of providing their users with the content that is interesting to them, for which they constantly monitor their users' activity. These recommender algorithms running in the background, filter out the information that seems irrelevant to the users' activity [6]. The performance of these algorithms to a large extent depends on users' activity or behavior. The other factors include the interaction of the algorithms with other algorithms, scalability of the algorithm, prediction accuracy, types of recommended items, etc [19].

At the same time, there seems to be something off with personalization. Users do not find their digital personal assistants to be right all the time. Most of the time users, seem to be losing their friends to the algorithmic abyss of social media's news feed. Many times the content users come across online seems to repeat the same topics [13].

### 1.1 Recommendation Algorithms and Filter Bubble

There is a general paradox that lies at the heart of personalization. Personalization is used as an aide to modify our interaction experience online concerning our interests. Simultaneously, our interactions online shape us, influence us and guide our everyday choices and actions. These incomprehensible algorithms sometimes make independent decisions on our behalf. Due to filtering, the number of visible choices is reduced thereby restricting our agency [13].

Eli Pariser coined the term “filter bubble” in 2011 in his book “Filter bubble – What the Internet is hiding from us”. [13] Pariser describes the filter bubble as a personalized information bubble that everyone is in. This bubble is the personal space that is not shared with others consisting mostly of the ideas and information that is interesting to us. It contains the different versions of the expected content that is presented to us by the different internet entities. However, being surrounded by the information that is only familiar to us and that is tailored to our tastes would deprive us of all the possible information that has been filtered out by the algorithms classifying them as unwanted. This would reinforce the confirmation bias that many of us already possess unconsciously [13]. Confirmation bias is the tendency to search for or interpret the information (either real or imagined) to confirm our previous ideas or views [19]. There is a general paradox that lies at the heart of personalisation. Personalisation is used as an aide to modify our interaction experience online with respect to our personal interests. Simultaneously, our interactions online shapes us, influence us and guide our everyday choices and actions. These incomprehensible algorithms sometimes make independent decisions on our behalf. Due to filtration, the number of visible choices is reduced thereby restricting our personal agency [13].

Our information bubble also exists offline. However, it becomes more apparent online as the user reactions can be magnified on a virtual context [13]. In this paper, we try to broaden the understanding of the filter bubble effect by developing agent based models to study whether the filter bubble affects the opinion formation in the society and how different would the opinion formation be affected if there was no filter bubble.

## 2 Related Work

The place where the filter bubble would be an advantage is the e-commerce applications, where narrowing the search results to match the preferences of the user is very critical. This increases the chances of matching potential products to its buyers [13]. However, the disadvantages of the filter bubble become evident

when it is connected with the process of fostering one's creativity. As being surrounded by familiar point of view would not provoke one's anxiety or instigate the curiosity that encourages to discover different view points [13,20].

This has given rise to the common criticism of recommendation algorithms in the recent years, that the algorithms may be responsible for causing filter bubbles as they filter the content choices over time effectively leading to polarised preferences [18]. However, this claim is in dispute, as Flaxman et al. found evidence that the recommendation algorithms both increase and decrease various aspects of filtering that leads to polarisation [8]. This has encouraged the research towards understanding the different design aspects of the recommendation system. In the literature, we find two common response to the filter bubble problem: algorithm centered and user centered [18].

## 2.1 Algorithm Based Approach

A conventional algorithm-based approach for the filter bubble is to develop more diversity aware recommendation algorithms [18]. The research mainly focuses on improving the diversity, novelty and relevance of the algorithms. The methods proposed are topic diversification approach [26], user centered clustering [1]. Many of the approaches proposed, although increased the diversity, affected the accuracy of the recommendation. The challenge in the research thus is to propose a method that improves the diversity of the recommendations without hampering the recommendation accuracy [18]. Smyth and Bridge found diversity based on the hamming distance on whether or not the items had been rated helped retrieve a target item most efficiently [16].

## 2.2 User-Based Approach

In the user based approach, the focus is more on developing diversity aware interfaces, where users receive the justification for the recommended item. Although developing such interfaces helped in tackling the filter bubble phenomena, it does not solve it completely [17]. The work of [11] showed that visualisation interfaces were used to increase the users understanding of the filter bubble phenomena. Still, it did not make an impact in trying to reduce it. It was however found that increasing the trust of the users by developing the interfaces that aid in the understanding of the recommendation system helped the users in giving better feedback that was in turn used in increasing the recommendation quality [9,17].

## 2.3 The Filter Bubble Debate

The topic of the filter bubble has divided the scientific community into two groups. There is an ongoing debate about the phenomena, as one community believes that reduced diversity in the information caused by the recommendation algorithms, to an extent where the challenging or the controversial content disappears virtually from the viewing systems of the users is constructing these

bubbles [27]. The other community, however, is concerned about the lack of scientific evidence for this phenomena. They claim that the amount of scientific evidence that filter bubble is leading to polarisation is not enough as the algorithm users can navigate through the information to identify the relevant information, thereby being the gatekeepers of the information they consume. In other words, this would give rise to highly individualised gates for the information that fit each users interests [10].

## 2.4 Impact of the Filter Bubble

Both the approaches mentioned above have a common goal of trying to reduce the effect of filter bubble either by improving the algorithms or by developing better interfaces. It is equally important to study the extent of the impact of the filter bubble. Nguyen et al., examine the longitudinal implications of recommender system on users and measure the filter bubble effect in terms of content diversity at the individual level [12]. Though this was a long term study, it had two exciting results, the users who used recommender systems found that the recommendations narrowed over time. But, the users who consumed the items recommended had reduced narrowing effects [12]. We try to address similar questions as Nguyen et al. [12], with the focus of opinion formation as the impact factor by using Agent-based modeling, as agent-based models fit perfectly for studying the emergent phenomena like filter bubbles.

## 3 Method

It is challenging to model human behavior. When building an agent-based model the challenge is to make the trade-off between how simple and traceable the model should be and how realistic and psychologically plausible should the behavior of the agents be modeled. We cannot find much guidance in this respect theoretically as the existing theories on human behaviour are mostly contradictory [7]. In this paper, we develop two simple agent-based models. The agents are characterized by attributes derived from theoretical approaches. The agents are modeled to be boundedly rational. They exhibit this behavior in the different cognitive levels of information processing and the different levels at which the consumed information could be effective. The focus on developing models lies in the implication of information distribution in filter bubble phenomena. But, the model can be expanded to include empirical studies as well as to study the different environmental factors like the social networks and influence of other agent's opinion.

### 3.1 Bounded Rationality - Decision Making of Agents

In many agent-based models, agents use multicriteria evaluation problems, for example in an agent-based simulation of planting crops, agents make the decision of choosing the land area in the simulation environment [4]. The main challenge

in modeling these agents that represent a real scenario of human decision-makers is to figure out how each agent solves the multicriteria problem of choosing locations in the simulated landscape as a function of varying spatial parameters with respect to the production activity. When evolutionary programming is used to solve this problem, the agent's decision making is represented as a form of bounded rationality [7].

Perfect rationality was one of the common theories of social sciences. However, a number of new alternate theories are being popular now, one of which is bounded rationality. While statistical regression models are used to express perfect rationality. Bounded rationality is best implemented with evolutionary programming. Bounded rationality was introduced by Simon, as the "rational choice that takes into account the cognitive limitations of the decision-maker - limitations of knowledge and computational capacity" [15]. We implement the bounded rationality in our model by introducing two attributes to the agent - cognitive threshold and effective threshold. The threshold values being generated randomly to represent the real scenario.

### 3.2 The Agent-Based Model

As mentioned above, we develop two agent-based models in this paper - one to model the filter bubble environment, the other to model the environment with no filtering. We use the LightGraphs package [14] and the Barabasi Albert model for network simulation [24]. The language used to write the simulation model is Julia [3]. The focus is to study the opinion formation in the network. We compare the two models to address the following questions: Does the filter bubble cause a significant impact on opinion formation? How is the opinion formation pattern different from the network with no filter bubble? How easy or difficult is it to get out of the filter bubble effect?

**Model1: Filter Bubble.** For the purpose of simplification, we make the following assumptions: the interaction of agents is only limited to sharing the topics or the messages. If an agent notices the message and then shares it, that signifies an opinion change. For implementing the message filtering, we have created messages initialized with a cognitive and affective value, the affective value being the message weight. If the message is shared by the agent the weight of message is increased by 10% and if the message is ignored by the agent then the weight of the message is decreased by 10% and if the message is seen by the agent then the weight of the message is kept same. Every agent has a bubble threshold, limit to interact with the message. A message is consumed by the agent only if the cognitive value of the message is higher than the bubble threshold. We use Mersennetwister to generate the random numbers that are allotted to the threshold values and to keep the values between 0 and 1.

**Model2: No Filter Bubble.** We keep the same assumptions mentioned in the filter bubble model here. For implementing the no filtering of messages we try three scenarios: first, we ran simulations initializing the bubble threshold to 0. Like in the filter bubble model, the weight of the message increased by 10% when the message was shared, decreased by 10% when the message was ignored and remained the same when the message was read. In the second case, we ran simulations with a bubble threshold value, but the weights of the messages did not change when the message was shared or ignored. In the third case, we ran simulations with the bubble threshold value set to 0 and no change in the weights of the messages.

All data is available at the open science foundation repository under <https://osf.io/xvna6/>.

## 4 Results

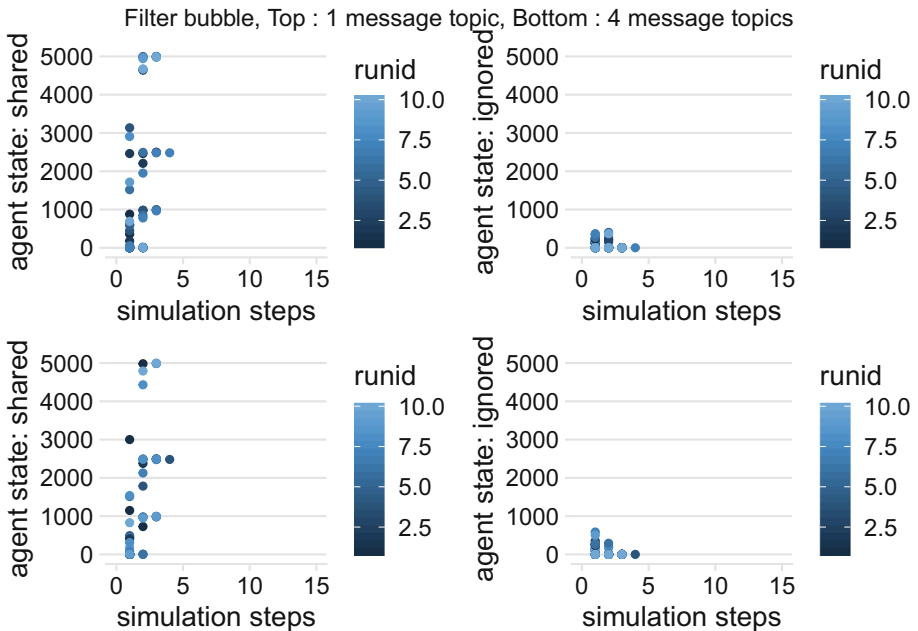
We ran each simulation with the agents ranging from 1000, 2500 and 5000 in 10 batches and 15 steps. We ran the simulations for two main cases: one where only one message was posted by the source agent and second where four different messages were posted. The simulations were run for both the filter bubble and no filter bubble. For the no filter bubble, three subcases were tested. The first case was with bubble threshold initialized to 0, the second case was with topic weights not being modified dynamically, the last case was with both bubble threshold set to 0 and topic weights not modified.

### 4.1 Simulation Run

The initial setup of the simulation experiment was initialized to have 1 message topic. An agent can be in one of the following states: read, sharing, shared, ignored and new. The simulation was run as explained in the steps below: 1. One agent is selected as a “source agent” at random. This agent spreads the messages and its state is new. 2. All other agents become the “target agents” that receive the messages with the state read, the agents that do not receive the message will have the state ignored. 3. The target may or may not choose to share the received message. 4. If the target agent decides to share the message, its state is sharing and the agents that have shared a message in the previous step would be in the state shared. The simulation is run until there is no more source agent or when all the agents have received the message. The next round of simulations was run for 4 message topics, the topics were differentiated based on their associated values and weights.

### 4.2 Analysis of the Results

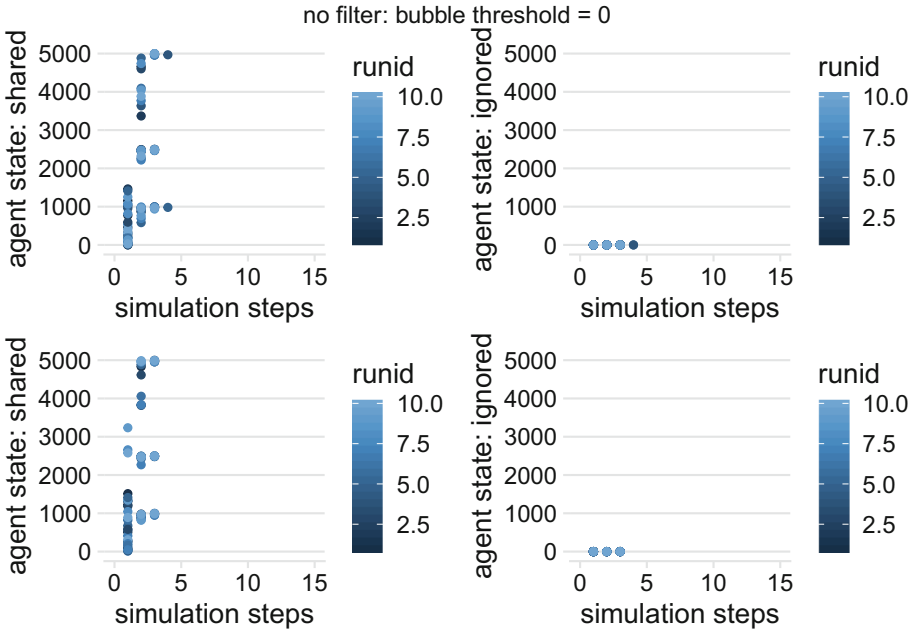
In Fig. 1 we see the graphs depicting the number of agents sharing the message and number of agents that did not receive the message in the filter bubble environment. From the plots, we can infer that all the agents have received a message by the end of the simulation as the number of agents in the ignored state is 0. An agent goes to the ignored state only when it does not receive a message.



**Fig. 1.** Plots showing the number of agents that have shared the message and the number of agents that did not receive any message in filter bubble environment. row 1: No. of message types = 1, row 2: 4 message types

The Fig. 2 shows the graphs depicting the number of agents sharing the message and number agents that did not receive the message in a no filter bubble environment for the first case, where bubble threshold is initialised to 0. From the plots we can infer that, when there is no bias from the agents then they would receive all the messages, even when the filtering mechanism is present in the system.

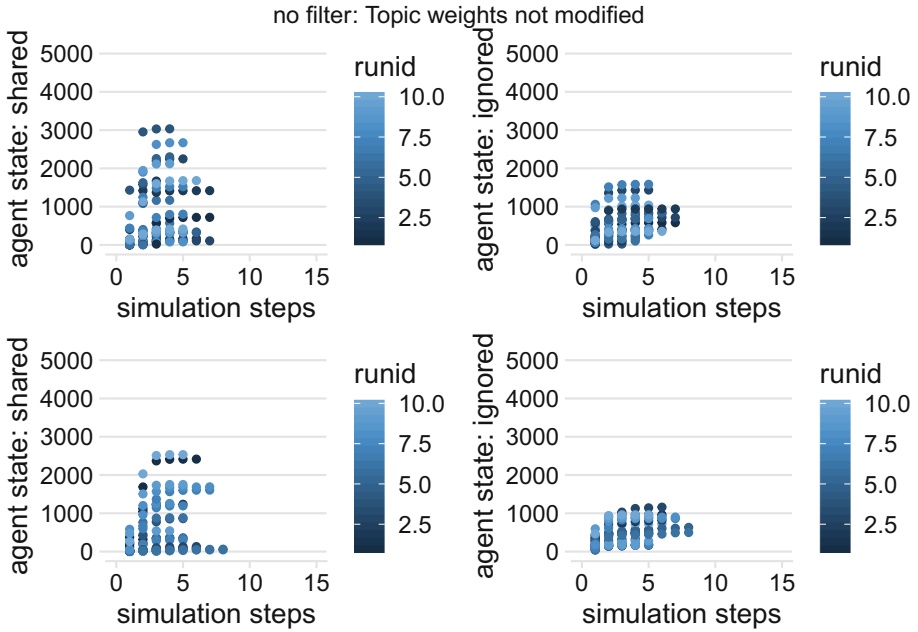




**Fig. 2.** Plots showing the number of agents shared the message and number of agents that did not receive the message in no filtering environment. Bubble Threshold = 0, row 1: No. of message types = 1, row 2: 4 message types

The Fig. 3 shows the graphs depicting the number of agents sharing the message and number agents that did not receive the message in a nofilter bubble environment for the second case, where the topics weights were not changed dynamically. From the plots it can be inferred that even when the filtering of the information is turned off, all the agents do not receive the messages. Even by the end of the simulation, the number of agents in the ignored state is not 0.

The Fig. 4, shows the graphs depicting the number of agents sharing the message and number agents that did not receive the message in a no filter bubble environment for the third case, where both the bubble threshold was initialised to 0 and the topic weights were not changed dynamically. From the plots it can be inferred that all the messages are received by the agents, as 0 agents stay in the ignored state from the first simulation run. This could be called an ideal case scenario, where there is no kind of filtering of information and no initial biases among the agents.

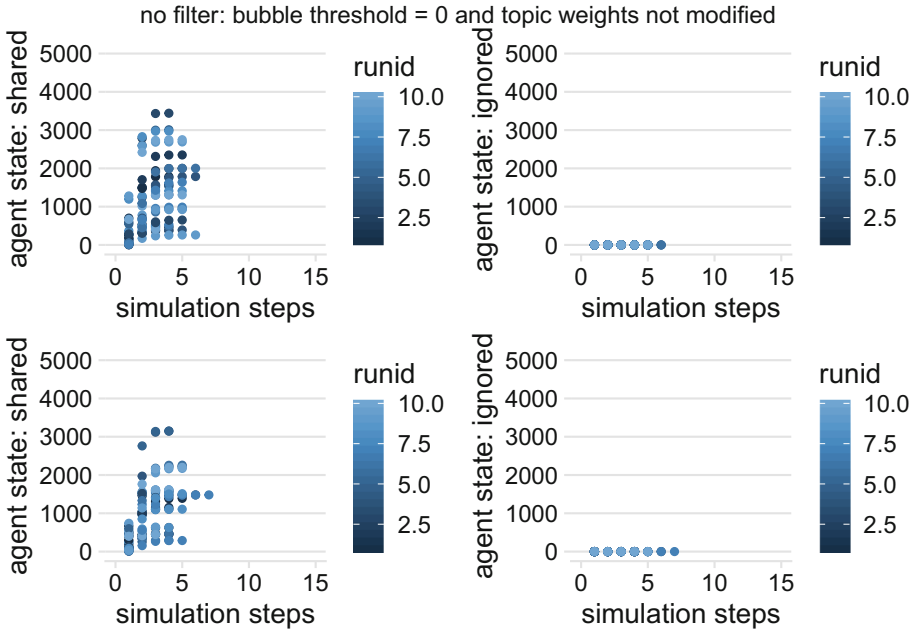


**Fig. 3.** Plots showing the number of agents shared the message and number of agents that did not receive the message in no filtering environment. Topic weights not modified, row 1: No. of message types = 1, row 2: 4 message types

## 5 Discussion

In this paper, we have attempted to simulate two environments (filter bubble and no filter bubble) and differentiate their impact on the consumption of information - impact on the diversity of the information. We see that the only way to achieve minimum filtering leading to maximum diversity in the information is when there is no threshold on the agents consuming the information and no bias formed after the consumption of the information which is the ideal case or when the agents have no bias. However, we see that the messages are filtered when there is a bubble threshold or weights associated with the message. It is interesting to note that, by the end of the simulation, all the agents in the filter bubble environment receive messages, implying that no agent remains in the ignored state. Whereas, in the no filter bubble environment where the topic weights are not modified dynamically, the agents remain in the ignored state at the simulation end. This is interesting as the bubble threshold indicating the agent's personal bias is taken into consideration and the topic weight modification representing the filtering of information is stopped making all the information reach every agent.

We have kept the model primitive, focusing only on the message (information) filtering aspect of the phenomena. It would be interesting to see the outcome when other variables like the influence of bias of the agent on one another, the



**Fig. 4.** Plots showing the number of agents shared the message and number of agents that did not receive the message in no filtering environment. Bubble threshold = 0 and topic weights not modified. Row 1: No. of message types = 1, row 2: 4 message types

different algorithmic filtering, topic interests of the agents are considered. We would like to improve the model by introducing the variables mentioned in the future.

Though the scientific evidence for filter bubbles is not enough. It does not mean that there is no reason to be concerned about the underlying problems. The more important matter that lies here is the concern about the algorithms that run in the background and the impact of the new data-driven forms of communication on the diversity of the content consumed in the media. The increasing importance in the role of social media in the exchange of information. Finally, when we consider the filter bubble, it is equally important to see the diversity of the information within the bubble - inclusion effect as it is to see the exclusion effect - the amount information that was filtered out because of algorithmic filtering, user interests, and other reasons.

**Acknowledgements.** This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia. We would further like to thank the authors of the packages we have used. We used the following packages to create this document: `knitr` [23], `tidyverse` [21], `rmdformats` [2], `kableExtra` [25], `scales` [22], `psych` [R-psych], `rmdtemplates` [5].

## References

1. Abbassi, Z., Mirrokni, V.S., Thakur, M.: Diversity maximization under matroid constraints. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 32–40 (2013)
2. Barnier, J.: rmdformats: HTML Output Formats and Templates for ‘markdown’ Documents. R package version 0.3.6 (2019). <https://CRAN.R-project.org/package=rmdformats>
3. Bezanson, J., et al.: Julia: a fresh approach to numerical computing. *SIAM Rev.* **59**(1), 65–98 (2017). <https://doi.org/10.1137/141000671>
4. Bonabeau, E.: Agent-based modeling: methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci.* **99**(suppl 3), 7280–7287 (2002)
5. Valdez, A.C.: rmdtemplates: rmdtemplates - an opinionated collection of R markdown templates. R package version 0.3.0.0 (2019). <https://github.com/statisticsforsocialscience/%20rmd.templates>
6. Valdez, A.C., Ziefle, M., Verbert, K.: HCI for recommender systems: the past, the present and the future. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp. 123–126 (2016)
7. Ebenhoh, E.: Agent-based modeling with boundedly rational agents. In: Handbook of Research on Nature-Inspired Computing for Economics and Management, pp. 225–245. IGI Global (2007)
8. Flaxman, S., Goel, S., Rao, J.M.: Filter bubbles, echo chambers, and online news consumption. *Pub. Opin. Q.* **80**(S1), 298–320 (2016)
9. He, C., Parra, D., Verbert, K.: Interactive recommender systems: a survey of the state of the art and future research challenges and opportunities. *Expert Syst. Appl.* **56**, 9–27 (2016)
10. Moeller, J., Helberger, N., et al.: Beyond the filter bubble: concepts, myths, evidence and issues for future debates (2018)
11. Nagulendra, S., Vassileva, J.: Understanding and controlling the filter bubble through interactive visualization: a user study. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media, pp. 107–115 (2014)
12. Nguyen, T.T., et al.: Exploring the filter bubble: the effect of using recommender systems on content diversity. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 677–686 (2014)
13. Pariser, E.: *The Filter Bubble: What the Internet is Hiding From You*. Penguin, London (2011)
14. Fairbanks, J., Bromberger, S., other contributors: JuliaGraphs/LightGraphs.jl: an optimized graphs package for the Julia programming language (2017). <https://doi.org/10.5281/zenodo.889971>
15. Simon, H.A.: *Models of Bounded Rationality: Empirically Grounded Economic Reason*, vol. 3. MIT Press, Cambridge (1997)
16. Smyth, B., McClave, P.: Similarity vs. diversity. In: Aha, D.W., Watson, I. (eds.) *ICCBR 2001. LNCS (LNAI)*, vol. 2080, pp. 347–361. Springer, Heidelberg (2001). [https://doi.org/10.1007/3-540-44593-5\\_25](https://doi.org/10.1007/3-540-44593-5_25)
17. Tintarev, N.: Presenting diversity aware recommendations: making challenging news acceptable (2017)
18. Tintarev, N., Dennis, M., Masthoff, J.: Adapting recommendation diversity to openness to experience: a study of human behaviour. In: Carberry, S., Weibelzahl, S., Micarelli, A., Semeraro, G. (eds.) *UMAP 2013. LNCS*, vol. 7899, pp. 190–202. Springer, Heidelberg (2013). [https://doi.org/10.1007/978-3-642-38844-6\\_16](https://doi.org/10.1007/978-3-642-38844-6_16)

19. Tintarev, N., et al.: Same, same, but different: algorithmic diversification of viewpoints in news. In: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, pp. 7–13 (2018)
20. Calero Valdez, A., Ziefle, M.: Human factors in the age of algorithms. understanding the human-in-the-loop using agent-based modeling. In: Meiselwitz, G. (ed.) SCSM 2018. LNCS, vol. 10914, pp. 357–371. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91485-5\\_27](https://doi.org/10.1007/978-3-319-91485-5_27)
21. Wickham, H.: tidyverse: easily install and load the ‘Tidyverse’. R package version 1.3.0 (2019). <https://CRAN.R-project.org/package=tidyverse>
22. Wickham, H., Seidel, D.: Scales: scale functions for visualization. R package version 1.1.0 (2019). <https://CRAN.R-project.org/package=scales>
23. Xie, Y.: knitr: a general-purpose package for dynamic report generation in R. R package version 1.26 (2019). <https://CRAN.Rproject.org/package=knitr>
24. Yook, S.-H., Jeong, H., Barabási, A.-L.: Modeling the Internet’s large-scale topology. *Proc. Natl. Acad. Sci.* **99**(21), 13382–13386 (2002)
25. Zhu, H.: kableExtra: construct complex table with ‘kable’ and pipe syntax. R package version 1.1.0 (2019). <https://CRAN.R-project.org/package=kableExtra>
26. Ziegler, C.-N., et al.: Improving recommendation lists through topic diversification. In: Proceedings of the 14th International Conference on World Wide Web, pp. 22–32 (2005)
27. Borgesius, F.Z., et al.: Should we worry about filter bubbles? *Internet Policy Rev. J. Internet Regul.* **5**(1) (2016)



# The Law of Live Streaming: A Systematic Literature Review and Analysis of German Legal Framework

Kaja J. Fietkiewicz<sup>(✉)</sup> 

Heinrich Heine University, Universitätsstr. 1, 40225 Düsseldorf, Germany  
kaja.fietkiewicz@hhu.de

**Abstract.** With evolved streaming technologies and faster mobile broadband, more and more live streaming platforms emerge online and become very popular among the users. From general platforms for streaming everyday life (e.g., YouNow) or reporting on news events (e.g., Periscope), through platforms for streaming video games (e.g., Twitch) or certain artistic performances (e.g., Picarto), the range of the services became very wide. As in most social media domains and with new developments on the digital market, the question arises whether the new trends also bear new challenges and issues of legal or ethical nature. This study is a systematic literature review of international scientific research on live streaming and potential legal problems (N = 22) conducted in order to pursue this question. It also entails a short review of legal issues with live streaming in Germany, a country with relatively strict consumer laws (e.g., data privacy) as well as first laws aiming at getting better control over the social media companies and users (e.g., Network Enforcement Act). The most prevalent legal domain within research on live streaming are copyright and sports broadcasting laws. The still understudied areas appear to be privacy, personality rights, and youth protection regulations. The most prominent issue within German legal discourse is the classification of live streaming as a telemedia offer or a broadcast, the second one entailing more restrictions and requirements (e.g., a broadcasting license).

**Keywords:** SLSSs · Live streaming · Law · Copyright · Personality rights · E-Sports · Privacy · Sport broadcasting

## 1 Introduction

With rapidly developing technologies for content production, more widespread and faster mobile broadband connection, services like live streaming platforms become more and more popular. The fundamental concept of live streaming services entails a streamer (producer of the content) and his/her viewers (consumers of the content, who might interact with the streamer in different ways). In a live stream the interaction between streamer and viewer happens in real time. The consumption of the specific content is therefore time-bound. Other popular streaming technologies that are not “live” include

such services like video on demand (VOD), e.g., YouTube videos (as example of user generated content) or Netflix (as example of commercial entertainment product), or music streaming (e.g., Spotify, Apple Music). The focus of this study is set on the live streaming of content. Given its recent raise in popularity and intensified social interaction and user engagement, it can potentially entail more legal issues than the more traditional VOD.

The live streaming services exist either in form of standalone live streaming platforms (e.g., Periscope, Twitch or YouNow) or as live-video features embedded in other services (e.g., on Facebook or Instagram Live) [1]. With the emergence of these streaming services, information production and information consumption behavior of social media users substantially changed [1–4]. The traditional behavioral patterns usual for watching TV evolved from the “lean-back” media usage to a space- and time-independent consumption and, in terms of live streaming services, to new production and participation behavior [2, 3]. What exactly makes live streaming services so special? Unlike on many other social media platforms, here, the inter-user communication occurs synchronously. The streamers and the viewers communicate in real-time with no time delay [4], which in turn leads to a versatile social interaction and user engagement [1]. Also, the viewers’ motivation has a stronger social and community basis [2, 5, 6]. Furthermore, there is an increasing popularity of live streaming within the e-sports domain [7, 8].

Such developments on the social media market do not only have impact on user behavior and social interaction. In fact, these rapid shifts might have increasing influence on the global economy, politics and legislation. New services and platforms change the structures and conditions of the digital market and economy. Too rapid modifications can be problematic for legal authorities regulating it, e.g., in terms of the competition law [9]. Furthermore, there is an increasing interaction between new technologies or online services and the law. On the one hand, this interaction exists because the law is supposed to set some boundaries for the platforms (e.g., in terms of consumer or data protection, antitrust regulations etc.), while still facilitating the technological progress. On the other hand, new technologies are being developed to actually enforce the law [10], like the Network Enforcement Act passed in Germany in October 2017, obliging bigger social media players (like Facebook or YouTube, hence, Google) to establish an effective reporting system and specific protocols for reporting illegal content (e.g., hate speech) [11]. Another example is the new EU Copyright Directive passed in 2019, which, among others, indirectly encourages platform providers (especially the ones where copyright infringing content is often uploaded by the users, e.g., YouTube) to implement the controversial “upload-filters” [12], which might be seen as a pre-censorship measure.

Given the increasing popularity of live streaming services and the prevalent interaction between technological developments and the law, the question arises what are the legal concerns and challenges in the context of live streaming services? The first part of this research focuses on legal aspects addressed in the international scientific literature on live streaming. In the second part of this paper, publications regarding legal framework in Germany relevant for live streaming services are being summarized. Germany’s legal system is known to be stricter, e.g., in terms of data protection or copyright regulations, than in countries like the USA. Furthermore, with legislation like the Network Enforcement Act, German legislature showed the interest and willingness to take more

measures in regulating the social media market. A systematic literature review as well as legal review like this one can help detecting topics of interest within the research community, but more importantly, research gaps as well as disparities between the current scientific research and the legal practice.

## 2 Methods

In this study, the systematic literature review (SLR) [13] was conducted to identify legal issues and challenges in context of live streaming dealt with by international researchers. The SLR procedure was based on the approach by Okoli and Schabram [13]. The purpose of the literature review was the identification of legal areas relevant for live streaming research as well as contexts in which they are being investigated (e.g., sports broadcasting or education). The review was conducted by one researcher.

The preliminary search revealed rather limited number of publications on the specific topic of “live streaming services” and “law.” The first round of literature research was conducted on December 19<sup>th</sup>, 2019 in the scientific databases Scopus and Web of Science (WoS) as well as on Google Scholar. While Scopus and WoS are commonly used for literature review and indicate certain qualitative standard of the research, Google Scholar was added to the sources in order to include legal research literature from the USA (especially US-American legal departments publishing primarily on their websites; these articles are not indexed by Scopus or WoS, but might be thematically and qualitatively suitable). The search query was recursively adjusted after the first search revealed some relevant articles, from which further keywords were extracted and used to expand the original query (Fig. 1).

For the first practical screening, only articles in English or German, and the ones with a full text available, were included. Furthermore, articles where no legal aspects were addressed (e.g., when “law” was only mentioned parenthetically in one sentence and not elaborated in any further way), were excluded. Also excluded were papers focusing solely on the P2P (Peer-to-Peer) technology and not live streaming platforms in general.

As for the quality appraisal, included were journal articles, conference proceedings, books and book chapters. Online student theses, blog article or similar postings were excluded. In the second round of the literature research, the references of the relevant publications were screened for further possibly relevant studies (the same inclusion and exclusion criteria as in the first round of research applied). As can be seen in Fig. 1, the literature research yield a total of 109 results, from which 22 were deemed as relevant for further analysis.

For the systematic review, several information pieces were intellectually extracted from the relevant studies. The data extraction included the respective legal areas (e.g., copyright law, broadcasting law), the context of the legal examination (e.g., live streaming in education or live streaming of broadcasted sports events). Furthermore, if available, the territorial applicability of the concerned regulation or law was noted. Finally, if available, the discussed challenges and outlooks were consolidated. The synthesis of the studies was conducted by the respective legal area and context. The review was partially summarized in a tabular form and described in a short summary.

Afterwards, a review of German legal framework applicable to (social) live streaming, based on legal literature and resources retrieved from German legal data base



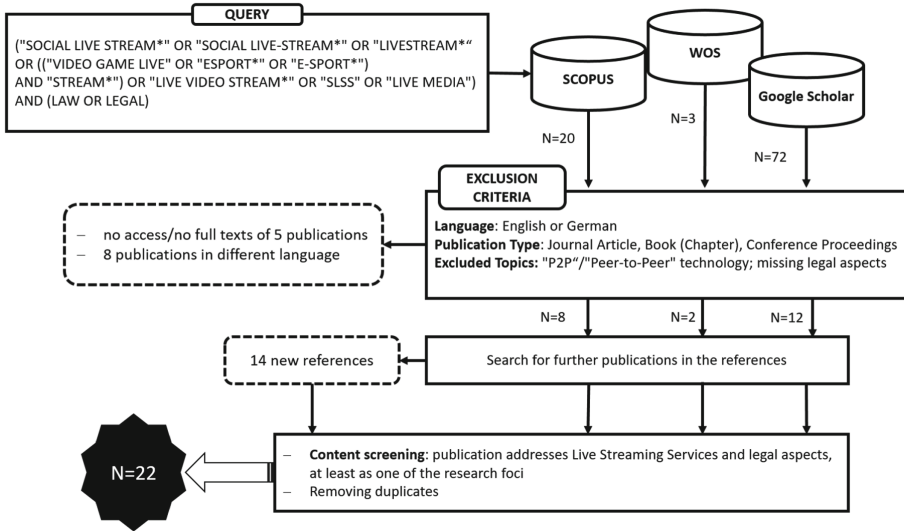


Fig. 1. Procedure of the literature search for the systematic literature review.

“beck-online”<sup>1</sup> was conducted. In the end, a comparison of the legal areas and possible challenges between international research and the legal status quo in Germany was performed.

### 3 Legal View on Live Streaming

#### 3.1 Results of the Systematic Literature Review

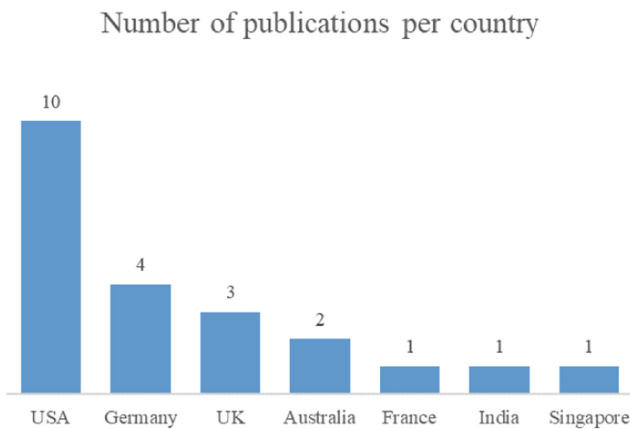
The literature research yield 22 relevant studies. During the search, no restriction regarding publication year was defined. The publications relevant for this review were published between 2011 and 2019 (Fig. 2), however, no publications on this topic were found for the years 2012 thru 2014. The highest research output was given in year 2017 and 2016. When looking at the countries where the publishing authors were based in (Fig. 3), most of the publications originated in the USA (10), followed by Germany (4) and UK (3). In general, the research output on this specific topic appears to be rather scarce, with the biggest research interest coming from the USA.

The reviewed literature was further classified by the legal area, context of the legal discourse and, if applicable, the regional focus (Table 1). Regional focus concerned primarily the territory where the discussed laws are applicable. As can be seen in Table 1, the most prominent legal area within the live streaming domain is the copyright law. Most of the reviewed publications either only discuss this legal area or refer to it together with other legal fields. Therefore, the summary of the literature will be structured primarily based on the context of the investigation (see Table 2 for shortly elaborated context categories). In Fig. 4 we can see a graphical depiction of the frequency of legal areas

<sup>1</sup> [www.beck-online.beck.de](http://www.beck-online.beck.de).



**Fig. 2.** Research on live streaming and law by year of publication (N = 22).



**Fig. 3.** Research on live streaming and law by country of the researchers (N = 22).

addressed in the reviewed literature. Despite copyright law, the mostly discussed areas were sports broadcasting laws and data privacy.

**Technology (Infringement Detection Systems).** Four of the reviewed studies include the description of a law infringement detection technology and/or method. Zhang, Song, Li, Zhang, and Wang [14] worked on a copyright infringement detection technology for live streaming videos. Live streaming poses a big challenge for infringement protection, since the content is generated in real-time, while most of the available protection techniques focus on or are better equipped for static content (e.g., ContentID by YouTube). The infringement detection within live streams is particularly difficult due to the more and more sophisticated behavior of the users, who are trying to bypass the detection system (e.g., by modifying the title or tweaking presentation of the video) [14]. Another aspect is the similarity of legal and copyright-infringing videos. Zhang et al. [14] worked on a solution for these challenges and developed a detection system called StreamGuard. It is

**Table 1.** Summary of the systematic literature review.

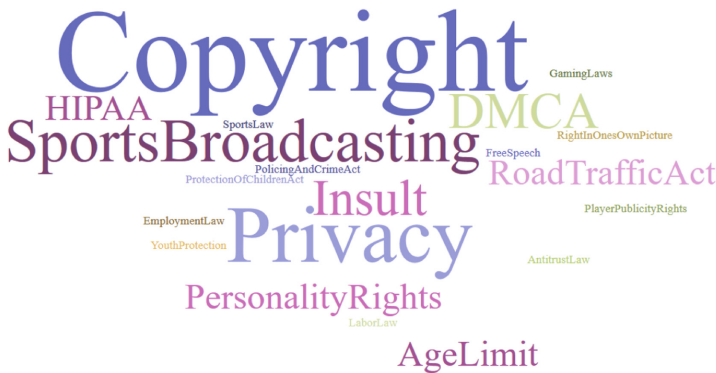
|      | Legal area(s)  | Context                             | Regional focus                            |
|------|--|-------------------------------------|---|
| [14] | Copyright  | Technology; Infringement detection  | N/A                                       |
| [15] | Copyright*   | Technology; Infringement detection  | N/A                                       |
| [16] | Copyright  | Discourse; Copyright issues         | EU  |
| [17] | Copyright  | Technology; Infringement detection  | N/A                                       |
| [18] | Copyright  | E-Sports; Legal discourse; Fair use | USA                                       |
| [19] | Copyright  | E-Sports; Fair use; Let's play      | USA                                       |
| [20] | Copyright  | Legal review; Webcaster's rights    | Italy, New Zealand, UK, USA               |
| [21] | Copyright  | Legal issues; Comparative analysis  | Australia, Canada, EU, Singapore, UK, USA |
| [22] | Copyright*; Free Speech; Privacy   | Discourse; Legal and ethical issues | EU, USA                                   |
| [23] | Copyright*; Privacy  | Ethical issues                      | N/A                                       |
| [24] | Copyright*; Privacy  | Education; Ethical and legal issues | USA                                       |
| [25] | Copyright; DMCA  | Sports broadcasting; Legal issues   | UK, USA                                   |
| [26] | Copyright, DMCA  | Sports broadcasting; Legal issues   | USA                                       |
| [27] | Copyright; Sports broadcasting   | Sports broadcasting; Legal issues   | Australia                                 |
| [28] | Copyright*; Sports broadcasting  | Sports broadcasting; Legal issues   | USA                                       |
| [29] | Sports-related laws  | E-Sports; Definition of "Sports"    | USA                                       |
| [30] | Copyright*; Gaming Laws; Player Publicity Rights; Antitrust Law; Labor Law; Employment Law; DMCA | E-Sports; Legal issues              | USA                                       |

*(continued)*

**Table 1.** (continued)

|      | Legal area(s)  | Context  | Regional focus |
|------|--|--|----------------|
| [31] | Copyright*; Age limit; Defamation; Personality Rights; Road Traffic Act; Privacy; Broadcasting of Sports Events; HIPAA | Law infringements; Content analysis of streams                 | Germany, USA   |
| [32] | Copyright; Personality rights; Sports broadcasting; Road Traffic Act; Insult; Data Protection; Age limit; HIPAA        | Law infringements; Content analysis of streams                 | Germany, USA   |
| [33] | Copyright; Right in one's own picture; Youth protection; Insult; Privacy   | Law infringements; Content analysis of streams                 | Germany, USA   |
| [34] | Policing and Crime Act; Protection of Children Act   | Technology; Examination methodology for Periscope; Child abuse | UK             |
| [35] | Privacy  | Legal issues; Privacy in public                                | USA            |

\*for the legal areas “Copyright” and “Intellectual Property” the term “Copyright” was used

**Fig. 4.** Legal areas addressed in the investigated literature.

based on an unsupervised Bayesian network and is supposed to overcome the mentioned challenges by “leveraging live chat messages from the audience.”

In another paper, Zhang et al. [17] again discuss the development of a crowdsourcing-based copyright infringement detection in live video streaming to address these challenges. Their system is based on investigation of clues from live chat messages. According to Zhang et al. [17], their scheme is significantly more effective and efficient than ContentID in detecting copyright infringing live videos on YouTube.

**Table 2.** Literature on live streaming and law by the context of investigation.

| Context                           | Literature       | Description   |
|-----------------------------------|------------------|---|
| 1. Technology (detection systems) | [14, 15, 17, 34] | Technology or methodology developed (and/or evaluated) to detect law infringements in live streams  |
| 2. E-Sports                       | [18, 19, 29, 30] | Legal issues (especially copyright) in the e-sports domain, hence, streaming of (commented) video games   |
| 3. Sports Broadcasting            | [25–28]          | Legal issues (especially copyright and broadcasting law) with re-streaming of live sports broadcasts  |
| 4. Education                      | [24]             | Implementation of live streaming for educational purposes and the accompanying legal implications   |
| 5. Content analysis               | [31–33]          | Actual analysis of the streams for possible law infringements; not based on automated approach, technology or methodology as under #1   |
| 6. Legal/ethical issues discourse | [16, 20–23, 35]  | Legal discourse, studies comparing different legislations and judicial systems; legal issues with live streaming in general not strictly falling under the remaining categories |

He, Maillé, and Simon [15], address the problem of illegal re-streaming of video streams. According to the authors, the existing solutions are based on watermarking the legal video to track the users who re-stream the stream on an illegal platform. Here, the problem is the fast detection of the illegal re-stream. He, Maillé, and Simon [15] worked on a fast and scalable algorithm for delivery of live watermarked video in content delivery network (CDN), which is supposed to expedite the process.

Horsman [34] addressed the challenges posed by the live streaming platform Periscope when detecting specific behavior in the stream and introduced an examination methodology for this application. This methodology is supposed to support those investigating cases of child abuse in a live stream. The practical enforcement of the offence of child abuse is still very difficult, the presented digital forensic analysis of Periscope's usage is supposed to facilitate the process.

**E-Sports.** The next four studies concerned e-sports (electronic sports), which is a form of competition using video games. With the rising popularity of these tournaments and video games in general, the streaming of (commented) video games on platforms like Twitch or YouTube (Let's Play) spread increasingly. Simultaneously, new legal issues (mainly copyrights-related) arise. Postel [18] discussed in his study, whether "fair use" defense is applicable in cases of Let's Play videos. According to Postel, a "court would most likely find that none of the four factors involved in the fair use defense analysis

support protecting ‘Let’s Play’ videos” [18]. In the end, he made a recommendation for a compulsory license for streamers in order to balance the interests between the gamer community and the video game companies. The same problem was addressed by Taylor [19], who also discussed the applicability of the “fair use” doctrine to “Let’s Plays.”

Holden, Edelman, and Baker III [30] conducted an in-depth analysis of legal issues that face the e-sports industry in the USA. Interestingly, the discourse addressed not only the copyright law, but also such areas as Player Publicity Rights, Antitrust Law, Labor Law or Employment Law.

Finally, Holden, Kaburakis, and Rodenberg [29], addressed the issue, whether e-sports falls under the definition of “sports,” which could have a meaningful impact on the application of a variety of different federal and state laws to this domain.

**Sports Broadcasting.** E-Sports, in terms of streaming of video games, is very popular content produced on streaming platforms. Another content type readily shared with help of the streaming technology are broadcasts of sports events. Here, again, the main problematic legal area are the infringements of copy- and broadcasting rights. In their study, Kariyawasam and Tasi [27] described the development of live streaming of sports broadcasts and its impact on the legal landscape. Furthermore, they addressed several uncertainties within copyright law in Australia. According to Kariyawasam and Tasi [27], “legislative reforms are required in order to balance the public’s ever-increasing desire for convenient ways to view digital materials against the legal rights of the owners of the material, while also aiming to maintain a forward-looking perspective in an attempt to foresee potential technological advancements that may pose considerable challenges to the traditional copyright law.”

Birmingham and David [25] also discussed the development of copyright infringement in sports broadcasting in the USA and UK, and conducted a comparison to copyright infringements in the music industry.

Edelman [28] discussed the potential impact of live streaming on the commercial sports industry. He analyzed, whether commercial sports enterprises have the legal power to stop live streaming of professional and collegiate sporting events (considering the federal copyright law, right of publicity law, and unfair competition doctrines).

Ainslie [26] examined the challenges that sports leagues and organizations face, when protecting their rights to live sports broadcasts. Here, the main focus lies again on the copyright law. Ainslie [26] discusses the “copyrightability” of sports telecasts, the direct copyright infringement liability for live streaming, the potential liability for websites and manufacturers under secondary infringement theories, and, the Digital Millennium Copyright Act takedown notices and safe harbor provisions [26].

**Education.** Even though there is a considerable amount of research on live streaming in education (in both, academic and medical context, the latter one being, e.g., live streams of medical procedures), only one publication on legal aspects has been found. Fuller, Mukhopadhyay, and Gardner [24] discussed live streaming for global pathology education and addressed several ethical and legal issues that need to be considered. As for the legal areas, they discussed the copyrights of the streamer and the privacy of the patients.

**Content Analysis.** Another topic addressed in three of the retrieved studies were the potential law infringements in the streams on popular live streaming services. The studies were based on content analyses of the streams and considered several possible law violations (like copyright infringement, insult, data privacy, etc.). According to Honka, Frommelius, Mehlem, Tolles, and Fietkiewicz [33], the most common law infringements observed in streams from Germany and the USA were copyright infringements, followed by infringing the right in one's own picture. Zimmer, Fietkiewicz, and Stock [32] conducted a content analysis of nearly 7,000 streams, from which 1,364 (almost 18%) contained possible law infringements. Most problematic area was again the copyright.

Finally, Scheibe, Zimmer, and Fietkiewicz [31] conducted a content analysis of streams from Germany, USA and Japan. The analysis was based on German legal framework, but also US-American legislation was addressed. In total, around 13% of the analyzed streams included some kind of potential law infringement. The most frequent one was, again, copyright infringement (around 80%), followed by not adhering to the minimum age restriction by the user (at least 13 years) and defamation.

**Discourse on Legal and/or Ethical Issues.** The remaining studies addressed in this literature review concern general ethical and legal discourse and do not fall under any of the aforementioned categories. Lim and Chik [21] examined the legality of Internet streaming and time-shifting technologies in the context of copyrights and in relation to the rights of 'reproduction' and 'communication to the public'. Their discourse is based on a comparative analysis of the jurisprudence in common law countries and the EU. According to Lim and Chik [21], the judicial decisions are not consistent, however, "they do provide some helpful guidance in our assessment of the relative strengths and weaknesses of the various arguments made on both sides of the divide as well as of the prevailing judicial sentiment towards new technologies."

Stewart and Littau [35] addressed the problematic issue of privacy and mobile video streaming. They discussed the gap between people's expectations of privacy in public and the wide-spread and uncontrolled use of mobile video streaming technologies, while the privacy laws cannot stand the pace of these technological developments. They covered the questions about whether people have any legal remedies when they are unwillingly the subjects of a live stream, or, whether the live stream can be protected under the First Amendment. For this purpose, Stewart and Littau [35] examined relevant court decisions and statutes as well as publications by legal scholars.

Sakthivel [20] addressed the popular topic of copyrights, however, took on a different perspective. He investigated the copyright protection of the webcaster (hence, streamer). He offered suggestions how to protect the webcasting (streaming) under the copyright law. This is an interesting take on the topic, since other relevant literature rather addressed the issue of copyright infringements committed by the streamer.

Faklaris et al. [22] explored the legal and ethical implications of (mobile) live streaming based on a review of public polices and human-computer interaction research. They identified relevant challenges in this area, namely, citizen videographers in conflict with the police (protection of free speech against unreasonable search and seizure), surveillance norms (e.g., "nanny cams") and voyeurism (both concerning the protection of privacy), and streaming of live events (i.e., protection of intellectual property, hence,

copyright). The authors also pointed out the problematic issue of global differences in legal aspects (here, especially between the USA and EU).

In their work-in-progress, Jung, Sell, and Stratmann [23] presented a design of an online Delphi study concerning ethical and legal issues of live streaming, which is supposed to involve international experts from seven different fields (like ethics, politics, law, journalism, software engineering, platform operators and users). Their study has two aims, first, to find a consensus about the ethical norms for live streaming in social media, and second, to discuss the derived codes of ethics for live streaming in social media in the context of ethical theories.

Finally, Borghi [16] conducted a discourse about copyright infringements in the streaming landscape by adhering to the legal situation in the EU. He discussed the European case law and several problematic aspects, like the public communication and reproduction rights. According to Borghi [16], the assessment of on-demand and live streaming through the lenses of copyright law is still unsettled.

### 3.2 Legal Issues with Live Streaming in Germany

The search in the German legal database “beck-online” revealed that the most prominent legal issue within live streaming domain to date is the German broadcasting law and the potential obligation of the streamers to apply for a broadcasting license. There are several critical aspects that need to be considered and clearly defined in order to classify a stream as a “broadcast.” Furthermore, the classification as “broadcast” is followed by more legal implications for the streaming industry, as compared to common telemedia offers [36, 37].

When it comes to broadcasting law, the main issue revolves around the question whether the regulations created for terrestrial broadcasting are transferable to the new services on the Internet without further amendments. Unlike in the past, today any content can be easily spread for simultaneous receipt via the Internet and, more importantly, it is no longer reserved for public and state controlled broadcasting companies, but is increasingly offered by private persons [37]. Indeed, some of them do not do it just for the fun of it, but actually earn their living this way (e.g., within the e-sports domain, [38]). What happens when these private streamers suddenly need to apply for a broadcasting license? Leeb and Seiter [37] investigated this issue in more detail by adhering to the press release by the German Commission on Licensing and Supervision (ZAK), according to which a stream called “PietSmietTV” was classified as broadcasting service and required a broadcasting license. Soon after, the Twitch-Channel “PietSmiet” also received such notice. The first channel (“PietSmietTV”) was streaming 24/7 and the content were reruns of already streamed videos (mostly Let’s Play). The second channel (“PietSmiet”) was a typical Twitch channel with no uninterrupted stream, but irregular live streams (with several days between each stream). Until the proceedings, both channels were viewed as license-free services. However, with classification as broadcast, further restrictions and sanctions can follow, e.g., content-related or of commercial nature. The only channel of “PietSmiet” not classified as broadcast remained the YouTube channel (video on demand) [37]. Hence, VOD is still classified as telemedia and not broadcast.

What is the difference between telemedia and broadcasts? According to the Interstate Broadcasting Treaty (RStV) Art. 2, broadcasting means a linear information and



communication service, provision and transmission of offers for the general public for simultaneous reception in moving images or sound along a schedule, using electromagnetic oscillations. Per definition, telemedia is an electronic information and communication service, which is not a broadcasting or telecommunication service. Based on this definition, VOD is easily not classified as broadcast, since the viewers decide by themselves when they want to watch the video. Hence, there is no linearity and simultaneity (viewers have technological influence on the times when the video is played). This also leads to lacking predefined schedule of the broadcast.

Unlike VOD, live-streams are intended for simultaneous reception by viewers at a certain time (chosen by the streamer and not the viewer). However, it is not clear whether they follow a predefined “schedule.” There is rarely a public announcement of a schedule for a live-stream (more like a sporadically streamed content announcement shortly beforehand). Another indicator can be the frequency and regularity of the streams, as it may indicate a consistent schedule. Here, more active streamers providing content regularly could fall under this definition and require a license. There are some exemptions, e.g., when the stream is offered to less than 500 potential users, however, meant is the technological possibility (which is usually given) and not the actual number of viewers. This requirement for an exemption will be rarely met on the Internet. The legal discourse and previous proceedings indicate rather vague and unclear definition of what is a regular broadcast streamed along a schedule. However, there is a possibility for streamers to have more legal certainty, because according to Art. 20 II RStV, providers of electronic information and communication services are entitled to apply for confirmation by the competent state media authority that a service would not raise objections under the broadcasting law [37, 39].

Finally, other legal areas discussed in the literature or being debated in court proceedings are the copyrights of the video games publishers (e-sports) [40], liability questions in case of re-streams of sports broadcast [41], and general re-streaming of a broadcast [42]. Finally, there is some attention given to the youth protection regulation, including restrictions to advertising (§ 6 JMStV, Interstate Treaty on the Protection of Human Dignity and the Protection of Minors in Broadcasting and in Telemedia), the obligation to appoint a youth protection officer (if the service is a broadcast and not telemedia), and establishment of time limits or other technical (pre-blocking) means so that children and youth of specific age groups cannot access any content that is development-impairing [43, 44].

## 4 Discussion

Recently, many new live streaming platforms emerged online and became very popular among different groups of users. As in most social media domains and with new developments on the digital market, the question arises whether the new trends also bear new challenges and issues of legal or ethical nature. This study was a systematic literature review of international scientific research on live streaming and potential legal problems (N = 22) conducted in order to pursue this question. Furthermore, it entailed a short review of legal issues with live streaming in Germany.

Based on the systematic literature review, the most prevalent legal domain within research on live streaming are copyright and sports broadcasting laws. Hence, the highest

interest is given to the commercial side of live streaming, where the regulations and their successful enforcement can lead to more financial gains. This is especially the case in the field of sports broadcasting and e-sports (video games). When considering studies on technologies for infringement detection, only one out of four papers focused on detection of child abuse, while the remaining addressed a better detection of copyright infringements. When looking at investigations including content analysis of streams, the most common infringements are indeed copyright related. Still, it is not feasible to analyze the entire population of live streaming users and the content they produce, which could give a more accurate picture of possible law infringements. There is a possibility that more problematic or disturbing content did not surface during the investigations in question. Furthermore, when disregarding commercial and/or financial importance of pursuing copyright infringements, the remaining legal areas of more ethical importance need to get more attention. These include the personality rights (e.g., data protection of people unknowingly and/or unwillingly included in a stream, protection against mobbing and defamation) as well as child and youth protection. The younger generation of Internet users is more at home when it comes to new trends on the social media market, however, they are also more vulnerable to predatory behavior and potentially more susceptible to inappropriate content. Therefore, more attention should be given to topics like detection of child abuse and control of development-impairing content.

The most prominent issue within German legal discourse was the classification of live streaming as a telemedia offer or a broadcast, the second one entailing more restrictions and requirements (e.g., a broadcasting license). This, however, does not only entail financial interests (e.g., the fee for broadcasting license), but also brings with it more restrictions and control regarding youth protection, which can be seen as a positive development. The retrieved commentaries and court proceedings addressed several critical issues, like the unclear definition of a broadcast and problems with a definite classification of live streaming, however, these will be most probably resolved in the coming months anyway. At the time of the investigation, the search in the legal database for live streaming-related publications did not yield any commentaries discussing the new Interstate Media Treaty (MStV), which will replace the Interstate Broadcasting Treaty and potentially clarify the legal situation of live streamers (e.g., there will be a minimum limit of 20,000 regular viewers to exclude smaller live streaming accounts). The bill will be voted on in the spring of 2020. The new regulation has to become effective in September 2020, since it constitutes the implementation of the amendments (Directive (EU) 2018/1808) to the European Audiovisual Media Services Directive. Future research should investigate the suitability and impact of the new regulation and reevaluate the legal certainty of live streamers.

In the future, more focus should be also set on content analysis (in order to get a better picture of currently pressing legal issues of live streaming) as well as areas other than copyright. The development of infringement detection technologies could entail more pursuit of detecting contents like child abuse or hate speech as well as technological apparatus for effective age control (restricting access to selected contents or platforms for children under 13 or other age groups).

## 5 Outlook and Limitations

The systematic literature review in this study had no restrictions regarding the publication year of the studies. However, new developments on social media market and legal developments (especially when considering court proceedings) are very time sensitive. Hence, it is questionable whether “older” research papers and the discussed issues are not already obsolete. This is especially noticeable for the review of German legal framework, where many of the published commentaries and proceedings concerned issues with the Interstate Broadcasting Treaty, which will be replaced in the second half of 2020 with the Interstate Media Treaty and, hopefully, most of the discussed problems for live streamers will be resolved. Another limitation could be the language of the analyzed papers, since there are many live streaming services in China and, potentially, many research papers on live streaming in Chinese. Furthermore, the choice of the search query as well as the three databases could have potentially restricted the number of retrieved studies. In the future research more focus should be set to current laws and court proceedings, e.g., in form of a comparative analysis. Furthermore, this legal review only focused on Germany. In the future other countries, or the general legal framework in the EU, should be considered.

## References






1. Fietkiewicz, K.J.: Guest editorial preface: special issue on live videos in social media. *Int. J. Interact. Commun. Syst. Technol.* **9**, vi–viii (2019)
2. Scheibe, K., Fietkiewicz, K.J., Stock, W.G.: Information behavior on social live streaming services. *J. Inf. Sci. Theory Pract.* **4**, 6–20 (2016). <https://doi.org/10.1633/jistap.2016.4.2.1>
3. Fietkiewicz, K.J., Stock, W.: Introduction to the minitrack on live streaming services. In: *Proceedings of the 52nd Hawaii International Conference on System Sciences*, pp. 2536–2537 (2019). <https://doi.org/10.24251/hicss.2019.305>
4. Fietkiewicz, K.J., Dorsch, I., Scheibe, K., Zimmer, F., Stock, W.G.: Dreaming of stardom and money: micro-celebrities and influencers on live streaming services. In: Meiselwitz, G. (ed.) *SCSM 2018. LNCS*, vol. 10913, pp. 240–253. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91521-0\\_18](https://doi.org/10.1007/978-3-319-91521-0_18)
5. Hilvert-Bruce, Z., Neill, J.T., Sjöblom, M., Hamari, J.: Social motivations of live-streaming viewer engagement on Twitch. *Comput. Hum. Behav.* (2018). <https://doi.org/10.1016/j.chb.2018.02.013>
6. Fietkiewicz, K.J., Scheibe, K.: Good morning... good afternoon, good evening and good night: adoption, usage and impact of the social live streaming platform YouNow. In: *Proceedings of the 3rd International Conference on Library and Information Science*, pp. 23–25 (2017)
7. Sjöblom, M., Törhönen, M., Hamari, J., Macey, J.: The ingredients of Twitch streaming: affordances of game streams. *Comput. Hum. Behav.* **92**, 20–28 (2019). <https://doi.org/10.1016/j.chb.2018.10.012>
8. Sjöblom, M., Hamari, J.: Why do people watch others play video games? An empirical study on the motivations of Twitch users. *Comput. Hum. Behav.* **75**, 985–996 (2017). <https://doi.org/10.1016/j.chb.2016.10.019>
9. Fietkiewicz, K.J., Lins, E.: New media and new territories for European law: competition in the market for social networking services. In: Knautz, K., Baran, K.S. (eds.) *Facets of Facebook: Use and Users*, pp. 285–324. De Gruyter, Berlin/Boston (2016). <https://doi.org/10.1515/9783110418163>

10. Specht, L.: Zum Verhältnis von (Urheber-) Recht und Technik. GRUR, pp. 253–259 (2019)
11. Kasakowskij, T., Fürst, J., Fischer, J., Fietkiewicz, K.J.: Network enforcement as denunciation endorsement? A critical study on legal enforcement in social media. *Telemat. Informatics*. **46** (2020). <https://doi.org/10.1016/j.tele.2019.101317>
12. Woollacott, E.: EU Copyright Directive Passed - Upload Filters and All. <https://www.forbes.com/sites/emmawoollacott/2019/03/26/eu-copyright-directive-passed-upload-filters-and-all/#2d3011e54c0f>
13. Okoli, C., Schabram, K.: A guide to conducting a systematic literature review of information systems research. *Sprouts Work. Pap. Inf. Syst.* **10**, 49 (2010). <https://doi.org/10.2139/ssrn.1954824>
14. Zhang, D.Y., Song, L., Li, Q., Zhang, Y., Wang, D.: StreamGuard: a Bayesian network approach to copyright infringement detection problem in large-scale live video sharing systems. In: *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pp. 901–910 (2019). <https://doi.org/10.1109/BigData.2018.8622306>
15. He, K., Maillé, P., Simon, G.: Delivery of live watermarked video in CDN: fast and scalable algorithms. In: *Proceedings of the 27th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video, NOSSDAV 2017*, pp. 79–84 (2017). <https://doi.org/10.1145/3083165.3083174>
16. Borghi, M.: Chasing copyright infringement in the streaming landscape. In: *IIC International Review of Intellectual Property and Competition Law*, vol. 42, pp. 316–343 (2011)
17. Zhang, D.Y., Li, Q., Tong, H., Badilla, J., Zhang, Y., Wang, D.: Crowdsourcing-based copyright infringement detection in live video streams. In: *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, pp. 367–374 (2018). <https://doi.org/10.1109/ASONAM.2018.8508523>
18. Postel, C.: “Let’s play”: YouTube & Twitch’s video game footage & a new approach to fair use. *Hastings Law J.* **68**, 1169–1192 (2017)
19. Taylor Jr., I.O.: Video games, fair use and the internet: the plight of the let’s play. *J. Law Technol. Policy* **2015**(1), 247–271 (2015)
20. Sakthivel, M.: Webcasters’ protection under copyright - a comparative study. *Comput. Law Secur. Rev.* **27**, 479–496 (2011). <https://doi.org/10.1016/j.clsr.2011.07.004>
21. Lim, S.C., Chik, W.B.: Whither the future of internet streaming and time-shifting? Revisiting the rights of reproduction and communication to the public in copyright law after Aereo. *Int. J. Law Inf. Technol.* **23**, 53–88 (2015). <https://doi.org/10.1093/ijlit/eav001>
22. Faklaris, C., Cafaro, F., Hook, S.A., Blevins, A., O’Haver, M., Singhal, N.: Legal and ethical implications of mobile live-streaming video apps. In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, MobileHCI 2016*, pp. 722–729 (2016). <https://doi.org/10.1145/2957265.2961845>
23. Jung, A.K., Sell, J.I., Stratmann, J.: Determining the ethical dimensions of live streaming: an explorative delphi study. In: *26th European Conference on Information System Beyond Digitization – Facets of Socio-Technical Change, ECIS 2018* (2018)
24. Fuller, M.Y., Mukhopadhyay, S., Gardner, J.M.: Using the periscope live video-streaming application for global pathology education: a brief introduction. *Arch. Pathol. Lab. Med.* **140**, 1273–1280 (2016). <https://doi.org/10.5858/arpa.2016-0268-SA>
25. Birmingham, J., David, M.: Live-streaming: will football fans continue to be more law abiding than music fans? *Sport Soc.* **14**, 69–80 (2011)
26. Ainslie, A.: The burden of protecting live sports telecasts: the real time problem of live streaming and app-based technology. *SSRN Electron. J.* (2016). <https://doi.org/10.2139/ssrn.2729641>
27. Kariyawasam, K., Tsai, M.: Copyright and live streaming of sports broadcasting. *Int. Rev. Law Comput. Technol.* **31**, 265–288 (2017). <https://doi.org/10.1080/13600869.2017.1299553>

28. Edelman, M.: From meerkat to periscope: does intellectual property law prohibit the live streaming of commercial sporting events. *Columbia J. Law Arts* **39**, 1–38 (2016)
29. Holden, J.T., Kaburakis, A., Rodenberg, R.M.: The future is now: Esports policy considerations and potential litigation. *J. Legal Aspects Sport* 46–78 (2017). <https://doi.org/10.2139/ssrn.2933506>
30. Holden, J.T., Edelman, M., Baker III, T.A.: A short treatise on esports and the law: how America regulates its next national pastime. *Univ. Ill. Law Rev.* **2020**(2), 509–582 (2020)
31. Scheibe, K., Zimmer, F., Fietkiewicz, K.J.: Das Informationsverhalten von Streamern und Zuschauern bei Social Live-Streaming Diensten am Fallbeispiel YouNow. [The information behavior of streamers and viewers on social live streaming services at the example of YouNow]. *Information-wiss. und Prax.* **68**, 352–364 (2017). <https://doi.org/10.1515/iwpp-2017-0065>
32. Zimmer, F., Fietkiewicz, K.J., Stock, W.G.: Law infringements in social live streaming services. In: Tryfonas, T. (ed.) *HAS 2017. LNCS*, vol. 10292, pp. 567–585. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-58460-7\\_40](https://doi.org/10.1007/978-3-319-58460-7_40)
33. Honka, A., Frommelius, N., Mehlem, A., Tolles, J.N., Fietkiewicz, K.J.: How safe is YouNow? – an empirical study on possible law infringements in Germany and the United States. *J. Macro Trends Soc. Sci.* **1**, 1–17 (2015)
34. Horsman, G.: A forensic examination of the technical and legal challenges surrounding the investigation of child abuse on live streaming platforms: a case study on periscope. *J. Inf. Secur. Appl.* **42**, 107–117 (2018). <https://doi.org/10.1016/j.jisa.2018.07.009>
35. Stewart, D.R., Littau, J.: Up, periscope: mobile streaming video technologies, privacy in public, and the right to record. *Journal. Mass Commun. Q.* **93**, 312–331 (2016). <https://doi.org/10.1177/1077699016637106>
36. Rundfunkrechtliche Zulassungspflicht für Live-Streams. *MMR*, 133 (2019)
37. Leeb, C.-M., Seiter, F.: Rundfunklizenzpflicht für Streaming-Angebote? *ZUM*, 573–581 (2017)
38. Törhönen, M., Hassan, L., Sjöblom, M., Hamari, J.: Play, playbour or labour? The relationships between perception of occupational activity and outcomes among streamers and YouTubers. In: *Proceedings of the 52nd Hawaii International Conference on System Sciences* (2019). <https://doi.org/10.24251/hicss.2019.308>
39. Martini, M.: TMG § 1 Anwendungsbereich. In: Gersdorf, H., Paal, P.B. (eds.) *BeckOK Informations- und Medienrecht*. Verlag C.H.Beck, München (2019)
40. Hentsch, C.-H.: Die Urheberrechte der Publisher bei eSport. *MMR*, 3 (2018)
41. Störerhaftung von EU-ausländischen Upstream-Providern für illegale Live-Streams von Spielen der Fußball-Bundesliga. *ZUM*, 67 (2016)
42. Unerlaubte Weitersendung via Internet. *ZUM*, 873 (2017)
43. Verstoß gegen den Jugendmedienschutz durch Live-Stream. *ZUM-RD*, 369 (2018)
44. Hopf, K., Brami, B.: Die Entwicklung des Jugendmedienschutzes 2016/2017. *ZUM*, 1 (2018)



# Social Media Use, Political Polarization, and Social Capital: Is Social Media Tearing the U.S. Apart?

James Hawdon<sup>(✉)</sup> , Shyam Ranganathan , Scotland Leman ,  
Shane Bookhultz , and Tanushree Mitra 

Virginia Tech, Blacksburg, VA 24061, USA

{hawdonj, shyam81, leman, sdbookhu, tmira}@vt.edu

**Abstract.** While some polarization is potentially beneficial for democracy, hyper-polarization can lead to political gridlock, tribalism, and even physical violence. Given the gravity of these concerns, we use data from 1,424 residents of Virginia, USA to investigate if media exposure is related to polarization. We explore if getting news from traditional media (e.g. television, radio, newspapers) or social media (e.g. Facebook, Twitter, news aggregators) predicts the likelihood of being polarized. Results reveal stark differences between liberals and conservatives. Polarized conservatives use radio talk shows and television for their news while polarized liberals are likely to get their news from newspapers, television, and various social media outlets. We then investigate if polarization influences social capital. We find that polarized conservatives express low levels of bridging capital while polarized liberals are more likely to express high levels of bonding capital. Media consumption also influences bridging and bonding capital. We also find that while being polarized does not predict civic engagement, media consumption does. We consider these results disturbing. At least among the political extremes, conservatives and liberals are informed by different sources. This lack of a shared information results in competing worldviews while providing little opportunity for finding common ground. This combination of high bonding, low bridging capital can explain the recent increase in “lethal partisanship” where groups not only disagree but also accept or even wish harm to their political opponents.

**Keywords:** Polarization · Social media · News sources · Social capital

## 1 Introduction

A group of foreign policy opinion leaders from government, think tanks, academia, the media, business, religious organizations, and NGOs cited political polarization as the most critical threat to the national security of the United States [1]. These experts fear that polarization is eroding the nation’s international standing and weakening its ability to lead efforts to confront global challenges. While some polarization is potentially beneficial for democracy, hyper-polarization often leads to political gridlock, tribalism, and the erosion of social capital. Polarization, so it is argued, rips at the fabric of society

and leads to a culture war, and hyper-polarization can result in social destabilization, civil unrest, and even physical violence.

Given the gravity of these concerns, it is important for us to ask what potentially leads to polarization. This research is designed to address this question by examining the relationship between media use and polarization and the relationship between political polarization and various forms of social capital. We begin by briefly considering polarization and its threats in the United States and cross-nationally. We then briefly review the relationship between media consumption and polarization, noting the potential role of “filter bubbles” or “echo chambers.” We then discuss the theoretical link between polarization and social capital. After this discussion, we use data from 1,424 residents of Virginia, USA to investigate if media exposure increases polarization. Specifically, we explore if getting news from traditional media sources (e.g. television, radio, newspapers) or social media sources (e.g. Facebook, Twitter, news aggregators) influences the likelihood of being polarized. We then turn to an analysis of the relationship between polarization and various forms of social capital and conclude by considering the implications of our research.

## 2 Literature Review

### 2.1 Polarization: Its Threats and Trends

Polarization is “a process whereby the normal multiplicity of differences in a society increasingly align along a single dimension, cross-cutting differences become instead reinforcing, and people increasingly perceive and describe politics and society in terms of “Us” versus “Them” [2]”. Even in cases where the extent of polarization is not extreme, it can lead to political gridlock [3, 4]. Taken to its extremes, polarization can lead to “lethal partisanship” where those on opposing sides rationalize harming their opponents, feel less restrained about harming their opponents, and feel less sympathy when their opponents are harmed or killed [5].

Noting the dangers of hyper-polarization, sociologists have worried about and tracked trends in polarization for some time [6–8]. The evidence suggests that polarization is increasing rather dramatically in the United States. For example, according to the Pew Research Center’s tracking of the values of the American public since 1994, political values are becoming more ideologically consistent and more strongly associated with partisanship. Overall, across 10 measures tracked by Pew, the average partisan gap increased from 15% points to 36% points between 1994 and 2017 [9]. Moreover, unlike in the past when political gaps were mostly based on education, religious attendance, gender, or race, the largest gap in values is now between political parties [9].

This hyper-polarization is occurring not only in the United States, but it is also evident in many nations. The rise of the hyper-nationalist Law and Justice (PiS) party in Poland reflects how Poland is now one of the most polarized societies in Europe [10, 11]. Similarly, the far-left Syriza party’s victory in Greece, the rise of Podemos in Spain, and the increasing shift toward rightwing politics in Austria, Germany, and the Czech Republic all demonstrate the growing levels of polarization in these nations [11–13]. The support for Brexit in England and the strong electoral performance of Le Pen’s far-right National Front in France, like the election of Donald Trump in the United States, also

reveal the rise of partisan populism and the polarization related to such populism [13]. Similar trends in political polarization have also been observed in Latin America, South Asia, and East Asia [14–17]. In short, polarization appears to be increasing around the globe.

## 2.2 Polarization and the Media

With respect to the relationship between media consumption and polarization, popular wisdom, mainstream media, and numerous scholars argue that social networking sites (SNS) such as Facebook and Twitter and online media sources such as news aggregators are driving polarization [18–21]. The 2016 U.S. elections and the use of bots to spread “fake news” and incite cultural debates are oft cited examples. Similarly, the rise of online “echo-chambers” or “filter bubbles” are also considered mechanisms by which social media contributes to polarization. The personalization of SNS refines users’ profiles to narrow the range of information they see, thereby reflecting their ideology and interests. The result of this personalization shrinks users’ social networks and exposure to competing information and alternative worldviews [18, 22, 23]. As this occurs, users develop online connections with likeminded people because people’s friendship networks tend to include those with similar habits, lifestyles, and cultural worldviews [24, 25]. This limiting of one’s online networks to include mostly those who share similar political views can lead people to believe their views are more widely held than they actually are and to discount those who disagree with them as being “out of touch” with what their network comes to know as “the truth” [22, 23].

There is evidence that suggests such echo chambers or filter bubbles do contribute to polarization [19–21, 26, 27]. For example, Bail and his colleagues [19] conducted a field experiment where participants followed bots that retweeted messages from political leaders and elected officials who held opposing political views. They found conservative participants became more conservative after following liberal bots, although the effect was not significant for liberal participants. Similarly, evidence suggests the network of political retweets are highly partisan and exhibit extremely limited connectivity between liberal- and conservative-leaning users [21]. In a different yet related line of research, Hong and Kim [20] found strong polarization when they analyzed the Twitter readership of members of the U.S. House of Representatives.

Although there is evidence that suggests social media can contribute to political polarization, an alternative “crosscutting interactions perspective” [19] argues that the openness of the Internet and social media allows for a variety of different opinions to be accessed and considered. As such, the use of SNS would not necessarily lead to polarization, and, in fact, could potentially increase political tolerance by allowing users to interact with and learn from those holding different political opinions. Because social media platforms facilitate interactions among individuals with weak ties who hold more politically heterogeneous perspectives than those in an individual’s primary networks, SNS use can actually increase political moderation instead of extremism [19]. There is also evidence supporting this position [19, 28–30]. For example, using a panel design to track the ideological composition of social media users’ online networks in Germany, Spain, and the United States, Barberá, [29] found most social media users to be embedded in ideologically diverse networks and this diversity was positively correlated



with political moderation. It is also noteworthy that political polarization has increased the most among the demographic groups who are least likely to use social media, thereby calling into question the overall claim that frequent SNS use leads to polarization [30]. At the least, this insight suggests that even if social media use is contributing to polarization, something other than SNS use is also driving the heightened levels of polarization.

One such additional driver of polarization is the increased political bias found in traditional forms of media such as newspapers and television news. Indeed, some scholars argue that these sources of media may have a far greater effect than social media on polarization [31–33]. Although most empirical findings consistently suggest that large, traditional media outlets in the United States express politically centrist views, some talk radio shows and cable news channels offer more ideologically extreme political opinions and presentations of news [33–35]. This finding can help account for increasing polarization among those groups who are not avid social media users. For example, Morris [32] finds that the network news audience is increasingly older Americans, and that the Fox News and CNN audiences are becoming increasingly polarized.

Consequently, there is evidence that both traditional and newer forms of media can contribute to polarization. Yet, polarization may not be as problematic as some suggest. Although hyper-polarization can lead to gridlock and even lethal partisanship, it can also simplify choices for voters [2] and stimulate political participation [2, 32, 36, 37]. Thus, it is necessary to consider when polarization moves from being socially healthy to socially dangerous. To help assess the possible dangers of polarization, we now consider the relationship between polarization and social capital.

### 2.3 Polarization and Social Capital

Polarization can potentially be dangerous because it can influence social capital in socially unhealthy ways. Social capital consists of networks where those interacting share norms of reciprocity and trust that facilitate cooperation [38]. Researchers have identified two distinct forms of social capital: bonding capital and bridging capital [38, 39]. Bonding capital is the trust of specific others, such as friends, co-workers, and neighbors. Conversely, bridging capital extends beyond one's immediate social circles to others with whom one has no direct ties or few personal connections. Bonding capital is personal, requires intimate contacts, and holds people together in groups. It is the "social glue" that keeps a community together, and it is similar to what Granovetter calls "strong ties." Conversely, bridging capital connects people across diverse groups and is similar to Granovetter's "weak ties" [40, 41]. In addition, civic engagement is a fundamental dimension of social capital as it is how communities accomplish important goals [39].

The language of social capital theory can help explain polarization and its potential threats to society. In short, polarization occurs when groups with high levels of bonding capital are not connected to each other through bridging capital. A tightly bonded yet unbridged group tends to become increasingly insular and disconnected from other groups. As this occurs, it becomes more homogeneous, and homogeneity in interactions tend to sustain consistency and definiteness in one's orientation and worldview [42–45]. Moreover, as relations with other groups become less frequent, inter-group differences and feelings of distrust increase. In short, research from a variety of settings shows that

involvement in like-minded groups amplifies the ideological tendencies of the group, diminishes opinion diversity, and creates greater distance between ideological opponents. Conversely, as Allport's [46] contact theory suggests when multiple groups who hold diverse worldviews and have access to varied forms of information are connected, distinctions between the groups become increasingly blurred. That is, the intra-group consistency and definiteness of values decreases while inter-group differences decrease [43, 44, 46, 47]. Thus, polarization occurs when strongly bonded groups lack bridging capital; it lessens when clustered networks are linked through bridging capital or "weak ties." Moreover, as a process, as interactions become less frequent between groups and intra-group homogeneity increases, the mistrust that typically develops between groups can destroy any bridging capital that did exist and limit the creation of new bridges that could link the groups. Thus, the process can create a feedback loop that leads to hyper-polarization.

The potential dangers associated with groups being highly bonded yet unconnected to others in the dominant social order is evident in literature on gangs, organized crime syndicates, terrorist organizations, and studies of genocide [48–50]. Not only do such conditions lead to "othering" and "lethal partisanship" known to be correlated with violence towards one's rivals, but these conditions also reduce the ability to find mutually acceptable resolutions to conflicts [51]. Thus, polarization becomes dangerous not when groups disagree; it is dangerous when groups disagree but do not interact!

Media consumption can affect both bonding and bridging capital. While non-partisan media would likely promote bridging capital among groups by creating common symbols that link them, media that celebrates in-group similarities while simultaneously highlighting out-group differences would likely promote high levels of bonding capital within various groups while decreasing bridging capital among them. Given that highly bonded, clustered networks that lack bridging capital are difficult to integrate into larger social discussions [41, 42], if opposing groups consume different media that present radically different versions of events, they are likely to become increasingly internally bonded but externally disconnected. This combination of high-bonding/low-bridging capital can further fuel polarization. If this polarization process remains unchecked and becomes extreme, it can lead to "lethal partisanship" [5], which can be a highly volatile and dangerous situation [23].

Given the above discussion, we ask: (1) Does exposure to traditional media sources or social media predict political polarization, and (2) Do political polarization and media consumption predict levels of social capital? We now turn to our analyses.

### 3 Methods

We begin by using a binomial logistic regression analysis to predict if respondents are political polarized liberally or conservatively. We then investigate if polarization and media consumption predicts levels of bonding and bridging capital. We then regress political participation on political polarization and media consumption.

#### 3.1 Sample

Our analyses are conducted on a sample of 1,424 adult residents of the Commonwealth of Virginia, USA. Data were collected using an online survey between March 7 and March 16, 2019. The sample was selected from demographically balanced panels of potential respondents who had previously volunteered to participate in research surveys. *Dynata*, the world's largest first-party data platform, administers the panels, and they recruit potential participants through a number of permission-based techniques, including random digit dialing and banner ads. *Dynata* sent email invitations to a sample of panel members stratified to reflect the adult population of Virginia, and the sample is within the expected margin of error in terms of important demographic characteristics such as race, ethnicity, and gender.

The sample was collected as part of a project funded by the National Science Foundation. This study was designed in part to analyze social media use, and these data are to supplement those data once they are collected. As such, the survey was limited to residents living in the three largest metropolitan statistical areas in Virginia (Greater Washington, Hampton Roads, and Richmond statistical areas). Over 80% of the Commonwealth's population live in these areas; however, this sampling strategy restricts our analysis to those living in larger urbanized areas and therefore is a limitation of our research.

#### 3.2 Measures

**Political Polarization.** We use Pew Research Center's measure of political polarization [9]. Respondents indicate which statement comes the closest to their view on a series of political issues. These items are presented in Table 1. Responses in the left column are considered "conservative," and those in the right column are "liberal." The presentation of these positions to respondents is randomized. To construct the measures of polarization, responses are coded as  $-1$  if the liberal response is selected and  $+1$  if the conservative response is selected, and the ten items are summed. Respondents who answer liberally to seven or more items (e.g. sum  $< -7$ ) are coded as "polarized liberal." Respondents who answer conservatively to seven or more items conservatively (e.g. sum  $> 7$ ) are coded as "polarized conservative."

**Table 1.** PEW political polarization questions.

| Which statement comes the closest to your view?  |    |  |
|--|----|--|
| Government is almost always wasteful and inefficient   | OR | Government often does a better job than people give it credit for                                      |
| Stricter environmental laws and regulations cost too many jobs and hurt the economy                      | OR | Stricter environmental laws and regulations are worth the cost   |
| Homosexuality should be discouraged by society   | OR | Homosexuality should be accepted by society  |
| Government regulation of business usually does more harm than good                                       | OR | Government regulation of business is necessary to protect the public interest                          |
| Poor people today have it easy because they can get government benefits without doing anything in return | OR | Poor people have hard lives because government benefits don't go far enough to help them live decently |
| The government today can't afford to do much more to help the needy                                      | OR | The government should do more to help needy Americans, even if it means going deeper into debt         |
| Blacks who can't get ahead in this country are mostly responsible for their own condition                | OR | Racial discrimination is the main reason why many black people can't get ahead these days              |
| Immigrants today are a burden on our country because they take our jobs, housing and health care         | OR | Immigrants today strengthen our country because of their hard work and talents                         |
| The best way to ensure peace is through military strength  | OR | Good diplomacy is the best way to ensure peace   |

**Media Use.** Respondents were asked from what media sources they got news during the week prior to completing the survey. The sources available for them to select included television, the Internet, the radio, or print, and they could select from none of these to all of these. Depending on the respondent's selected options, she or he was then asked to specify specific sources within the general media type. Thus, those who indicated they received news from the television were asked to specify from which station or stations they received news, those who specified the radio were asked to identify the type of radio broadcast, etc.

The television stations that could be selected included the major television news stations available in the U.S. (FOX, CNN, ABC, NBC, CBS, PBS, BBC), the respondent's local news station, and the possibility of some other television station. Each of these were coded as 0 (did not receive news from this station) or 1 (received news from this station). A similar strategy was used for radio, print, and Internet sources. The possible choices for radio included a talk show, a news show, or a radio station that is mostly devoted to music or sports but includes news segments. Print choices included the respondent's local newspaper, the New York Times, the Washington Post, USA Today, the Wall Street Journal, the National Inquirer, and a news weekly such as Time or Newsweek. Internet sources included a news aggregator such as Google News or Smart News; news websites

such as MSN, CNN, or Fox; Facebook; Twitter; YouTube or a similar service; Reddit; or some other Internet source.

**Social Capital.** We include measures of bridging and bonding capital, and we operate in the tradition of measuring social capital as various forms of trust [52]. *Bridging capital* is measured with a single item. Respondents were asked, “Generally speaking, would you say that most people can be trusted or that you can’t be too careful in dealing with people,” and the response “you can’t be too careful” was coded as 0 and the response “people can be trusted” was coded as 1. *Bonding capital* was measured with an index of four, four-point Likert items that asked about the level of trust respondents had for people in their neighborhood, the police in their neighborhood, people work in the stores where you shop, and people who are of a similar race to you. Responses ranged from (1) trust not at all to (4) trust a lot. The index had a Cronbach’s alpha of .808. *Civic engagement* was measured using an indicator of eight items asking if respondents did any of the following: worked on a community project; attended a public meeting to discuss town affairs; attended a political meeting or rally; participated in a political group; participated in demonstrations, boycotts, or marches; participated in a charity; participated in a religious group; and participated in a club or organization unrelated to their work. Responses for each item were coded as 0 or 1. The index had a Cronbach’s alpha of .775.

**Control Variables.** We also control for a number of factors known to be correlated with political involvement, social capital, and media use. Age was measured continuously. Income was measured as an ordinal variable with five categories (under \$30,000; between \$30,000 and \$50,000; between \$50,000 and \$75,000; between \$75,000 and \$100,000; and over \$100,000). Education was measured with the question “what was the highest grade of school or year of college you have completed,” and responses ranged from less than high school to a graduate or professional degree. Religiosity was measured using a 5-point Likert scale asking how frequently the respondent attended religious services (1 = never, 5 = every week or more). Respondents were also asked, “How much do you enjoy talking about government and politics with friends and family.” This item was a 4-point Likert scale ranging from 1 “not at all” to 4 “a lot.” We also included indicator variables for sex (male = 1), race (white = 1), and citizenship status (citizen = 1).

## 4 Results

Univariate statistics for all variables used in the analysis are reported in the appendix. Due to space considerations, traditional media sources that were unrelated to any form of polarization or social capital are excluded from the analyses, but we retain all of the measured social media sources because of theoretic interests. Comparisons of the more parsimonious models and the full models confirm that eliminating these variables from the model did not result in any substantive changes to the results. As such, only the trimmed models are presented. Complete results of all analyses that include all media sources are available from the corresponding author.

#### 4.1 Polarization

We begin with two logistic regression models predicting if respondents are polarized liberally or conservatively. We are interested in the relationship between the use of various forms of media to receive news and polarization, and we investigate these relationships while controlling for factors known to be related to political engagement because those interested in politics are more likely to be politically polarized [2]. The results of these models are presented in Table 2.

**Table 2.** Logistic regression of liberal and conservative polarization on media use

|                           | Polarized liberally |             |             | Polarized conservatively |             |              |
|---------------------------|---------------------|-------------|-------------|--------------------------|-------------|--------------|
|                           | B                   | S. E.       | Odds ratio  | B                        | S.E.        | Odds ratio   |
| News from CNN             | <b>.453</b>         | <b>.161</b> | <b>1.57</b> | <b>-1.339</b>            | <b>.335</b> | <b>0.26</b>  |
| News from FOX             | <b>-1.61</b>        | <b>.209</b> | <b>0.20</b> | <b>1.414</b>             | <b>.221</b> | <b>4.11</b>  |
| News from BBC             | .171                | .246        | 1.19        | -.109                    | .400        | 0.89         |
| Local TV news             | -.069               | .176        | 0.93        | -.181                    | .268        | 0.83         |
| Radio talk show news      | -.337               | .228        | 0.71        | <b>1.046</b>             | <b>.251</b> | <b>2.85</b>  |
| Local paper news          | -.176               | .175        | 0.83        | .046                     | .253        | 1.05         |
| New York Times            | <b>.671</b>         | <b>.242</b> | <b>1.96</b> | -.686                    | .661        | 0.50         |
| Washington Post           | <b>.821</b>         | <b>.177</b> | <b>2.27</b> | <b>-1.033</b>            | <b>.359</b> | <b>0.36</b>  |
| A news aggregator         | <b>.354</b>         | <b>.155</b> | <b>1.43</b> | .017                     | .249        | 1.02         |
| Netnews (e.g. MSN)        | .106                | .146        | 1.11        | -.160                    | .232        | 0.85         |
| News from Facebook        | <b>-.544</b>        | <b>.188</b> | <b>0.58</b> | -.040                    | .279        | 0.96         |
| News from Youtube         | <b>-.490</b>        | <b>.280</b> | <b>0.61</b> | -.121                    | .524        | 0.89         |
| News from Twitter         | -.010               | .259        | 0.99        | -.259                    | .457        | 0.77         |
| News from Reddit          | <b>.899</b>         | <b>.318</b> | <b>2.46</b> | -1.467                   | 1.091       | 0.23         |
| White                     | -.194               | .168        | 0.82        | <b>1.057</b>             | <b>.337</b> | <b>2.88</b>  |
| U.S. Citizen              | .132                | .438        | 1.14        | 1.27                     | 1.13        | 3.56         |
| Male                      | <b>-.614</b>        | <b>.146</b> | <b>0.54</b> | <b>.604</b>              | <b>.229</b> | <b>1.83</b>  |
| Religiosity               | <b>-.082</b>        | <b>.038</b> | <b>0.92</b> | <b>.269</b>              | <b>.057</b> | <b>1.31</b>  |
| Education                 | <b>.213</b>         | <b>.056</b> | <b>1.24</b> | <b>-.179</b>             | <b>.086</b> | <b>0.84</b>  |
| Enjoy politics            | <b>.279</b>         | <b>.077</b> | <b>1.32</b> | <b>.566</b>              | <b>.123</b> | <b>1.76</b>  |
| Income (in \$10,000 s)    | -.002               | .021        | 1.00        | <b>.089</b>              | <b>.033</b> | <b>1.09</b>  |
| Age                       | .001                | .005        | 1.00        | .008                     | .007        | 1.01         |
| Constant                  | <b>-2.32</b>        | <b>.527</b> | <b>0.10</b> | <b>-7.65</b>             | <b>1.29</b> | <b>.0004</b> |
| -2 Log Likelihood         | 1365.9              |             |             | 639.2                    |             |              |
| Nagelkerke R <sup>2</sup> | .264                |             |             | .319                     |             |              |

Bolded =  $p < .05$

As seen in Table 2 for the liberally polarized model, those who get their news from CNN are 57% more likely to be liberally polarized (odds ratio = 1.57), while those who watch FOX news are five times less likely to be liberally polarized (OR = 0.20). There are also strong relationships between getting one's news from the New York Times (OR = 1.96) and the Washington Post (OR = 2.27) and being liberally polarized. In terms of social media sources, those who are liberally polarized are more likely to get their news from a news aggregator (OR = 1.43) and Reddit (OR = 2.46), but they are less likely to get news from Facebook (OR = .58) or YouTube (OR = .61). A number of other variables are related to being liberally polarized as expected. For example, males and the more religious are less likely to be liberally polarized. Conversely, education and liking politics are positively related to liberal polarization. Race, age, income and citizenship status were not particularly strong predictors of liberal polarization.

Looking at the model for polarized conservatives, it is clear that they tend to get news from FOX (OR = 4.11) and radio talk shows (OR = 2.85), and they are very unlikely to get their news from CNN (OR = 0.26) or the Washington Post (OR = 0.36). None of the social media sources are good predictors of being conservatively polarized, although all of the effects except for a news aggregator (which is virtually no effect) indicate polarized conservatives do not use these as news sources. For example, the odds ratio for Reddit, while not statistically significant by conventional standards, is nevertheless noteworthy (OR = .23). The other variables in the model are related to conservative polarization as expected with Whites, males, the religious, those with higher incomes, and those who enjoy politics being more likely to be polarized in a conservative manner. U.S. citizens are also more likely to be polarized conservatively, although the effect is not significant by conventional standards, the odds ratio of 3.56 is noteworthy. One likely reason that the effect did not achieve statistical significance is because approximately 96% of the sample were U.S. citizens. Education is inversely related to conservative polarization.

## 4.2 Social Capital

Turning to the question of whether media consumption and polarization influence social capital, we first conduct a logistic regression on generalized trust, which is our measure of bridging capital. We then conduct two ordinary least squares regressions. The first investigates the relationship between bonding capital, polarization, and media use. The second model regresses civic engagement on our polarization, media, and control variables. The results are presented in Table 3.

We see in Table 3 that those who are polarized conservatively tend to express lower levels of bridging capital (OR = 0.63) while those who are polarized liberally report higher levels of bridging capital (OR = 1.25), although the latter effect is not statistically significant using conventional standards. In terms of media effects, the only traditional media source that is a significant predictor of bridging capital is receiving news from a local T.V. station (OR = 0.67). Social media sources, however, do appear to be good predictors of bridging capital. Those who receive news from a news aggregator are 39% more likely to express high levels of bridging capital (OR = 1.39), but those who receive news from Facebook (OR = 0.74) and Reddit (OR = 0.31) are more likely to report low

levels of bridging capital. Whites, males, the religious, those with more education, and those who enjoy discussing politics all express higher levels of bridging capital than do their counterparts.

**Table 3.** Logistic regression of bridging capital on polarization and media use

|                           | Bridging capital |             |             |
|---------------------------|------------------|-------------|-------------|
|                           | B                | S. E.       | Odds ratio  |
| Polarized Liberal         | .219             | .143        | 1.25        |
| Polarized Conservative    | <b>-.460</b>     | <b>.217</b> | <b>0.63</b> |
| News from CNN             | .172             | .140        | 1.19        |
| News from FOX             | -.063            | .137        | 0.94        |
| BBC News                  | -.183            | .207        | 0.83        |
| Local TV news             | <b>-.408</b>     | <b>.147</b> | <b>0.67</b> |
| Radio talk show news      | -.218            | .177        | 0.80        |
| Local paper news          | .099             | .143        | 1.10        |
| New York Times            | .095             | .230        | 1.10        |
| Washington Post           | .203             | .160        | 1.23        |
| A news aggregator         | <b>.330</b>      | <b>.135</b> | <b>1.39</b> |
| Netnews (e.g. MSN)        | -.167            | .124        | 0.85        |
| News from Facebook        | <b>-.295</b>     | <b>.151</b> | <b>0.74</b> |
| News from Youtube         | .306             | .227        | 1.36        |
| News from Twitter         | -.174            | .226        | 0.84        |
| News from Reddit          | <b>-1.177</b>    | <b>.352</b> | <b>0.31</b> |
| White                     | <b>.698</b>      | <b>.149</b> | <b>2.01</b> |
| U.S. Citizen              | -.447            | .359        | 0.64        |
| Male                      | <b>.274</b>      | <b>.121</b> | <b>1.32</b> |
| Religiosity               | <b>.106</b>      | <b>.032</b> | <b>1.11</b> |
| Education                 | <b>.108</b>      | <b>.046</b> | <b>1.12</b> |
| Enjoy politics            | <b>.327</b>      | <b>.065</b> | <b>1.39</b> |
| Income (in \$10,000 s)    | .026             | .017        | 1.03        |
| Age                       | .002             | .004        | 1.00        |
| Constant                  | <b>-2.33</b>     | <b>.440</b> | <b>0.10</b> |
| -2 Log Likelihood         | 1800.5           |             |             |
| Nagelkerke R <sup>2</sup> | .146             |             |             |

Bolded = p < .05



Table 4 reports the OLS regression results for both bonding capital and civic engagement. Being polarized liberally is associated with elevated levels of bonding capital ( $b = .362$ ) while being polarized conservatively is largely unrelated to bonding capital ( $b = -.013$ ). In terms of media effects, getting news from a local paper ( $b = .500$ ) and YouTube ( $b = .502$ ) are positively related to bonding capital, while getting news

**Table 4.** Regression of bonding capital and civic engagement on polarization and media use

|                                    | Bonding capital |             |              |              | Civic engagement                   |             |              |              |
|------------------------------------|-----------------|-------------|--------------|--------------|------------------------------------|-------------|--------------|--------------|
|                                    | B               | S. E.       | T            | Beta         | B                                  | S. E.       | T            | Beta         |
| Polarized left                     | <b>.362</b>     | <b>.143</b> | <b>2.52</b>  | <b>.065</b>  | -.011                              | .131        | -0.09        | -.002        |
| Polarized right                    | -.013           | .217        | -.059        | -.001        | -.173                              | .198        | -0.87        | -.022        |
| CNN                                | .148            | .139        | 1.06         | .027         | -.022                              | .128        | -0.17        | -.004        |
| FOX                                | .135            | .136        | 0.99         | .025         | .028                               | .125        | 0.22         | .006         |
| BBC News                           | -.276           | .209        | -1.32        | -.033        | <b>.484</b>                        | <b>.191</b> | <b>2.54</b>  | <b>.063</b>  |
| Local TV news                      | .100            | .144        | 0.69         | .017         | -.085                              | .132        | -0.65        | -.016        |
| Radio talk show                    | -.100           | .176        | -0.57        | -.014        | .239                               | .160        | 1.49         | .036         |
| Local paper                        | <b>.500</b>     | <b>.145</b> | <b>3.45</b>  | <b>.089</b>  | <b>.349</b>                        | <b>.132</b> | <b>2.63</b>  | <b>.068</b>  |
| New York Times                     | -.152           | .233        | -0.65        | -.017        | .327                               | .212        | 1.53         | .040         |
| Washington Post                    | .109            | .164        | 0.69         | .019         | <b>.484</b>                        | <b>.150</b> | <b>3.22</b>  | <b>.090</b>  |
| News aggregator                    | .185            | .135        | 1.37         | .034         | <b>.243</b>                        | <b>.123</b> | <b>1.98</b>  | <b>.049</b>  |
| Net News                           | -.026           | .124        | -0.23        | -.005        | -.132                              | .113        | -1.17        | -.029        |
| Facebook                           | .145            | .146        | 0.99         | .024         | .218                               | .133        | 1.63         | .040         |
| Youtube                            | <b>.502</b>     | <b>.221</b> | <b>2.27</b>  | <b>.057</b>  | .294                               | .203        | 1.45         | .037         |
| Twitter                            | .130            | .215        | 0.61         | .015         | .238                               | .197        | 1.21         | .030         |
| Reddit                             | <b>-.610</b>    | <b>.296</b> | <b>-2.06</b> | <b>-.052</b> | -.108                              | .267        | -0.40        | -.010        |
| White                              | <b>1.213</b>    | <b>.142</b> | <b>8.53</b>  | <b>.225</b>  | -.034                              | .130        | -0.26        | -.007        |
| U.S. Citizen                       | -.632           | .360        | -1.76        | -.042        | -.144                              | .325        | -0.44        | -.011        |
| Male                               | <b>-.256</b>    | <b>.120</b> | <b>-2.14</b> | <b>-.053</b> | .145                               | .109        | 1.32         | .033         |
| Religiosity                        | <b>.164</b>     | <b>.031</b> | <b>5.26</b>  | <b>.128</b>  | <b>.388</b>                        | <b>.029</b> | <b>13.58</b> | <b>.333</b>  |
| Education                          | .078            | .046        | 1.68         | .046         | <b>.097</b>                        | <b>.042</b> | <b>2.30</b>  | <b>.064</b>  |
| Enjoy politics                     | <b>.254</b>     | <b>.064</b> | <b>3.97</b>  | <b>.098</b>  | <b>.459</b>                        | <b>.059</b> | <b>7.83</b>  | <b>.194</b>  |
| Income                             | <b>.054</b>     | <b>.017</b> | <b>3.14</b>  | <b>.086</b>  | -.001                              | .016        | -0.09        | -.002        |
| Age                                | <b>.029</b>     | <b>.004</b> | <b>7.14</b>  | <b>.213</b>  | <b>-.022</b>                       | <b>.004</b> | <b>-5.83</b> | <b>-.175</b> |
| Constant                           | <b>8.629</b>    | <b>.433</b> | <b>19.93</b> | ...          | <b>2.36</b>                        | <b>.393</b> | <b>6.01</b>  | ...          |
| F = 17.53<br>R <sup>2</sup> = .227 |                 |             |              |              | F = 18.52<br>R <sup>2</sup> = .241 |             |              |              |

Bolded =  $p < .05$

from Reddit ( $b = -.610$ ) is inversely related to bonding capital. Whites, the religious, the more educated, those who are older and with higher incomes, and those who enjoy discussing politics all express higher levels of bonding capital. Males and U.S. citizens express lower levels of bonding capital.

In terms of civic engagement, polarization appears to be unrelated to civic engagement. However, getting the news from the BBC ( $b = .484$ ), the local newspaper ( $b = .349$ ), the Washington Post ( $b = .484$ ), and a news aggregator ( $b = .243$ ) are all positively related to civic engagement. Religiosity, education, enjoying politics are also positively related to civic engagement, while age is inversely related to civic engagement. Although no other variable achieves statistical significance using traditional values, it is notable that the standardized coefficients for the New York Times, Facebook, and YouTube (Betas = .04, .04, and .037, respectively) suggest these have a similar relationship with civic engagement as does news aggregator ( $\beta = .049$ ).

## 5 Discussion

Numerous commentators have noted how political polarization is increasing in the United States and other Western democracies. Polarization, so it is argued, rips at the fabric of society and leads to a culture war, and hyper-polarization can potentially lead to social unrest and physical violence. Popular wisdom, mainstream media, and numerous scholars argue that social networking sites such as Facebook and Twitter are driving polarization, as these allow partisan opinions and “fake news” to spread rapidly and incite intensely polarized cultural debates. Similarly, the rise of online “echo chambers” are also considered mechanisms by which social media contributes to polarization. Yet, others find political bias in more traditional forms of media are problematic and may even have a greater effect than social media on polarization. We aimed to investigate these claims and add to the discussion by further investigating the relationship between polarization, media use, and social capital.

Our results revealed stark differences between liberally polarized and conservatively polarized respondents in terms of their news consumption. Polarized conservatives got their news from radio talk shows and FOX News while avoiding newspapers and television news perceived to be liberal (e.g. CNN). Interestingly, using social media for news was unrelated to being polarized in a conservative direction. In contrast, while polarized liberals were likely to get their news from newspapers and television (watching CNN and avoiding FOX News), there were also significant social media effects. Specifically, polarized liberals got news from Reddit and news aggregators and avoided getting news from Facebook and YouTube or similar SNS sites.

Finding these media effects on polarization, we then investigated if polarization influences various forms of social capital. We first analyzed the relationship between polarization and generalized trust, which is often used as a measure of “bridging capital.” Here we found that polarized conservatives are less likely to trust people in general while there was no relationship between being a polarized liberal and trusting others. We also found that getting news from one’s local television station, Facebook, and Reddit was associated with lower levels of expressed trust in others; however, receiving news from a news aggregator was positively related to our measure of bridging capital. We also

considered the relationship between polarization and trust among intimates or “bonding capital.” In this case, polarized liberals were more likely to express high levels of bonding capital while being a polarized conservative was unrelated to bonding capital. Receiving news from one’s local paper or Youtube was positively related to bonding capital, while getting news from Reddit was inversely related to bonding capital. Finally, we also explored if polarization predicts civic engagement. Polarization appears to be unrelated to civic engagement, but several media sources—including watching BBC, reading the local paper or the Washington Post, and using a news aggregator—were positively related to civic engagement.

Consequently, our first contribution to the literature concerns how various forms of media may be contributing to polarization differently. Polarized conservatives appear to get their news from traditional sources, and while polarized liberals also use traditional news sources, they apparently supplement these sources with newer forms of news media. Thus, as for fear of social media creating a “filter bubble” or “echo-chamber” where an individual’s ideas are echoed back to them by a personalized Internet experience limited to those who share his or her worldviews, it appears that this is more likely to be happening among polarized liberals than polarized conservatives. Conversely, conservatives appear to be living in a “FOX news bubble” and “conservative talk radio bubble” while actively avoiding other forms of media often accused of being overly liberal. Thus, the digital divide extends to media consumption: liberals are embracing new sources of media while conservatives cling to traditional forms of media. Therefore, at least among the political extremes, the divide between polarized conservatives and liberals results in opposing groups who are informed by different sources. This lack of a shared information results in competing worldviews while providing little opportunity for finding common ground.

We also find that polarization varies in its relationship with various forms of social capital. While it may increase bonding capital for liberals, it may decrease bridging capital for conservatives. We consider this finding particularly disturbing, as it appears polarization may influence social capital in dangerous ways. While polarized liberals appear to be becoming “tribal” in the sense that they express high levels of bonding capital, polarized conservatives appear to be distrustful of those they do not know well. While polarized “tribes” who are unconnected are problematic in that there are limited avenues for pursuing dialogue among them, this becomes especially problematic when we consider this in combination with our findings that these groups are receiving information from very different sources. Given research that suggests growing media bias even in “mainstream” news [e.g. 33], it becomes increasingly likely that any dialogue that occurs between these opposing groups is going to begin with a different set of “facts”; and, if we cannot agree on what the facts are, it is unlikely we will ever agree on how facts should be interpreted. This combination of high bonding, low bridging capital coupled with the lack of a common source of news can explain the recent increase in lethal partisanship where groups not only disagree but also accept or even wish harm to their political opponents [5].

With this said, we want to warn against becoming overly alarmist. First, we should note that fully two-thirds of our sample *were not politically polarized*, and the non-polarized are receiving news from a variety of traditional and newer forms of media. Next, there is evidence that even among those who are polarized, people still interact

with those who do not share their political views. Respondents were asked if most of their close friends shared their views on government and politics, and while 54% of those who were polarized said they did, 46% had close friends with differing views. More encouragingly, among those who were not polarized, 67% had friends with differing views. Thus, at least for most people, we do have access to differing political perspectives, and these perspectives are shared through traditional media and social media. Consequently, while there may be cause for concern, those claiming we are so polarized that compromise is impossible are likely overstating their case. Nevertheless, civil discourse would undoubtedly benefit from more dialogue across the political divide.

## 6 Conclusion

While claims that Western democracies are too polarized to function adequately are likely overly alarmist, there is evidence of a growing gulf between those who view the world through a political liberal lens and those who view it more conservatively. The media may be contributing to this growing divide, but we cannot blame only new forms of media. Although the personalized online experience may result in echo chambers, traditional forms of media appear to be leading to bubbles for some on the right. Thus, finding a way to bridge the information divide will be challenging, as humans have always tended to sort themselves along political lines. Even those who warn against filter bubbles and echo chambers [e.g. 22, 23] recognize this. Yet, it would behoove anyone interested in preserving democracy to continue paying attention to polarization and the influence media has on it because we find evidence that it is related to high levels of bonding capital but low bridging capital, which history has shown can be extremely dangerous. Finding ways to make social media achieve its promise of promoting the free-flow of information to stimulate discussions across opposing political camps should continue to be a goal of service providers, politicians, researchers, and the public at large. Our democracies can only benefit from doing so.

## Appendix: Univariate Statistics

| N = 1424                | Minimum | Maximum | Mean | Standard deviation |
|-------------------------|---------|---------|------|--------------------|
| Polarized liberal       | 0       | 1       | 0.25 | 0.44               |
| Polarized conservative  | 0       | 1       | 0.09 | 0.28               |
| Bridging capital        | 0       | 1       | 0.40 | 0.49               |
| Bonding capital         | 4       | 16      | 12.4 | 2.42               |
| Civic engagement        | 0       | 8       | 3.8  | 2.21               |
| News from CNN           | 0       | 1       | 0.28 | 0.45               |
| News from FOX           | 0       | 1       | 0.28 | 0.45               |
| News from BBC           | 0       | 1       | 0.09 | 0.29               |
| News from local station | 0       | 1       | 0.22 | 0.42               |

(continued)

*(continued)*

| N = 1424                  | Minimum | Maximum | Mean   | Standard deviation |
|---------------------------|---------|---------|--------|--------------------|
| News from radio talk show | 0       | 1       | 0.13   | 0.33               |
| News from local paper     | 0       | 1       | 0.25   | 0.43               |
| News from New York Times  | 0       | 1       | 0.08   | 0.27               |
| News from Washington Post | 0       | 1       | 0.21   | 0.41               |
| News from news aggregator | 0       | 1       | 0.26   | 0.44               |
| News from net (e.g. MSN)  | 0       | 1       | 0.35   | 0.48               |
| News from Facebook        | 0       | 1       | 0.21   | 0.41               |
| News from YouTube         | 0       | 1       | 0.08   | 0.28               |
| News from Twitter         | 0       | 1       | 0.09   | 0.28               |
| News from Reddit          | 0       | 1       | 0.04   | 0.23               |
| White                     | 0       | 1       | 0.72   | 0.45               |
| U.S. Citizen              | 0       | 1       | 0.96   | 0.16               |
| Male                      | 0       | 1       | 0.48   | 0.50               |
| Religiosity               | 0       | 5       | 2.2    | 1.89               |
| Education                 | 1       | 6       | 4.1    | 1.45               |
| Enjoy discussing politics | 1       | 4       | 2.4    | 0.94               |
| Income                    | 20,000  | 125,000 | 83,805 | 38,484             |
| Age                       | 18      | 90      | 52.02  | 17.74              |

## References




1. Smeltz, D., Busby, J., Tama, J.: National Security Network of Foreign Policy Opinion Leaders. Chicago Council on Global Affairs, Chicago, IL (2018)
2. McCoy, J., Rahman, T., Somer, M.: Polarization and the global crisis of democracy: common patterns, dynamics, and pernicious consequences for democratic polities. *Am. Behav. Sci.* **62**(1), 16–42 (2018)
3. Jones, D.R.: Party polarization and legislative gridlock. *Polit. Res. Q.* **54**(1), 125–141 (2001)
4. Thurber, J.A., Yoshinaka, A.: American Gridlock: The Sources, Character, and Impact of Political Polarization. Cambridge University Press, Cambridge (2015)
5. Kalmoe, N.P., Mason, L.: Lethal mass partisanship: prevalence, correlates, and electoral contingencies. In: American Political Science Association Conference, Washington, D.C., pp. 1–41 (2018)
6. Baldassarri, D., Gelman, A.: Partisans without constraint: political polarization and trends in American public opinion. *Am. J. Sociol.* **114**(2), 408–446 (2008)
7. DiMaggio, P., Evans, J., Bryson, B.: Have American's social attitudes become more polarized? *Am. J. Sociol.* **102**(3), 690–755 (1996)
8. Iyengar, S., Westwood, S.J.: Fear and loathing across party lines: new evidence on group polarization. *Am. J. Polit. Sci.* **59**(3), 690–707 (2015)

9. PEW 2017 The Partisan Divide on Political Values Grows Even Wider: Sharp shifts among Democrats on aid to needy, race, immigration. <https://www.people-press.org/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/>. Accessed 03 Jan 2020
10. Applebaum, A.: Warning from Europe: The Worst is yet to come (polarization, conspiracy theories, attacks on free press). <https://www.theatlantic.com/magazine/archive/2018/10/pol-and-polarization/568324/>. Accessed 02 Jan 2020
11. Pisani-Ferry, J. Responding to Europe's Political Polarization France Stratégie. <https://www.strategie.gouv.fr/english-articles/responding-europes-political-polarization>. Accessed 02 Jan 2020
12. Müller, S., Schnabl, G.: The European Central Bank Drives the Political Polarization in Europe. Thinkmarkets. <https://thinkmarkets.wordpress.com/2017/11/09/the-european-central-bank-drives-the-political-polarization-in-europe/>. Accessed 02 Jan 2020
13. Bonikowski, B.: Three lessons of contemporary populism in Europe and the United States. *Brown J. World Aff.* **23**(1), 9–24 (2016)
14. Bulut, E., Yörük, E.: Mediatized populisms|Digital populism: trolls and political polarization of Twitter in Turkey. *Int. J. Commun.* **11**(1), 4093–4117 (2017)
15. Vachudova, M.A.: From competition to polarization in central Europe: how populists change party systems and the European Union. *Polity* **51**(4), 689–706 (2019)
16. Dalton, R.J., Tanaka, A.: The patterns of party polarization in East Asia. *J. East Asian Stud.* **7**(2), 203–223 (2007)
17. Populism and polarisation threaten Latin America After dictatorships gave way to democracy trouble is brewing again. *The Economist*. <https://www.economist.com/briefing/2019/05/09/populism-and-polarisation-threaten-latin-america>. Accessed 30 Dec 2019
18. Sunstein, C.R.: *Going to Extremes: How like Minds Unite and Divide*. Oxford University Press, New York (2009)
19. Bail, C.A., et al.: Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci.* **115**(37), 9216–9221 (2018)
20. Hong, S., Kim, S.H.: Political polarization on twitter: implications for the use of social media in digital governments. *Gov. Inf. Q.* **33**(4), 777–782 (2016)
21. Conover, M.D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., Flammini, A.: Political polarization on twitter. In: *Fifth international AAAI Conference on Weblogs and Social Media*, pp. 83–96 (2011)
22. Pariser, E.: *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. Penguin, London (2011)
23. Hawdon, J.: Applying differential association theory to online hate groups: a theoretical statement. *Res. Finnish Soc.* **5**(1), 39–47 (2012)
24. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. *Ann. Rev. Sociol.* **27**, 415–444 (2001)
25. Vaisey, S., Lizardo, O.: Can cultural worldviews influence network composition? *Soc. Forces* **88**(4), 1595–1618 (2010)
26. Adamic, L.A., Glance, N.: The political blogosphere and the 2004 U.S. election: divided they blog. In: *Proceedings of the 3rd International Workshop on Link Discovery*, pp. 36–43. ACM, New York (2005)
27. Gruzd, A., Roy, J.: Investigating political polarization on twitter: a Canadian perspective. *Policy Internet* **6**(1), 28–45 (2014)
28. Garrett, R.K., Carnahan, D., Lynch, E.K.: A turn toward avoidance? Selective exposure to online political information, 2004–2008. *Polit. Behav.* **35**(1), 113–134 (2011)
29. Barberá, P.: *How social media reduces mass political polarization. Evidence from Germany, Spain, and the US*. Job Market Paper, New York University, N.Y. (2014)

30. Boxell, L., Gentzkow, M., Shapiro, J.M.: Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proc. Natl. Acad. Sci.* **114**(40), 10612–10617 (2017)
31. Iyengar, S., Hahn, K.S.: Red media, blue media: evidence of ideological selectivity in media use. *J. Commun.* **59**(1), 19–39 (2009)
32. Morris, J.S.: Slanted objectivity? Perceived media bias, cable news exposure, and political attitudes. *Soc. Sci. Q.* **88**(3), 707–728 (2007)
33. Cassino, D.: *Fox News and American Politics: How One Channel Shapes American Politics and Society*. Routledge, New York (2016)
34. Baum, M.A., Groeling, T.: New media and the polarization of American political discourse. *Polit. Commun.* **25**(4), 345–365 (2008)
35. Prior, M.: Media and political polarization. *Annu. Rev. Polit. Sci.* **16**(1), 101–127 (2013)
36. Changjun, L., Shin, J., Hong, A.: Does social media use really make people politically polarized? Direct and indirect effects of social media use on political polarization in South Korea. *Telematics Inform.* **35**(1), 245–254 (2018)
37. Prior, M.: News vs entertainment: how increasing media choice widens gaps in political knowledge and turnout. *Am. J. Polit. Sci.* **49**(3), 577–592 (2005)
38. Putnam, R.D., Leonardi, R., Nanetti, R.Y.: *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton University Press, Princeton (1994)
39. Putnam, R.D.: *Bowling Alone: The Collapse and Revival of American Community*. Simon and Schuster, New York (2001)
40. Granovetter, M.S.: The strength of weak ties. *Am. J. Sociol.* **78**(6), 1360–1380 (1973)
41. Granovetter, M.: The strength of weak ties: a network theory revisited. *Sociol. Theory* **1**(1), 201–233 (1983)
42. Hawdon, J.: Cycles of deviance: structural change, moral boundaries, and drug use, 1880–1990. *Sociol. Spectr.* **16**(1), 183–207 (1996)
43. Pettigrew, T.F., Tropp, L.R.: A meta-analytic test of intergroup contact theory. *J. Pers. Soc. Psychol.* **90**(5), 751–783 (2006)
44. Pettigrew, T.F., Tropp, L.R., Wagner, U., Christ, O.: Recent advances in intergroup contact theory. *Int. J. Intercultural Relat.* **35**(3), 271–280 (2011)
45. Schmid, K., Ramiah, A.A., Hewstone, M.: Neighborhood ethnic diversity and trust: the role of intergroup contact and perceived threat. *Psychol. Sci.* **25**(3), 665–674 (2014)
46. Allport, G.W.: *The Nature of Prejudice*. Addison-Wesley, Reading (1954)
47. Kadushin, C.: *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford University Press, New York (2012)
48. Hawdon, J., Ryan, J., Lucht, M.: *The Causes and Consequences of Group Violence: From Bullies to Terrorists*. Lexington Press, Lanham (2014)
49. Deuchar, R.: People look at us, the way we dress, and they think we're gangsters': bonds, bridges, gangs and refugees: a qualitative study of inter-cultural social capital in Glasgow. *J. Refugee Stud.* **24**(4), 672–689 (2011)
50. Helfstein, S.: Social capital and terrorism. *Defence Peace Econ.* **25**(4), 363–380 (2014)
51. Quigley, P., Hawdon, J.: *Reconciliation After Civil Wars: Global Perspectives*. Routledge, New York (2018)
52. van Staveren, I., Knorringa, P.: Unpacking social capital in economic development: how social relations matter. *Rev. Soc. Econ.* **65**(1), 107–135 (2007)



# Designing an Experiment on Recognition of Political Fake News by Social Media Users: Factors of Dropout

Olessia Koltsova , Yadviga Sinyavskaya  , and Maxim Terpilovskii

Laboratory for Social and Cognitive Informatics, National Research University Higher School of Economics, Saint-Petersburg, Russia

ysinyavskaya@hse.ru

**Abstract.** Although social networking sites (SNS) offer functionalities for large-scale online research, user behavior and, in particular, scale and factors of their dropout from SNS-administered research have hardly been studied. In this paper we present an SNS-based experiment and survey tool and report the results of our investigation of user dropout from a research that uses this tool. This research is a pilot stage of a cross-country comparative study of political fake news recognition. At this stage Facebook and Vkontakte users from Russia have been recruited via SNS ad managing systems, asked to evaluate the truthfulness of the displayed news items and to answer a number of questions. We find that although we had to perform thousands of ad displays, among those who clicked the ad dropout rate was 60 and 65% in Vkontakte and Facebook respectively. 1,816 complete questionnaires were collected within a few days. More educated respondents, people living in or near megalopolises and those who agreed to grant access to their Vkontakte account data were significantly more inclined to complete the survey, but the major predictor of dropout was high individual speed – an indicator of low interest. Neither device type (mobile vs desktop) nor the number of questions per screen (one vs two) affected dropout. The number of leavers declined from the first to the last screens of our tool, but transition from the experiment to the survey and demographic questions produced clear peaks in the dropout curve.

**Keywords:** Online experiment · Online survey · Factors of dropout · Social media

## 1 Introduction

Proliferating social networking sites (SNS) open great opportunities for academic surveys and experiments in social sciences. SNS ad managing services allow for control of sample parameters, while the advertised data collection applications may employ a variety of experimental designs and gamification techniques. Thus, web-based research allows combining the collection of experimental data, self-reported data, and observational data from user accounts. However, they are also not free from the typical problems faced by researchers in “paper-and-pencil” surveys and laboratory experiments.



One of the most important tasks in online research (similar to offline non-response rate reduction) is dropout prevention.

Dropout is a respondent's premature termination of a survey session. Dropout rate is the share of respondents who started the questionnaire but abandoned it on different stages of completing, among all respondents [1].

The complete (unit nonresponse) or partial (item nonresponse) loss of data due to respondent's refusal to participate seems prominent methodological issue causing the *nonresponse* errors, i.e. respondents being significantly different from nonrespondents [2]. This, in turn, poses a threat for the reliability and generalizability of results gained via Web-based surveys.

The comparison of dropout values in offline and online research reveals the higher rates among the latter: it varies from 15% to 80% depending on the recruiting method in web-based studies in contrast to 5% termination rate in face to face or telephone surveys [3].

Such striking contrast in dropout rates can be explained by the lack of communication between researchers and respondents in online settings and, therefore, limited ability to establish the rapport, provide real time feedback to respondents' actions and increase participants' loyalty by all other means available in a classical laboratory setting.

Two separate strategies (or their combination) are usually implemented by researchers for reducing the dropout in web-based studies. The first one aims at increasing the motivation of respondent to participate by proposing incentives and awards, such as money or new knowledge. The empirical evidence on the effect of monetary rewards on both dropout and data quality is inconclusive [4–7] while the non-monetary motivation, such as *interest* in research topic, seem to lead to lower incidents of respondents leaving the study [8].

The second strategy implies manipulation of the features of the web-based tool so that to increase its usability and convenience for a respondent. Interface structure and design [1, 2, 9, 10], tool's stability across different platforms and devices [11]; questionnaire features, such as its length [3], type of questions [8, 12], their order, including the request of personal information in the beginning [7]; the way the questions are displayed on the screen [3, 5, 13] – all have been reported as substantial factors of dropout in web-based research.

Decision field theory [14] provides the perspective for explaining the survey behavior by combining these two strategies. The decision to continue or to abort a survey session depends on the ratio between the strength of factors increasing or decreasing respondent's motivation.

Research on dropout factors is of special importance for politically sensitive research topics, such as our experiment on fake political news recognition which also offers a number of sensitive political questions. This goal, first, is a competing task with ethical data collection based on informed consent, second, it contradicts the goal of research thoroughness, and third, it can stumble into dropout factors that are dramatically different from what researchers know from their offline experience.

## 1.1 Objectives

In this study, we investigate respondents' behavior in a self-administered online experiment on recognition of political fake news. First, we compare the dynamics of dropout in two large SNSs – Facebook and Vkontakte – most popular in the studied country (Russia). Second, we examine what are some of the factors that may affect users' decision to leave the experiment. The first important group of factors addressed in the study is the properties of respondents, such as socio-demographic features, privacy preferences or perception of personal performance during the experiment. The other group is the effects of web interface, which includes screen type and the number of questions per screen among others. This study is a pilot stage of our future comparative cross-country research on factors of fake news recognition.

## 2 Research Design

In this section we briefly describe the design of our future research on fake news whose instrument we are testing in this paper and the approach to its testing. The goal of the future research is to experimentally find out factors influencing Facebook and Vkontakte users' ability to recognize fake news in a situation with or without international tensions. The research is supposed to be carried in three countries: Russia, Kazakhstan and Ukraine. Both in the future research and in the current pilot research we employ  $2 \times 2 \times 2$  experimental design in which each SNS user receives eight news randomly retrieved from our database with varying truthfulness (true/false), news source (from user's country/from the country covered in the news) and news frame (dominant/alternative). A user then receives two questions on whether he/she has checked any of the news or has seen any of them before, six demographic questions, three questions on news consumption, three questions on generalized trust, four questions on conspiracy thinking, and three on political interest and attitudes (5-point Likert scales). Questions are organized in 14 screens, 1–2 questions per screen. They are followed by the user's score, a half-serious comment on her/his fake-detection abilities and an offer to learn correct answers after the end of the experiment. For Vkontakte users, the instrument is presented as an application that also requests the data from user accounts, namely friend lists and group lists (not required for participation). For Facebook users, it is presented as a stand-alone website that does not collect any account data (due to restrictions of Facebook). Since Vkontakte is much more popular than FB in Russia and in Kazakhstan, but not in Ukraine, one of the goals of our instrument testing is to compare dropout in both social networks. It is also important to understand that VK terms of use allow third parties to download certain user data from accounts not protected by privacy settings without special requests. These data may include gender, age, city, views and interests and any other data from the "about" section that the user chooses not to hide. These data are also used in this research.

Instrument testing began from technical tests by the Lab members who revealed several dozens of bugs and glitches; at this stage we also refused from any screens that involved scrolling since it obviously irritated all participants, and from drop-out menus since they obviously incentivized participants to choose the first option. An earlier

offline pre-test on 100 students that compared convenience of scales and binary questions revealed no significant difference, which is why we opted for scales as more informative options.

In the present research we defined our populations as all FB and VK users from Russia, which is why we randomly targeted these groups in both networks using respective ad management systems. As we estimate our final VK sample size in Russia as 3–5 thousands, we took 400 (approx. 10%) as the target size for our pilot research, and the size of FB sample twice as small, according to its smaller penetration in Russia. After having collected half of both samples we analyzed the results and permuted the order of one of the questions in each of the instrument versions to test two additional hypotheses that emerged after the interim data analysis. All the pilot experiments were carried out in October–December 2019.

### 3 Assumptions and Hypotheses

In this study we test several statistical hypotheses, as well as a series of less rigorously formulated assumptions that are nonetheless important for calibrating our instrument. In formulating them, we based on the literature on survey non-response and dropout, on user churn in gaming and human-computer interaction. We begin with the description of two most general expectations.

First, we expect that *the pattern of dropout in our study will be similar to those detected in previous works on churn dynamic* [3]. The power-law distributed curve with only a small loyal minority of “survivors” is typical for both online game and app users [15, 16] and is expected to be unavoidable. Therefore, fluctuations can be detected correctly only against this general trend.

Second, since the online experiment as a method opens a wide range of opportunities for increasing the response rate by means of interactive stimuli, gamification of a process, presence of graphic elements in the design etc., we suppose that combination of an experimental and a more standardized survey sections in one research may provoke a dropout of the respondents. Thus, *while mixing the entertaining experimental part with the more standard survey-part (like in the present study), the highest number of dropouts will occur at the moment of transition of a respondent from experimental to survey part.*

More specifically, we focused around two lines of inquiry, the first relating to the features of respondents and the second relating to the features of data gathering process. While the second line obviously aims to reveal ways to improve our instrument, the first one is aimed to uncover potential sample biases that should be taken into account during the future research.

#### 3.1 Respondents’ Features as a Factor of Dropout

**Socio-demographic Factors.** Age of a user might have a two-way impact on a person’s behavior during online survey. On the one hand, the interaction with the online survey application requires technical skills which may be less common among so-called “digital immigrants” who encountered new digital technologies being middle-aged [17]. On the other hand, as Galesic [3] has shown, older respondents are more cooperative, with the

probability of dropout decreasing by around 1% with the each additional year of age. Taking into account these two lines, we hypothesize that:

*H1a: Dropout rate will increase with users' age.*

*H1b: Dropout rate will decrease with users' age.*

Along with the age, *gender differences* in survey behavior may occur because of differences in online behavior, such as frequency of SNS use [18], purposes of Internet usage [19] and levels of online self-disclosure [20, 21]. In addition, some scholars nominated women as more cooperative respondents [22] with higher levels of interest in survey process [3]. In consistence with these findings, we expect that:

*H2: Dropout rate will be higher among male than among female respondents.*

Finally, following the argument of Galesic [3] that lower level of education might lead to additional difficulties during surveys, we hypothesize that:

*H3: Dropout rate will be higher among respondents with lower education.*

**Individual Performance as a Factor of Dropout.** Respondent's perception of her own performance in an experimental task might be a crucial factor of attrition as it has been shown to happen in case of online learning [23, 24] and in online game contexts [16, 25]. Thus, we suppose that high individual speed of performance in the experiment may serve as a proxy for the low level of difficulty and, consequently, for high level of satisfaction that should incentivize a respondent to move further. At the same time, high speed might also indicate low quality of answers that has been found to be related to the lack of interest, while the latter, in turn, has been found to increase dropout [3]. Therefore:

*H4a: The higher the personal speed the lower the probability of dropout.*

*H4b: The lower the personal speed the lower the probability of dropout.*

**Privacy and Online Activity as a Factor of Dropout.** As functional complexity of social media and its popularity grows, privacy issues are becoming increasingly salient. Recruitment of respondents on social networking sites (SNSs) as the territories often perceive as private, thus, may be hindered by users' privacy concerns. In this context requests for personal information, inherent to survey process, are additionally increasing dropout rate [7, 26].

Considering the complex relationship between privacy concerns and the level of online user self-disclosure [27], we aim at testing a range of hypotheses on how privacy-related requests to the respondents along with the presence of sensitive questions in the study, might affect respondents' decisions to leave the online experiment. Therefore, we expect that:

*H5: Dropout rate will be higher among those who denied access to any of their data (Vkontakte only).*

If this is the case *and* those who have given access are substantially different from those who have not, this may result in serious sample biases. Therefore, it is important to learn whether the second condition holds:

*H6: Age (a), gender (b) and education (c) of users who have given access to their private data will be significantly different from those who have not (Vkontakte only).*

Next, we assume that respondents' privacy-related anxiety might increase when they are presented questions about their political views. In non-democratic regimes such views are often perceived as threatening individuals' security, in case their identity is disclosed, while such disclosure may, in turn, be perceived as probable due to low trust to social institutions.

*H7: The two sensitive questions (a) on political loyalty towards the government and (b) on political relations with a neighboring country will be skipped by the majority of users.*

**Political Engagement.** Since the topic of the experiment is highly politicized, politically engaged individuals might have more incentives to complete the questionnaire. In non-democratic regimes grassroots political engagement is usually of contentious or even protest nature. Since the levels of protest across Russia are very uneven, with Moscow and St. Petersburg taking the lead [28], we expect regional differences in dropout:

*H8a: Dropout rate among users from Moscow and Saint Petersburg will be significantly lower than among users from other regions.*

*H8b: Dropout rate among users from Moscow, Moscow region, Saint Petersburg, or Leningrad region will be significantly lower than users from other regions.*

### 3.2 Features of Data Collection Tool as a Factor of Dropout

The second group of the investigated factors is patterns of the visual organization which might affect the dropout rate of respondents [29].

**SNS and Interface Type.** In our research we had to employ different web interfaces for different SNSs – an app for VK and a stand-alone website for FB. This suboptimal choice was made because, on the hand, we aim at analyzing both experimental data and the data retrieved from user accounts, where possible (Vkontakte), but, on the other hand, we cannot exclude Facebook. Although we tried to make both interfaces as similar as possible, they are not graphically identical, which is why we cannot isolate the effect of the SNS from the effect of web-interface. However, we assume that one or both should matter. On the one hand, the fact that the website does not request personal data may contribute to dropout prevention. On the other hand, redirection to an unknown external website with minimalistic layout, as opposed to staying on the familiar SNS, may work in the opposite direction. Therefore, we propose two alternative hypotheses:

*H9a: Dropout rate will be higher among the users of Vkontakte as compared to Facebook.*

*H9b: Dropout rate will be lower among the users of Vkontakte as compared to Facebook.*

**Number of Questions on the Screen.** Previous research on this factor is inconclusive. On the one hand, several studies carried out on different samples have established that single-item-per-screen design leads to a slightly smaller dropout rate than scroll design (in which the entire questionnaire is presented on a single online page) [30, 31]. On the other hand, this contradicts the earlier results of Couper et al. [32] who showed that the multi-item-per-screen design requires significantly less time to complete and produces lower data loss than single-item-per-screen version. Here, we find it important to separate two factors: number of questions per screen and the necessity to scroll. Therefore, we have designed and compared only screens with one or two questions fully visible without scrolling on any device:

*H10: The probability of dropout will be higher on the pages that contain more than one question.*

**Screen Size as a Factor of Dropout.** Wenz [11] has found that smaller screens increase the probability of respondents' dropout from web-surveys. However, a major reason for that may be that he has tested this hypothesis on a web-interface that was not adapted to different devices, while our tool has been adapted. Wenz's research thus does not differentiate between screen size per se and other factors, such as font size or the necessity to zoom and scroll in different directions. With our device-adaptable instrument, we can isolate and test the screen size effect, to the extent to which we have information on it (mobile vs desktop):

*H11: The probability of dropout among mobile device users will be higher than that among desktop users.*

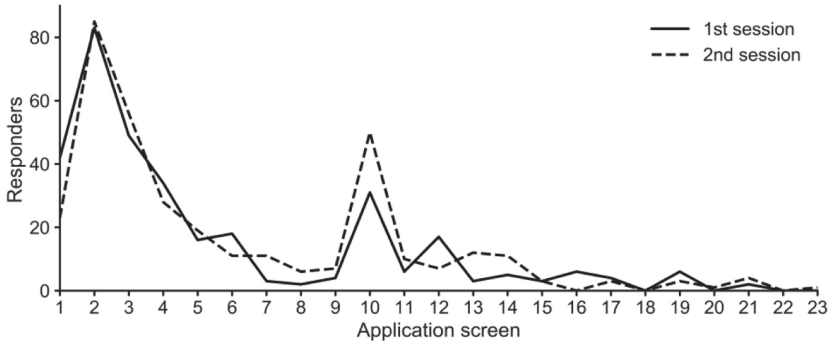
## 4 Results

### 4.1 Descriptive Analysis

**Vkontakte Campaign.** Vkontakte recruitment campaign lasted for three days (28–30 Oct 2019). The ad was displayed 198,609 times to the news feeds of 179,051 unique users (data provided by the platform ad managing service). Of them, 1,138 users clicked on the ad and 453 completed both the experimental and survey parts. The campaign was divided into two sessions, with roughly equal numbers of users in each (231 and 222 respectively). In total, 61.8% of responders were male, 18.9% were female, and 19.3% chose not to specify. The respondents aged from 18 to 55+ years (median = 19.0, IQR = 6.0). As expected, mobile device users dominated, amounting to 85.8% of all users (N = 976).

Vkontakte dropout rate constitutes 60.2% (685 users): 9.5% users launched the application by clicking on the ad and left without proceeding further (screen 1), 63.1% users left at the experimental stage (screens 2–9), and 27.4% users left at the survey stage (screens 10–23). As it can be seen from Fig. 1, the dropout decreases from the second to

the last screen, but peaks on screen 10. This was observed already after the first session when we stopped recruitment for express data analysis, however, it was not clear whether this peak was attributable to the shift between the experiment and the survey or to the content of the question displayed on screen 10. To isolate the former effect from the latter, we swapped the questions displayed on screens 10 and 15 (picking up those that would not substantially break the logic of the questionnaire). As can be seen in Fig. 1, user behavior did not change fundamentally after that.

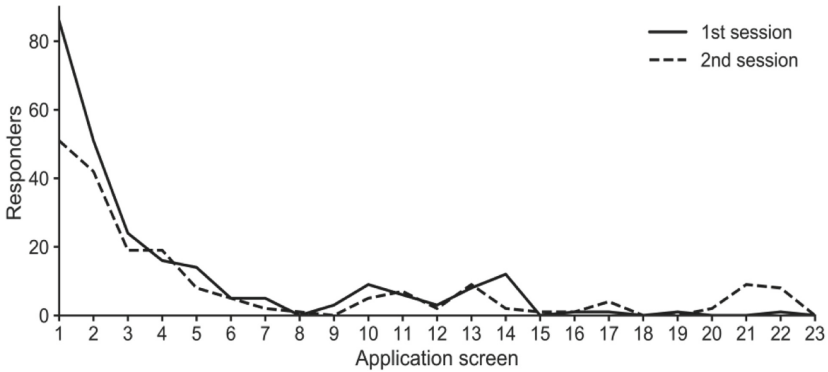


**Fig. 1.** Number of Vkontakte users who left the application on different screens (user dropout).

**Facebook Campaign.** Facebook recruitment campaign lasted for three days as well (2–4 Dec 2019) resulting in 6,644 displays of the ad. 678 participants clicked on it of whom 235 completed both the experimental and the survey parts. In total, 34.8% of respondents were male, 52.5% were female, and 12.7% chose not to specify. Mobile device users constituted 63.9% of all users (N = 433). The respondents aged from 18 to 55+ years, although the distribution of respondents’ ages dramatically differed from what we observed in Vkontakte campaign (median = 52.0, IQR = 16.0). Furthermore, judging from the data on age distributions of overall populations of both VK and FB provided by the respective ad managing services, both samples also diverge from the populations, but in the opposite directions. While our VK sample age distribution was skewed in favor of younger groups, as compared to the overall VK audience age distribution, our FB sample was, on the contrary, older than the overall FB audience (see Table. 1). This indicates highly non-random character of ad displays across the selected audiences either on one of on both SNSs.

Facebook dropout rate was equal to 65.3% (443 users): 30.9% users ran the application but did not participate (left on screen 1), 48.3% users left the application at the experimental stage (screens 2–9), 20.8% users left the application at the survey stage (screens 10–23). Maximum dropout rate was observed at the starting screen (screen 1), while on VK it was on screen 2 (first news item).

After the first session—with 114 users completing both the experiment and the survey part—we analyzed the obtained results and found out that the FB dropout curve, just like VK curve, in fact had a second smaller peak, although not around the politically sensitive questions, as expected, but around the two screens with four socio-demographic questions. When moved to the end of the questionnaire in the second session, these questions still produced a peak (screens 21–22) (Fig. 2).



**Fig. 2.** Number of Facebook users who left our application on different screens (user dropout).

**Overall Data.** To analyze the overall data, we combined Vkontakte and Facebook datasets. A total of 1,816 users participated in our pilot trials. Out of 767 responses with specified gender data, there were 52.7% male respondents ( $n = 404$ ), 30.2% female ( $n = 232$ ), and 17.1% chose not to specify ( $n = 131$ ). The participants had a median age of 23.0 (IQR = 33.0) and came from 76 regions of Russia. They mostly used mobile devices (77.6%,  $n = 1409$ ). The respondents reported the following education levels: 21% finished secondary school or college, 15.8% were studying at the university, 31.2% already received a bachelor or master’s degree, 4.1% had a doctor’s degree, 9.2% chose not to specify. Interestingly, 18.7% of respondents claimed to be studying at the secondary school, although we had limited the age of users to 18+ in ad campaigns.

The overall dropout rate was equal to 62.1% ( $n = 1128$ ): 17.9% left at the starting screen (screen 1) and did not enter the experimental stage, 57.3% left at the experimental stage (screens 2–9), 24.8% left at the survey stage (screens 10–23). Approximately 37.9% respondents finished both experimental and survey stages ( $n = 688$ ).



**Table 1.** The age distribution of Vkontakte and Facebook SNSs users.

| SNS       | Feature | Category | Respondents |            | Total audience |
|-----------|---------|----------|-------------|------------|----------------|
|           |         |          | N           | Percentage | Percentage     |
| Vkontakte | Age     | 18–24    | 379         | 75.2       | 27.9           |
|           |         | 25–34    | 34          | 6.7        | 32.0           |
|           |         | 35–44    | 16          | 3.2        | 14.4           |
|           |         | 45–54    | 18          | 3.6        | 6.0            |
|           |         | 55+      | 57          | 11.3       | 6.7            |
| Facebook  | Age     | 18–24    | 9           | 3.5        | 6.7            |
|           |         | 25–34    | 15          | 5.8        | 27.2           |
|           |         | 35–44    | 55          | 21.2       | 32.3           |
|           |         | 45–54    | 65          | 25.1       | 18.6           |
|           |         | 55+      | 115         | 44.4       | 14.8           |

### 4.2 Hypotheses Testing

To test our hypotheses, we ran a series of logistic regressions and  $\chi^2$  tests. The brief summary of hypotheses testing is given in Table 2.

**User Features as a Factor of Dropout.** Here, we start with demographic feature of our users.

*H1: Dropout rate will (a) or (b) increase with users’ age.*

The analysis based on logistic regression model did not reveal any association between dropout rate and users’ age on Facebook and Vkontakte datasets ( $\beta = 0.0037$ ,  $p = 0.456$ ). Thus, the hypotheses *H1a* and *H1b* were rejected. Additional analysis showed that the distribution of dropout rate by age resembled a quadratic function, however, this resemblance was not statistically significant either.

*H2: Dropout rate will be higher among male than among female respondents.*

In addition to gender data provided at the survey stage, we had previously obtained a random sample of 1.2 million open Vkontakte accounts including gender information (specified in user profiles). Gender data agreement between survey and user profile data subsets was estimated at 77.0%. We tested the hypothesis that dropout would be higher among women than men on both self-reported and observed data from respondents’ profiles in VK. In both cases no statistically significant difference was revealed (self-reported data:  $\chi^2 = 0.09$ ,  $p$ -value = 0.38; VK data:  $\chi^2 = 0.3$ ,  $p$ -value = 0.7). In this relation, the hypothesis H2 was rejected.

*H3: Dropout rate will be higher among respondents with lower education.*

The hypotheses did not reveal any association between dropout and user's level of education when tested separately on Facebook and Vkontakte datasets. At the same time, the overall data support the hypothesis *H3*, suggesting dropout dependency on user education. We used logistic regression to gain insight into this dependency pattern. Model fitting showed that education level was negatively associated with user dropout ( $\beta = -0.22$ ,  $p = 0.014$ ): a lower level of a user's education is associated with a higher probability of dropout. Thus, the hypothesis *H3* was supported.

*H4a: The higher the personal speed the lower the probability of dropout.*

*H4b: The lower the personal speed the lower the probability of dropout.*

Our tool recorded the time at which each user entered and left the experimental stage (screens 2–9). Personal speed was calculated as the mean time a respondent spent per experimental item, and this time was 6.3 times smaller among those who left the survey. We applied logistic regressions for each SNS sample and for the joint dataset and confirmed that the more time user spends at the experimental stage, the less is the probability of user dropout ( $\beta = -0.015$ ,  $p = 0.033$  for Facebook;  $\beta = -0.049$  and  $-0.036$ ,  $p = 5.5e-12$  and  $3.2e-11$  for Vkontakte and overall data). Thus, we accepted hypothesis 4b. Additional exploratory analysis showed that the distributions of speeds among those who left and those who stayed were a similarly left-skewed bell-shaped curve which means that there was no latent group of slow performers among the leavers.

*H5: Dropout rate will be higher among those who denied access to any of their data (Vkontakte only).*

As our application requested access to the private user data in Vkontakte, but not in Facebook, hypotheses 6 and 7 were tested on VK subsamples only. Dropout rate among users who provided full access to their private data was equal to 51.8% ( $n = 397$  of 582), but it was higher by 16.4% (68.2%,  $n = 288$  of 556) among users who denied access to the data. Based on the results of a chi-squared test, the hypothesis *H5* was accepted ( $\chi^2 = 31.29$ ,  $p = 2.2e-08$ ). Thus, dropout rate was significantly higher among the users who did not provide their private data.

*H6: Age (a), gender (b) and education (c) of users who have given access to their private data will be significantly different from those who did not (Vkontakte only).*

Access to the requested private data (friend list and group list), was provided by approximately a half of Vkontakte users (49.5%,  $n = 556$ ). However, all VK respondents who completed the survey indicated their age, gender, education and region, which allows us to compare access “deniers” and “providers” by these features. Statistically significant difference was obtained for user age as we compared the age distributions of both user groups with the two-sample Kolmogorov-Smirnov test ( $D = 0.26$ ,  $p = 1.6e-07$ ). In addition, logistic regression model revealed the negative relation between user age and user willingness to provide full access to their private data in Vkontakte ( $\beta = -0.013$ ,

$p = 0.00086$ ). Gender and education were not significantly associated with user intent to provide the private data ( $\chi^2 = 2.73$ ,  $p = 0.26$  and  $\chi^2 = 11$ ,  $p = 0.08$  respectively). Thus, hypothesis 6 was partly supported.

*H7: The two sensitive questions (a) on political loyalty towards the government and (b) on political relations with a neighboring country will be skipped by the majority of users.*

Hypothesis 7 was rejected as only 118 users of 705 skipped the question on the political loyalty ( $\chi^2 = 310.67$ ,  $p = 1.0$ ) and 123 users of 708 chose not to answer the question about the political relations ( $\chi^2 = 300.17$ ,  $p = 1.0$ ).

**Political Engagement.** The regional group of Moscow and Saint Petersburg among our respondents was found to be significantly underrepresented as compared with general Vkontakte population, as determined by our random sample of 1.2 million users ( $\chi^2 = 230.64$ ,  $p < 2.2e-16$ ). This actually contradicted our expectations as we had assumed that inhabitants of both cities, as more politicized, would be more interested in the topic of the experiment which had also lead us to hypothesis 8.

*H8a: Dropout rate among users from Moscow and Saint Petersburg will be significantly lower than among users from other regions.*

*H8b: Dropout rate among users from Moscow, Moscow region, Saint Petersburg, or Leningrad region will be significantly lower than users from other regions.*

Logistic regression analysis has shown that, indeed, users of all SNSs coming from Moscow, Saint Petersburg, and the corresponding surrounding regions finish both stages of the experiment with higher probability than users from other regions (Facebook:  $\beta = 5.39$ ,  $p = 9.3e-08$ ; Vkontakte:  $\beta = 2.88$ ,  $p = 2.0e-06$ ; overall data:  $\beta = 3.99$ ,  $p = 5.9e-15$ ). At the same time, this hypothesis was rejected for users coming from just Moscow and Saint Petersburg (Facebook:  $\beta = 18.49$ ,  $p = 0.97$ ; Vkontakte:  $\beta = 16.02$ ,  $p = 0.96$ ; overall data:  $\beta = 17.18$ ,  $p = 0.95$ ). Thus, only the hypothesis H8b was supported.

**Features of Data Collection Tool as a Factor of Dropout.** In this section we report the factors that theoretically can be manipulated.

*H9: Dropout rate will be (a) higher or (b) lower among the users of Vkontakte as compared to Facebook*

The tests of hypotheses H9 revealed the association between user dropout and SNS, with Vkontakte dropout rate being significantly lower than that of Facebook ( $\chi^2 = 4.47$ ,  $p = 0.017$ ). Thus, the hypothesis H9b was supported.

*H10: The probability of dropout will be higher on the pages that contain more than one question.*

We further explored how the number of questions on the screen might be related to the users' dropout. Logistic regression model failed to establish any significant association between these variables ( $p = 0.987$ ). Thus, the hypothesis H10 was rejected.

*H11: The probability of dropout among mobile device users will be higher than that among desktop users.*

Our instrument automatically collected user device data (operational system, browser, device type) that allowed us to estimate the mobile device percentage. Based

**Table 2.** Hypotheses testing results

| Hypotheses: Predictor; association with dropout rate or probability (except H6, H7)   | Status                      |   |                             |
|---|-----------------------------|---|-----------------------------|
|   | Facebook                    | Vkontakte                               | Overall                     |
| H1a: age; positive.<br>H1b: age; negative   | Rejected<br>Rejected        | Rejected<br>Rejected                    | Rejected<br>Rejected        |
| H2: gender (male), positive.  | -                           | Rejected                                | -                           |
| H3: education level; negative   | Rejected                    | Rejected                                | <b>Accepted</b>             |
| H4a: Speed; negative<br>H4b: Speed, positive  | Rejected<br><b>Accepted</b> | Rejected<br><b>Accepted</b>             | Rejected<br><b>Accepted</b> |
| H5: Access denial; positive   | —                           | <b>Accepted</b>                         | —                           |
| H6: Users who have given access to their private data will be significantly different from those who did not, by:<br>- Age (young)<br>- Gender (any)<br>- Education (any) | —                           | <b>Accepted</b><br>Rejected<br>Rejected | —                           |
| H7: The two sensitive questions will be skipped by the most users.  | Rejected                    | Rejected                                | Rejected                    |
| H8a: region (Moscow & St. Petersburg); negative<br>H8b: region (Moscow, St. Petersburg & surroundings); negative  | Rejected<br><b>Accepted</b> | Rejected<br><b>Accepted</b>             | Rejected<br><b>Accepted</b> |
| H9a: SNS (Vkontakte); positive<br>H9b: SNS (Vkontakte); negative  | -                           | -                                       | Rejected<br><b>Accepted</b> |
| H10: N of Qs per page; positive   | Rejected                    | Rejected                                | Rejected                    |
| H11: Device (mobile); positive  | Rejected                    | Rejected                                | Rejected                    |

on all SNS data and the overall dataset, mobile devices dominated over desktop (Facebook:  $\chi^2 = 51.58$ ,  $p < 3.4e-13$ , Vkontakte:  $\chi^2 = 580.82$ ,  $p < 2.2e-16$ , overall data:  $\chi^2 = 551.76$ ,  $p < 2.2e-16$ ). The highest percentage of mobile devices was observed among Vkontakte users (85.8%). However, the analysis did not reveal any association between dropout and the type of device which user used for completing the survey on either Facebook or Vkontakte datasets (Facebook:  $\chi^2 = 0.005$ ,  $p = 0.528$ , Vkontakte:  $\chi^2 = 5.08e-30$ ,  $p = 0.5$ ). Thus, the hypothesis H11 was rejected.

## 5 Discussion

The results of this research have several important implications for dropout reduction and for a broader methodology of online experiments and surveys carried out via social networking sites.

First, although we targeted the entire Russian populations of both VK and FB, the resulting samples had significant differences from those populations in a number of demographic characteristics, the first of which is age. If age bias were the same in both networks, one could assume more interest in the experiment topic among certain age groups. However, the biases were both very strong and opposite in the two SNSs which suggests that ad managing systems employ both highly non-random and non-transparent targeting strategies. Therefore, it is highly recommended that researchers control demographic proportions of their samples while administrating SNS-based surveys. Under-representation of Moscow and St. Petersburg in our VK sample most probably has a different cause. Here we used our own data on population distribution (the only available) dating back to 2015. Since then the penetration of VK beyond big cities might have grown. Also, these two cities are likely to host the largest share of business and spam accounts that might have naturally self-deselected from our research.

Second, we observed a slightly lower level of dropout in VK than in FB (60.2% vs 65.3%), although the latter SNS, being more politicized, could be expected to contain more users interested in the topic of our experiment. A possible explanation might be derived from the fact that the largest number of dropouts among FB users happened on the very first page of the website, while among VK users this maximum happened on the next page containing the first news item. This suggests that additional losses from FB might be explained by the website design or by the fact of it being stand-alone; however, this hypothesis should be tested by isolation of this effect from that of SNS.

Third, the dropout curves were similar in SNSs declining from the beginning to the end and produced two peaks: at the transition from the experimental part to the survey and on the screens with socio-demographic questions. While the first peak cannot be avoided, the sensitive questions can be shifted to the end. This may not increase the share of users willing to answer them, but it will increase the share of surveys complete in all other respects. It is also interesting that political questions were not perceived as sensitive by respondents, even if they were asked after socio-demographic data had been obtained.

Fourth, although we were afraid that mobile device users would be either distracted from the experiment by their environments or would have difficulties reading from the smaller screens, in fact, dropout among them was the same as among desktop users. The

overall high prevalence of mobiles over desktops may suggest that by now users have been already well adapted to small screens and to noisy and changing environments, and that mobiles can be used for research along with desktops. Likewise, once all the content of each page could fit the screen and no scrolling was needed, the number of questions per screen did not influence dropout.

Fifth, we have found out that individual speed or time per item is by no means an indicator of perceived difficulty; instead, high speed reliably indicates lack of interest and motivation. It is by far the best predictor of dropout.

Sixth, demographic features of respondents had different effects on their dropout. Gender was definitely unimportant, while people with higher education were visibly more inclined to complete our survey. This might be explained by its topic (politics), as well as by the fact that it appeared as an intellectual challenge. Influence of age demands further research. Although we find that dropout was higher among both the oldest and the youngest groups, this trend has not only been statistically confirmed, and, furthermore, may ultimately be some artifact of targeting which has resulted in overrepresentation of these two age groups in our FB and VK samples respectively.

Finally, as those users who denied access to their data were more likely to drop out, and simultaneously were more likely to be older (as judged from the age of those “deniers” who reached the stage of indicating their age), older age groups might be systematically underrepresented in our VK sample. This is consistent with the direction of age bias of our sample as compared to the overall VK population and may serve as an additional explanation for it, along with the features of VK ad managing service algorithm.

Overall, we can conclude that SNS-administered online experiments are quite efficient, fast and relatively inexpensive instruments, although they are not without limitations. Despite low rates of conversion of ad displays into clicks and fairly high dropout rates, such instruments can quickly collect large amounts of completed surveys. The highest caution should be exercised in relation to sample construction and recruitment. The biggest challenge for social scientists and psychologists is, however, task formulation for software developers and usability testing.

**Acknowledgements.** The research was implemented in the framework of the Russian Scientific Fund Grant № 19-18-00206 (2019–2021) at the National Research University Higher School of Economics.

## References

1. Vicente, P., Reis, E.: Using questionnaire design to fight nonresponse bias in web surveys. *Soc Sci Comput Rev* **28**(2), 251–267 (2010). <https://doi.org/10.1177/0894439309340751>
2. Dillman, D., Tortora, R.L., Conradt, J., Bowker, D.: Influence of plain vs. fancy design on response rates for Web surveys. In: *Proceedings of the Joint Statistical Meetings*. American Statistical Association, Alexandria (1998)
3. Galesic, M.: Dropouts on the web: effects of interest and burden experienced during an online survey. *J. Off. Stat.* **22**(2), 313 (2006)
4. Deci, E.L.: *Intrinsic Motivation*. Plenum Press, New York (1975)

5. O'Neil, K.M., Penrod, S.D., Bornstein, B.H.: Web-based research: methodological variables' effects on dropout and sample characteristics. *Behav. Res. Methods Instrum. Comput.* **35**(2), 217–226 (2003). <https://doi.org/10.3758/bf03202544>
6. Bosnjak, M., Tuten, T.L.: Prepaid and promised incentives in web surveys—an experiment. *Soc. Sci. Comput. Rev.* **21**, 208–217 (2003). <https://doi.org/10.1177/0894439303021002006>
7. Frick, A., Bächtiger, M.T., Reips, U.D.: Financial incentives, personal information and drop-out rate in online studies. In: Reips, U.D., Bosnjak, M. (eds.) *Dimensions of Internet Science*, pp 209–219. Pabst Science Publishers, Lengerich (1999). <https://doi.org/10.5167/uzh-19758>
8. Knapp, F., Heidingsfelder, M.: Drop-out analysis: effects of the survey design. In: Reips, U.D., Bosnjak, M. (eds.) *Dimensions of Internet Science*, pp. 221–230. Pabst Science Publishers, Lengerich (1999)
9. Healey, B.: Drop downs and scroll mice: the effect of response option format and input mechanism employed on data quality in web surveys. *Soc. Sci. Comput. Rev.* **25**(1), 111–128 (2007). <https://doi.org/10.1177/0894439306293888>
10. Couper, M.P., Traugott, M.W., Lamias, M.J.: Web survey design and administration. *Public Opin. Q.* **65**(2), 230–253 (2001). <https://doi.org/10.1086/322199>
11. Wenz, A.: Completing web surveys on mobile devices: does screen size affect data quality? (No. 2017-05). ISER Working Paper Series (2017)
12. Heerwegh, D., Loosveldt, G.: An evaluation of the effect of response formats on data quality in web surveys. *Soc. Sci. Comput. Rev.* **20**(4), 471–484 (2002). <https://doi.org/10.1177/089443902237323>
13. Crawford, S.D., Couper, M.P., Lamias, M.J.: Web surveys: perceptions of burden. *Soc. Sci. Comput. Rev.* **19**(2), 146–162 (2001). <https://doi.org/10.1177/089443930101900202>
14. Busemeyer, J.R., Townsend, J.T.: Decision field theory: a dynamic cognition approach to decision making. *Psychol. Rev.* **100**, 432–459 (1993). <https://doi.org/10.1037//0033-295x.100.3.432>
15. Perriánñez, Á., Saas, A., Guitart, A., Magne, C.: Churn prediction in mobile social games: towards a complete assessment using survival ensembles. In: *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp 564–573. IEEE Press (2016). <https://doi.org/10.1109/dsaa.2016.84>
16. Hadji, F., Sifa, R., Drachen, A., Thureau, C., Kersting, K., Bauchhage, C.: Predicting player churn in the wild. In: *IEEE Conference on Computational Intelligence and Games*, pp 1–8. IEEE Press (2014). <https://doi.org/10.1109/cig.2014.6932876>
17. Prensky, M.: Digital natives, digital immigrants. *On Horiz.* **9**(5), 1–6 (2001). <https://doi.org/10.1108/10748120110424816>
18. Correa, T., Hinsley, A.W., de Zúñiga, H.G.: Who interacts on the Web?: The intersection of users' personality and social media use. *Comput. Hum. Behav.* **26**(2), 247–253 (2010). <https://doi.org/10.1016/j.chb.2009.09.003>
19. Muscanell, N.L., Guadagno, R.E.: Make new friends or keep the old: gender and personality differences in social networking use. *Comput. Hum. Behav.* **28**(1), 107–112 (2012). <https://doi.org/10.1016/j.chb.2011.08.016>
20. Fogel, J., Nehmad, E.: Internet social network communities: risk taking, trust, and privacy concerns. *Comput. Hum. Behav.* **25**, 153–160 (2009). <https://doi.org/10.1016/j.chb.2008.08.006>
21. Raacke, J., Bonds-Raacke, J.: MySpace and Facebook: applying the uses and gratifications theory to exploring friend-networking sites. *Cyberpsychol. Behav.* **11**(2), 169–174 (2008). <https://doi.org/10.1089/cpb.2007.0056>
22. Groves, R.M., Couper, M.P.: *Nonresponse in Household Interview Surveys*. Wiley, New York (1998)

23. Chyung, S.Y.: Systematic and systemic approaches to reducing attrition rates in online higher education. *Am. J. Distance Educ.* **15**(3), 36–50 (2001). <https://doi.org/10.1080/08923640109527092>
24. Terrell, R.S.: A longitudinal investigation of the effect of information perception and focus on attrition in online learning environments. *Internet High. Educ.* **8**(3), 213–219 (2005). <https://doi.org/10.1016/j.iheduc.2005.06.003>
25. Borbora, Z., Srivastava, J., Hsu, K.W., Williams, D.: Churn prediction in MMORPGs using player motivation theories and an ensemble approach. In: *IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp 157–164. IEEE Press (2011). <https://doi.org/10.1109/passat/socialcom.2011.122>
26. Heerwegh, D., Loosveldt, G.: An evaluation of the semi-automatic login procedure. *Soc. Sci. Comput. Rev.* **21**, 223–234 (2003). <https://doi.org/10.1177/0894439303021002008>
27. Kokolakis, S.: Privacy attitudes and privacy behavior: a review of current research on the privacy paradox phenomenon. *Comput. Secur.* **64**, 122–134 (2017). <https://doi.org/10.1016/j.cose.2015.07.002>
28. Information Agency Regnum. <https://regnum.ru/news/polit/2588639.html>
29. Ganassali, S.: The influence of the design of web survey questionnaires on the quality of responses. *Surv. Res. Methods* **2**, 21–32 (2008). <https://doi.org/10.18148/srm/2008.v2i1.598>
30. Manfreda, L.K., Batagelj, Z., Vehovar, V.: Design of web survey questionnaires: three basic experiments. *J. Comput. Mediat. Commun.* **7**, 3 (2002). <https://doi.org/10.1111/j.1083-6101.2002.tb00149.x>
31. Peytchev, A., Couper, M., McCabe, S., Crawford, S.: Web survey design: paging versus scrolling. *Public Opin. Q.* **70**, 596–607 (2006). <https://doi.org/10.1093/poq/nfl028>
32. Couper, M.P., Traugott, M., Lamias, M.: Effective survey administration on the Web. In: *Midwest Association for Public Opinion Research Conference*, Chicago, Illinois (1999)





# Illicit Drug Purchases via Social Media Among American Young People

Atte Oksanen<sup>1</sup> , Bryan Lee Miller<sup>2</sup> , Iina Savolainen<sup>1</sup> , Anu Sirola<sup>1</sup> ,  
Jakob Demant<sup>3</sup> , Markus Kaakinen<sup>4</sup> , and Izabela Zych<sup>5</sup> 

<sup>1</sup> Faculty of Social Sciences, Tampere University, Kalevantie 5, 33014 Tampere, Finland  
atte.oksanen@tuni.fi

<sup>2</sup> Clemson University, Clemson, USA

<sup>3</sup> University of Copenhagen, Copenhagen, Denmark

<sup>4</sup> University of Helsinki, Helsinki, Finland

<sup>5</sup> University of Cordoba, Cordoba, Spain

**Abstract.** Illicit drugs are sold online. Besides cryptomarkets, young people today are also using social media to buy and sell different drugs. The aim of this nationwide study was to investigate the phenomenon of buying drugs from social media among American young people. Relatively few studies have investigated young people buying drugs online and, therefore, it is important to know more about the psychological and social risk factors of this behavior. The participants of the study were 15–25-year-olds from the U.S. ( $M = 20.05$ ; 50.17% female). Buying drugs online was utilized as an outcome variable. The covariates included measures of impulsivity and delay of gratification, sense of belonging to online communities, online homophily, friends sharing risk material online, psychological distress, and measures for addictive behaviors including hazardous drinking, problem gambling, and compulsive Internet use. Results showed that buying drugs online is still a relatively rare phenomenon, but many of those buying drugs online used social media services to do so. Buying drugs online was associated with higher impulsivity and lower measures of delay discounting indicating self-control problems. Online buyers also had multiple problems with mental wellbeing, as they reported more psychological distress, problem gambling, and compulsive Internet use than those drug users who had not bought drugs online. The existence and comorbidity of these problems suggest that drug availability online might worsen their situation. As impulsive decisions are especially easy to make on social media, more focus should be placed on youth behavior on mainstream social media services.

**Keywords:** Drugs · Internet · Social media · Young adults · Wellbeing

## 1 Introduction

The Internet offers easy access to legal and illegal activities taking place on open social media services and encrypted services that use, for example, the Tor network [1, 2]. Over recent years, online drug sales have been investigated with studies focusing on

the Silk Road darknet market and its successors [3–6]. These studies have employed surveys and interviews [7, 8] along with web crawlers [2, 3, 5, 9, 10]. Drug buyers often prefer cryptomarkets because they are designed to protect user anonymity [1, 2, 5, 6]. Buyers also prioritize cryptomarkets over street markets for personal safety and better quality of drugs [7, 11]. Studies have viewed cryptomarket users as a “technological drug subculture” [8]. However, there is a lack of studies comparing them to other drug users and non-users, and hence, claims about these users often remain anecdotal in published studies.

Despite the considerable amount of attention given to cryptomarkets, there are indications that people might also use mainstream social media sites, such as Instagram, to buy drugs. Recently, a Nordic project employed qualitative interviews and online ethnography to analyze the phenomenon in Denmark, Finland, Iceland, Norway and Sweden. Results showed that drugs were sold via Facebook in Denmark, Iceland and Sweden, but not in Finland and Norway where other social media services, such as Instagram, were utilized [12]. There are, however, very sparse studies on potential drug purchases on mainstream social media sites using nationally representative data. Within a EMCDDA report on Danish young people it was reported that among buyers of cannabis, 6% do it online and 36% of the buyers of drugs other than cannabis purchased drugs online [13]. Therefore, it is important to understand this phenomenon from a broader perspective and in relation to other online drug purchases made on cryptomarkets.

We ground our study in social psychological and addiction research, both of which help in understanding the psychological and social risk factors related to buying drugs online. The relevance of self-control on human behavior has been widely noted in both social psychology and criminology [14, 15]. Integrative meta-analyses showed that high self-control is widely related to positive outcomes and low self-control to deviant and addictive behaviors [16]. Addicts are well known for their tendency towards immediate gratification, and they prefer smaller immediate rewards over larger delayed rewards [17–19]. This phenomenon related to impulsivity has been examined with delay of gratification, which concerns individuals’ ability to resist a readily available stimulus, or with a form of delay discounting that measures individuals’ cognitive processes in devaluating a hypothetical large reward over an immediate small one [20]. Delay discounting is tested by asking people whether they would prefer a smaller sum of money right away or a larger sum after some time [21, 22]. Present-biased preference in choices is more common among young people who prioritize fast rewards [23].

Peer influence is recognized as a risk factor for deviant behavior and substance abuse among young people—especially in social learning theory, which underlines the idea that people learn about crime through social interaction [24]. Social learning theory could be especially powerful in explaining deviant behavior online, as it is easy to find like-minded people online and be influenced by them [25]. Especially homophily (i.e. the tendency of people to form relationships with others who are similar to them) [26] is often even encouraged on social media sites due to algorithms connecting people with each other and different possibilities to connect with each other [27]. Homophily is shown to increase when possibilities for social selectivity expand [28]. Homophily and easy access to like-minded people online have been previously noted in research on deviant behavior [27, 29, 30]. For example, online communities potentially have a role

in contributing to various forms of risk-behavior among young people, including online gambling [31].

Problems with mental wellbeing coincide and develop with addictive behaviors, including excessive drug use [19]. Psychological distress is defined as unpleasant feelings of depression and anxiety. According to previous studies, psychological distress is associated with drug use [32–34]. Moreover, drug use has a high comorbidity with hazardous drinking [33], problem gambling [35], and compulsive or addictive use of the Internet [36]. From addiction research perspective, drug users are looking for effects, but when the use of substances increase, they start falling back to where they eventually started, meaning that they are more psychologically distressed and less happy [37]. In our view, buying drugs is a potential indicator of substance abuse problems, and we expect it to be associated with psychological distress, lower happiness, hazardous drinking, problem gambling and compulsive Internet use. Our hypotheses for this study were:

H1: Among drug users, lower self-control is associated with higher likelihood of buying drugs online.

H2: Among drug users, peer influence online is associated with higher likelihood of buying drugs online.

H3: Among drug users, mental wellbeing issues are associated with higher likelihood of buying drugs online.

## 2 Method

### 2.1 Participants

The participants of the study were 15–25-year-olds ( $N = 1212$ ) living in the U.S. ( $M_{\text{age}} = 20.1$ ,  $SD = 3.2$ ; 50.2% female). We recruited them in January 2018 from an online panel administrated by Dynata (formerly Survey Sampling International), which offers access to demographically balanced panels for research purposes [18, 38, 39]. Data were confirmed to mirror the US population aged 18 to 25 years in terms of age, gender and geographical area. The participants were from 50 different states, with highest response rates coming from California (12.51%) and New York (7.39%), Texas (6.37%), Pennsylvania (5.35%) and Florida (5.01%). The participants were 56.17% White (non-Hispanic), 18.01% Hispanic, 12.91% Black or African American, and 9.51% Asian American, and 94.66% of them had a high school diploma. We found these figures to be close to population estimates of the 15–25-year-olds in the U.S. [38]. We did not apply weights due to the close resemblance to the population estimates.

### 2.2 Procedure

We conducted the survey with LimeSurvey software that was run with the University server. The study was reviewed and approved by The Academic Ethics Committee of the Tampere Region in Finland (decision 62/2016). The voluntary participants were informed about the aims of the study, and they had the possibility to withdraw, totally or partially, from the survey at any time during the completion process. The participants

were also provided with information on how to follow the progress of the study. The data collection ensured the anonymity of the participants and the datasets were de-identified after the data collection. Median survey response time for the full survey was 875 s (14.58 min). Data quality checks were run with both response time and attention check questions included in the questionnaire. The data do not include missing data as each question was set mandatory.

### 2.3 Measures

**Drug Use and Buying Drugs Online.** We asked the participants whether they had “used or experimented with substances other than alcohol or tobacco to get high?” The response options were yes and no. We then asked them to specify the types of drugs used and the frequency of use. After this, the participants answered the question: “Did you use the Internet for purchasing these drugs?” (*yes/no*). Next, we asked to identify different online resources for purchasing drugs, ranging from darknet marketplaces to various social media platforms, for example Facebook and Instagram. Types of drugs asked for were: a) cannabis, b) synthetic cannabinoids, LSD, magic mushrooms, or other comparable hallucinogens, c) amphetamines, ecstasy, cocaine or other stimulants, d) opiates, e) pharmaceutical opioids, f) gamma, GBL and other similar drugs, and g) other pharmaceuticals. Those reporting “I use regularly” were categorized into regular cannabis users and regular users of other drugs (e.g., stimulants, opiates, and hallucinogens).

**Self-control.** We used two measures of self-control constructs. These were the Eysenck Impulsiveness Scale (EIS) [40]. The measure showed an acceptable inter-item reliability of .69, items were responded on a dichotomous yes/no scale and the total score ranged from 0 to 5 with higher figures indicating higher impulsivity. Table 1 shows the details for this and other measures used in the analysis of those young people who have experimented with drugs ( $n = 253$ ). The other measure for self-control was delay of gratification that was measured with 3-items concerning whether the participants would either receive a certain sum of money immediately or a larger sum after 33 days. The received lower sum varied from 28 euros to 40 euros, and the larger sum ranged from 62 euros to 80 euros. Our measure was grounded on behavioral economics literature on delay discounting [21, 41] and a similar test for delay of gratification has also been used in social psychological studies on impulsivity and economic behavior [22]. The scale had good internal consistency ( $\alpha = .87$ ) and the total score ranged from 0 to 3. A higher figure indicates higher delay of gratification.

**Peer Influence.** We asked the respondents about their sense of belonging to online communities. This was measured with a single-item: “How strongly do you feel you belong to an online community?” The response scale was from 1 (*not at all*) to 10 (*very strongly*). The item has been previously validated in several studies [18, 27, 30, 42]. Homophily online was measured with the homophily subscale of the Identity Bubble Reinforcement Scale (IBRS-9) consisting of 3 items with a response scale ranging from 1 (*does not describe me at all*) to 10 (*describes me completely*): 1) “In social media, I prefer interacting with people who are like me”, 2) “In social media, I prefer interacting with people who share similar interests with me” and 3) “In social media, I prefer interacting

with people who share my values.” [25] The measure showed good inter-item reliability ( $\alpha = .85$ ). The measure involving potential peer influence online concerned seeing friends sharing risk material. This was asked with a question: “do your friends share gambling content in social media?” Response scale ranged from 0 (*never*) to 7 (*daily*).

**Mental Wellbeing.** We measured psychological distress with the 12-item General Health Questionnaire (GHQ-12), which has been widely used in general population studies [43]. The scale had excellent internal consistency of .90. Likert scoring (0-1-2-3) was applied [43, 44], and the total score ranged from 0 to 36, with higher scores indicating higher psychological distress. We also used a standard single-item measure of happiness: “All things considered, how happy would you say you are?” The response scale ranged from 1 (*extremely unhappy*) to 10 (*extremely happy*). This question has been previously applied to studies on young people and social media [27, 30, 45].

**Table 1.** Descriptive statistics on U.S. drug users (n = 253).

| Categorical variables                | Coding | %     | n     |          |
|--------------------------------------|--------|-------|-------|----------|
| Bought drugs online                  | No     | 89.72 | 227   |          |
|                                      | Yes    | 10.28 | 26    |          |
| Continuous variables                 | Range  | M     | SD    | $\alpha$ |
| <i>Self-control</i>                  |        |       |       |          |
| Impulsivity                          | 0–5    | 2.21  | 1.65  | .69      |
| Delay of gratification               | 0–3    | 1.72  | 1.32  | .87      |
| <i>Peer influence</i>                |        |       |       |          |
| Sense of belonging to friends online | 1–10   | 4.91  | 2.74  | –        |
| Homophily online                     | 1–10   | 6.86  | 2.12  | .85      |
| Friends share risk content online    | 0–7    | 1.58  | 2.15  | –        |
| <i>Mental wellbeing</i>              |        |       |       |          |
| Psychological distress               | 0–36   | 15.56 | 7.64  | .90      |
| Happiness                            | 1–10   | 6.30  | 2.55  | –        |
| <i>Addictive behaviors</i>           |        |       |       |          |
| Hazardous drinking                   | 0–12   | 3.15  | 2.54  | .80      |
| Problem gambling                     | 0–20   | 2.06  | 3.77  | .87      |
| Compulsive Internet use              | 0–56   | 24.25 | 13.82 | .94      |

**Addictive Behaviors.** We measured hazardous drinking with the 3-item AUDIT-C [46], which had good internal consistency ( $\alpha = .80$ ). Problem gambling was measured with the South Oaks Gambling Screen for problem gambling [47] ( $\alpha = .87$ ) and compulsive Internet use with the Compulsive Internet Use Scale [48] ( $\alpha = .94$ ).

**Control Factors.** We used gender, age and social media activity as controls. We measured social media activity with a set of 12 questions involving how often the respondents used the most popular social media sites. Then, we turned an aggregated sum variable into a dummy variable on the basis of the median (0 = *low*, 1 = *high*).

## 2.4 Statistical Techniques

The analysis focused first on showing prevalence and different ways of buying drugs online by using descriptive analysis methods. We report these findings in the text. The main analyses focused on those participants who had experimented with drugs ( $n = 253$ ). Within this group we conducted a comparison between those who had bought drugs online and those who had not. We report chi-square tests results for categorical variables and a mean comparison based on t-test and Kruskal-Wallis test. Due to the small number of young people buying drugs online in the sample, we used penalized maximum likelihood logistic regression (i.e., Firth method) to reduce potential small-sample bias [49–51]. As our sample sizes were over 200 and the events presenting more than 10%, our analyses could have been done with standard logistic regression. However, Firth method provides more robust findings in cases when either sample size or events are low. The analyses were run with the `firthlogit`-command in Stata 16.0 and age, gender, and social media activity were used as controls in each model. We report odd ratios (OR), their 95% confidence intervals and p-values for statistical significance.

## 3 Results

Results showed that about 21% (253/1212) of participants reported they had used or experimented with substances other than alcohol or tobacco to get high. Cannabis was clearly the most experimented drug among those in the study. About 10% (26/253) of the drug experimenters and users reported using the Internet to buy drugs. This is only 2% of the whole sample (26/1212) meaning that, on the population level, buying drugs online is a marginal phenomenon. Buying drugs online was more common among males (3.31%, 20/584) than females (0.99%, 6/602,  $p = .005$ ). Buyers were older than non-buyers (age 22.62 vs. 20.00,  $p < .001$ ) and more commonly regular users than other users (65.38 vs. 34.62%,  $p < .001$ ). Among regular cannabis users, only less than 4% (3.75%, 3/80) had bought drugs online, while among regular users of other drugs, the percentage of online buyers was 44% (43.75%, 14/32,  $p < .001$ ).

In total, 69% of participants who bought drugs online used social media sites. The remainder (31%) of those who purchased drugs online did so only via darknet services. Respondents were also asked which services or sites they used when purchasing drugs online. The most common sites were Instagram (42%), Facebook (38%), and Craigslist (19%) with one respondent indicating the use of an online legal cannabis delivery service. About half of all the respondents had used several different sites and services to buy drugs.

Our main analyses focused on those who had experimented with drugs ( $n = 253$ , see Table 2). The descriptive analyses on comparison of means with Kruskal-Wallis test showed that those buying drugs online reported higher impulsivity ( $p < .001$ ), lower delay of gratification ( $p = .001$ ), higher frequency of friends sharing risk material online ( $p < .001$ ), higher psychological distress ( $p < .001$ ), higher hazardous drinking ( $p < .001$ ), higher problem gambling ( $p < .001$ ) and higher compulsive Internet use ( $p = .001$ ) than other drug users. Sense of belonging to friends online, homophily online and happiness were not statistically significant predictors of buying drugs online.

The penalized maximum likelihood logistic regression models were adjusted for age, gender, social media activity and regular drug use. Online drug buyers reported higher

**Table 2.** Predictors of buying drugs online among young U.S. drug users (n = 253).

| Buying drugs online                  |       |       |       |       |       |                            |       |
|--------------------------------------|-------|-------|-------|-------|-------|----------------------------|-------|
|                                      | No    |       | Yes   |       | p     | Logistic regression models |       |
|                                      | M     | SD    | M     | SD    |       | OR [95% CI]                | p     |
| <i>Self-control</i>                  |       |       |       |       |       |                            |       |
| Impulsivity                          | 2.09  | 1.66  | 3.27  | 1.15  | <.001 | 1.58 [1.61–2.15]           | .004  |
| Delay of gratification               | 1.81  | 1.30  | 0.85  | 1.19  | .001  | 0.61 [0.43–0.87]           | .006  |
| <i>Peer influence</i>                |       |       |       |       |       |                            |       |
| Sense of belonging to friends online | 4.82  | 2.69  | 5.73  | 3.07  | .1435 | 1.03 [.87–1.20]            | .753  |
| Homophily online                     | 6.91  | 2.06  | 6.49  | 2.59  | .49   | 0.91 [0.75–1.10]           | .322  |
| Friends share risk content online    | 1.39  | 2.02  | 3.27  | 2.52  | <.001 | 1.22 [1.01–1.46]           | .034  |
| <i>Mental wellbeing</i>              |       |       |       |       |       |                            |       |
| Psychological distress               | 14.97 | 7.51  | 20.69 | 6.92  | <.001 | 1.10 [1.04–1.17]           | .002  |
| Happiness                            | 6.38  | 2.41  | 5.58  | 3.52  | .334  | 0.86 [0.74–1.01]           | .060  |
| <i>Addictive behaviors</i>           |       |       |       |       |       |                            |       |
| Hazardous drinking                   | 2.93  | 2.44  | 5.04  | 2.65  | <.001 | 1.16 [.98–1.36]            | .079  |
| Problem gambling                     | 1.32  | 2.65  | 8.54  | 5.62  | <.001 | 1.34 [1.20–1.50]           | <.001 |
| Compulsive internet use              | 23.35 | 13.45 | 32.12 | 14.78 | .006  | 1.04 [1.01–1.07]           | .023  |

impulsivity (OR = 1.58,  $p = .004$ ), lower delay of gratification (OR = 0.61,  $p = .006$ ), higher frequency of friends sharing risk material online (OR = 1.22,  $p = .034$ ), higher psychological distress (OR = 1.10,  $p = .002$ ), higher hazardous drinking (OR = 1.16,  $p = .079$ ), higher problem gambling (OR = 1.34,  $p < .001$ ) and higher compulsive Internet use (OR = 1.04,  $p = .023$ ) than other drug users.

## 4 Discussion

This study investigated the phenomenon of buying drugs from social media among American young people. Very few previous studies have analyzed this phenomenon using nationwide samples and our study provided important new evidence on the prevalence and risk factors of this phenomenon. The results indicated that buying drugs online is a rare phenomenon, with only 2% of young people aged 15 to 25 using online resources to buy drugs. Yet, every tenth young drug user had bought drugs online. The prevalence figures of our study, despite being low, are also understandable given that around five to ten percent of individuals commit at least half of crimes [52]. It is also remarkable that in total 69% of those buying drugs online had done so via social media. This result confirms recent findings based on qualitative studies on social media drug sales [12]

and is also comparable to results based on a Danish sample [13]. They also indicate that online drug studies should not only focus on cryptomarkets.

Results underline the relevance of low self-control which confirms our first hypothesis. This result is generally consistent with literature on low self-control and impulsivity [16, 24]. In our case, both impulsivity and low delay of gratification were significantly related to buying drugs online in all analyses and also in final regression models adjusting for age, gender, Internet use and regular drugs. These results also contrast previous literature on online drug buyers that have been based on limited qualitative samples and have portrayed online buyers as technologically savvy users who can regulate themselves [11, 53]. Future studies should continue investigating online drug purchases using solid survey designs, instead of very selected qualitative samples or otherwise biased samples.

Our second hypothesis concerned the potential role of peer influences. Negative peer influence coming from online communities has been found in several previous studies [27, 29, 31, 54, 55]. We found some indication of this, but it was limited to our data to peers sharing risk content online. We did not find evidence that simple perceived homophily or strong belonging to online groups or communities would have a role in buying drugs online. Future studies should continue to investigate the dynamics of how peers distribute different material online and how this helps us to understand the phenomenon of buying drugs online.

Our third hypothesis concerned mental wellbeing. As expected, those buying drugs online reported psychological distress. Happiness was not a statistically significant predictor, but we found that addictive behavior in general was strongly associated with buying drugs online. Different forms of addictive behaviors included hazardous drinking, problem gambling and compulsive Internet use. These indicate that those buying drugs online have multiple mental health issues. These findings also confirm previous research results on the associations of drug use in general with different problem behaviors [31, 33, 34]. However, our analysis also found that self-reported happiness was not significantly related to buying drugs online.

Our analysis was limited to a cross-sectional design and a small number of young people buying drugs online. The strength of the study was that it employed a relatively large nationwide sample and is among the first studies to focus on buying drugs online, an emerging problem behavior. Our results were also very consistent and based on validated scales. Hence, we believe that these results fill important gaps in knowledge and will guide future studies.

We conclude that online buyers have multiple self-control and mental health problems. Easy availability of drugs online and especially on open social media might worsen their situation as impulsive decisions are especially easy to make on social media. Our results call for more attention to youth behavior on mainstream social media services.

**Acknowledgements.** This study was funded by the Finnish Foundation for Alcohol Studies (Problem Gambling and Social Media Project, 2017–2019, PI: Atte Oksanen).

## References

1. Martin, J.: *Drugs on the Dark Net: How Cryptomarkets are Transforming the Global Trade in Illicit Drugs*. Palgrave Macmillan, Basingstoke (2014)



2. Nurmi, J., Kaskela, T., Perälä, J., Oksanen, A.: Seller's reputation and capacity on the illicit drug markets: 11-month study on the Finnish version of the Silk Road. *Drug Alcohol Depend.* **178**, 201–207 (2017)
3. Aldridge, J., Décary-Héту, D.: Hidden wholesale: the drug diffusing capacity of online drug cryptomarkets. *Int. J. Drug Policy* **35**, 7–15 (2016)
4. Décary-Héту, D., Giommoni, L.: Do police crackdowns disrupt drug cryptomarkets? A longitudinal analysis of the effects of Operation Onymous. *Crime Law Soc. Change* **67**(1), 55–75 (2017). <https://doi.org/10.1007/s10611-016-9644-4>
5. Demant, J., Munksgaard, R., Houborg, E.: Personal use, social supply or redistribution? Cryptomarket demand on Silk Road 2 and Agora. *Trends Organized Crime* **21**(1), 42–61 (2016). <https://doi.org/10.1007/s12117-016-9281-4>
6. Martin, J., Munksgaard, R., Coomber, R., Demant, J., Barratt, M.J.: Selling drugs on darkweb cryptomarkets: differentiated pathways, risks and rewards. *Br. J. Criminol.* (2019). <https://doi.org/10.1093/bjc/azz075>
7. Barratt, M.J., Ferris, J.A., Winstock, A.R.: Use of silk road, the online drug marketplace, in the United Kingdom, Australia and the United States. *Addiction* **109**(5), 774–783 (2014)
8. Van Hout, M.C., Bingham, T.: 'Surfing the Silk Road': a study of users' experiences. *Int. J. Drug Policy* **24**(6), 524–529 (2013)
9. Demant, J., Munksgaard, R., Décary-Héту, D., Aldridge, J.: Going local on a global platform: a critical analysis of the transformative potential of cryptomarkets for organized illicit drug crime. *Int. Crim. Justice Rev.* **28**(3), 255–274 (2018)
10. Hardy, R.A., Norgaard, J.R.: Reputation in the Internet black market: an empirical and theoretical analysis of the Deep Web. *J. Inst. Econ.* **12**(3), 515–539 (2015)
11. Barratt, M.J., Lenton, S., Maddox, A., Allen, M.: 'What if you live on top of a bakery and you like cakes?'—drug use and harm trajectories before, during and after the emergence of silk road. *Int. J. Drug Policy* **35**, 50–57 (2016)
12. Demant, J., Bakken, S.A., Oksanen, A., Gunnlaugsson, H.: Drug dealing on Facebook, Snapchat and Instagram: a qualitative analysis of novel drug markets in the Nordic countries. *Drug Alcohol Rev.* **38**(4), 377–385 (2019)
13. Demant, J., Bakken, S.A.: Technology-facilitated drug dealing via social media in the Nordic countries. Background paper commissioned by the EMCDD. The European Monitoring Centre for Drugs and Drug Addiction, Lisbon (2019)
14. Baumeister, R.F., Heatherton, T.F.: Self-regulation failure: an overview. *Psychol. Inq.* **7**(1), 1–15 (1996)
15. Vazsonyi, A.T.M., Mikuška, J., Kelley, E.L.: It's time: a meta-analysis on the self-control-deviance link. *J. Crim. Justice* **48**, 48–63 (2017)
16. De Ridder, D.T., Lensvelt-Mulders, G.: Taking stock of self-control: a meta-analysis of how trait self-control relates to a wide range of behaviors. In: Baumeister, R.F. (ed.) *Self-regulation and Self-control*, pp. 221–274. Routledge, Abingdon (2018)
17. Ainslie, G.: *Breakdown of Will*. Cambridge University Press, Cambridge (2001)
18. Oksanen, A., Sirola, A., Savolainen, I., Kaakinen, M.: Gambling patterns and associated risk and protective factors among Finnish young people. *Nord. Stud. Alcohol Drugs* **36**(2), 161–176 (2019)
19. Orford, J.: *Excessive Appetites: A Psychological View of Addictions*. Wiley, West Sussex (2001)
20. Reynolds, B., Schiffbauer, R.: Delay of gratification and delay discounting: a unifying feedback model of delay-related impulsive behavior. *Psychol. Rec.* **55**(3), 439–460 (2005)
21. Green, L., Myerson, J.: A discounting framework for choice with delayed and probabilistic rewards. *Psychol. Bull.* **130**(5), 769–792 (2004)
22. Mittal, C., Griskevicius, V.: Sense of control under uncertainty depends on people's childhood environment: a life history theory approach. *J. Pers. Soc. Psychol.* **107**(4), 621–637 (2014)

23. Steinberg, L., Graham, S., O'Brien, L., Woolard, J., Cauffman, E., Banich, M.: Age differences in future orientation and delay discounting. *Child Dev.* **80**(1), 28–44 (2009)
24. Pratt, T.C., et al.: The empirical status of social learning theory: a meta-analysis. *Justice Q.* **27**(6), 765–802 (2019)
25. Kaakinen, M., Sirola, A., Savolainen, I., Oksanen, A.: Shared identity and shared information in social media: development and validation of the identity bubble reinforcement scale. *Media Psychol.* **23**(1), 23–51 (2020)
26. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. *Annu. Rev. Sociol.* **27**(1), 415–444 (2001)
27. Keipi, T., Näsi, M., Oksanen, A., Räsänen, P.: *Online Hate and Harmful Content: Cross-National Perspectives*. Routledge, Abingdon (2017)
28. Bahns, A.J., Pickett, K.M., Crandal, C.S.: Social ecology of similarity: big schools, small schools and social relationships. *Group Process. Intergr. Relat.* **15**(1), 119–131 (2011)
29. Oksanen, A., Hawdon, J., Räsänen, P.: Glamorizing rampage online: school shooting fan communities on YouTube. *Technol. Soc.* **39**, 55–67 (2014)
30. Oksanen, A., Näsi, M., Minkkinen, J., Keipi, T., Kaakinen, M., Räsänen, P.: Young people who access harm-advocating online content: a four-country survey. *Cyberpsychology J. Psychosoc. Res. Cyberspace* **10**(2) (2016). Article no. 6. <https://cyberpsychology.eu/article/view/6179/5909>
31. Sirola, A., Kaakinen, M., Oksanen, A.: Excessive gambling and online gambling communities. *J. Gambl. Stud.* **34**(4), 1313–1325 (2018)
32. Edlund, M.J., et al.: Opioid abuse and depression in adolescents: results from the national survey on drug use and health. *Drug Alcohol Depend.* **152**, 131–138 (2015)
33. Grant, B.F., et al.: Epidemiology of DSM-5 alcohol use disorder. *JAMA Psychiatry* **72**(8), 757–766 (2015)
34. Lai, H.M.X., Cleary, M., Sitharthan, T., Hunt, G.E.: Prevalence of comorbid substance use, anxiety and mood disorders in epidemiological surveys, 1990–2014: a systematic review and meta-analysis. *Drug Alcohol Depend.* **154**, 1–13 (2015)
35. Peters, E.N., et al.: Relationship of gambling with tobacco, alcohol, and illicit drug use among adolescents in the USA: review of the literature 2000–2014. *Am. J. Addict.* **24**(3), 206–216 (2015)
36. Fisoun, V., Floros, G., Siomos, K., Geroukalis, D., Navridis, K.: Internet addiction as an important predictor in early detection of adolescent drug use experience—implications for research and practice. *J. Addict. Med.* **6**(1), 77–84 (2012)
37. Oksanen, A.: Deleuze and the theory of addiction. *J. Psychoactive Drugs* **45**(1), 57–67 (2013)
38. Oksanen, A., Savolainen, I., Sirola, A., Kaakinen, M.: Problem gambling and psychological distress: a cross-national perspective on the mediating effect of consumer debt and debt problems among emerging adults. *Harm Reduction J.* **15**, 45 (2008). <https://doi.org/10.1186/s12954-018-0251-9>
39. Savolainen, I., Oksanen, A., Kaakinen, M., Sirola, A., Paek, H.J.: A three-country study on the role of perceived loneliness in youth addictive behaviors. *J. Med. Internet Res. Ment. Health* **7**(1), e14035 (2020)
40. Eysenck, S.B.G., Eysenck, H.J.: Impulsiveness and venturesomeness: their position in a dimensional system of personality description. *Psychol. Rep.* **43**, 1247–1255 (1978)
41. Bickel, W.K., Marsch, L.A.: Toward a behavioral economic understanding of drug dependence: delay discounting processes. *Addiction* **96**(1), 73–86 (2001)
42. Kaakinen, M., Keipi, T., Räsänen, P., Oksanen, A.: Cybercrime victimization and subjective well-being: an examination of buffering effect hypothesis among adolescents and young adults. *Cyberpsychol. Behav. Soc. Netw.* **21**(2), 129–137 (2018)

43. Banks, M.H., Clegg, C.W., Jackson, P.R., Kemp, N.J., Stafford, E.M., Wall, T.D.: The use of the general health questionnaire as an indicator of mental health in occupational studies. *J. Occup. Psychol.* **53**(3), 187–194 (1980)
44. Goldberg, D.P., et al.: The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychol. Med.* **27**, 191–197 (1997)
45. Minkkinen, J., et al.: Does social belonging to primary groups protect young people from the effects of pro-suicide sites? A comparative study of four countries. *Crisis J. Crisis Interv. Suicide Prev.* **37**(1), 31–41 (2016)
46. Bush, K., Kivlahan, D.R., McDonell, M.B., Fihn, S.D., Bradley, K.A.: The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. *Arch. Intern. Med.* **158**(16), 1789–1795 (1998)
47. Lesieur, H.R., Blume, S.B.: The South Oaks Gambling Screen (SOGS): a new instrument for the identification of pathological gamblers. *Am. J. Psychiatry* **144**(9), 1184–1188 (1987)
48. Meerkerk, G.J., van Den Eijnden, R.J., Vermulst, A.A., Garretsen, H.F.: The compulsive internet use scale (CIUS): some psychometric properties. *Cyberpsychol. Behav.* **12**(1), 1–6 (2009)
49. Firth, D.: Bias reduction of maximum likelihood estimates. *Biometrika* **80**(1), 27–38 (1993)
50. Greenland, S., Mansournia, M.A.: Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat. Med.* **34**(23), 3133–3143 (2015)
51. King, G., Zeng, L.: Logistic regression in rare events data. *Political Anal.* **9**(2), 137–163 (2001)
52. Farrington, D.P., West, D.J.: Criminal, penal and life histories of chronic offenders: risk and protective factors and early identification. *Crim. Behav. Ment. Health* **3**(4), 492–523 (1993)
53. Maddox, A., Barratt, M.J., Allen, M., Lenton, S.: Constructive activism in the dark web: cryptomarkets and illicit drugs in the digital ‘demimonde’. *Inf. Commun. Soc.* **19**(1), 111–126 (2016)
54. Oksanen, A., et al.: Pro-Anorexia and anti-pro-anorexia videos on YouTube: sentiment analysis of user responses. *J. Med. Internet Res.* **17**(11), e256 (2015)
55. Oksanen, A., Garcia, D., Räsänen, P.: Pro-anorexia communities on social media. *Pediatrics* **137**(1), e20153372 (2016)



# I Do It Because I Feel that...Moral Disengagement and Emotions in Cyberbullying and Cybervictimisation

Oronzo Parlange<sup>1</sup>(✉), Enrica Marchigiani<sup>1</sup>, Stefano Guidi<sup>1</sup>, Margherita Bracci<sup>1</sup>,  
Alessandro Andreadis<sup>2</sup>, and Riccardo Zambon<sup>2</sup>

<sup>1</sup> Department of Social, Political and Cognitive Sciences, University of Siena, Via Roma 56,  
Siena, Italy

{oronzo.parlangeli, enrica.marchigiani,  
margherita.bracci}@unisi.it, stefano.g73@gmail.com

<sup>2</sup> Department of Information Engineering and Mathematics, University of Siena, Via Roma 56,  
Siena, Italy

{alessandro.andreadis, riccardo.zambon}@unisi.it

**Abstract.** Few studies have jointly explored the role of factors such as the use of social media, the personality characteristics of young people, the use of thinking mechanisms aimed at moral disengagement, and the emotions experienced in relation to cyberbullying and cybervictimisation behaviour. The analysis presented here, carried out through a questionnaire distributed online and filled in by 655 Italian high school students, allowed to highlight the relationships between these variables. In particular, it emerged that the phenomena of cyberbullying and cybervictimisation are related to the time spent online and to the mechanisms of moral disengagement, which in turn are related to the personality trait of agreeableness. Emotions experienced are most clearly positive in cases of cyberbullying and negative for the victims. This correspondence, however, is reversed in bullies who resort more to thoughts aimed at moral disengagement and feel more negative emotions. The same reversal seems to occur in the victims who, in correspondence with an increased use of the mechanisms of moral disengagement, report to feel more positive emotions.

**Keywords:** Cyberbullying · Cybervictimisation · Emotions · Moral disengagement · Personality traits · Big five · Social media · Social network

## 1 Introduction

The data on the spread of cyberbullying is disturbing. Already in 2012 it was reported that in a representative sample of American adolescents, at least 10% reported having experienced cyberbullying [1]. These figures may vary slightly from country to country, or as the age of the respondents varies. For example, also in 2012, Scheneider and colleagues [2] reported a bullying rate of 6.4% in a sample of students. But if we focus on high school students these percentages can reach 21% and 30% [3, 4].

In Italy a study was recently conducted on a sample of high school students. Of these, 23% reported having posted or written embarrassing things to offend someone on the internet at least once [5].

The phenomenon of cyberbullying is not only relevant in view of its numbers. The main issues are related to the causes that determine it, the psychological aspects of who is responsible for it and the psychological consequences that this can cause in the victims. With good agreement among scholars, we generally speak of bullying when we witness acts of repeated aggression by one or more individuals, in those cases where there is a disparity between the power (any form of power) of those who carry out the aggression and the victim [6]. Traditional bullying acts do not differ substantially from those that take place online: in both cases the aggression may be verbal and involve forms of social victimization. The forms of physical aggression are clearly more referable to real contexts, although virtual contexts may also be antecedents of physical aggression. But what ultimately differentiates traditional bullying from cyberbullying is the means by which the latter takes place. Cyberbullying has in fact as its intrinsic expression the use of communication tools made available on the net - social media, instant messaging, e-mail - both to offend and to exclude [7].

In an attempt to fully understand the cyberbullying phenomenon, with the aim of trying to mitigate it, some issues still remain unclear. Several studies have found a clear relationship between actual and online bullying behaviour. More specifically, online aggressors, those who engage in victimization behaviour using virtual means, also show a tendency to engage in aggressive behaviour in reality [8–10]. This would lead to the consideration that there are stable factors, such as personality characteristics, at the origin of aggressive behaviour. Contrary to this assumption, however, Parlangeli and colleagues [5] found no difference, for instance, in the level of empathy (globally considered, but also in its components as emotional empathy and cognitive empathy) between those who reported aggressive behaviour and those who did not. However, other scholars [11], have more directly related traditional and online bullying behaviours considering personality traits in reference to the big five theory and constructs such as Machiavellianism, narcissism, and psychopathy (the so-called “dark triad”) as well as sadism [11, 12]. From the results of this analysis it would seem to emerge a difference between traditional bullying and cyberbullying because the former is related to agreeableness, Machiavellianism, psychopathy and sadism. The latter, instead, would be related only to the agreeableness trait and sadism [11].

Personality characteristics can however be expressed, or modulated, by cognitive styles, or by thought processes aimed at restructuring experiences in accordance with reference values. Bullying is a reprehensible act, therefore it is foreseeable that both the perpetrators of offensive acts and those who suffer them will undertake thought processes aimed at mitigating the negative self-image that follows these acts.

The mechanisms of thought that allow to loosen the strength of moral principles have long been studied in literature [13–15]. These essentially involve the attempt, by those who commit reprehensible actions, to formulate apparently rational justifications related to the causes that have produced the act. These include the thought that even if an action may seem reprehensible, it is instead carried out with reference to higher principles (for example, the safeguarding of the compactness of a team), or the consideration

of the victim's faults (a punishment must be given otherwise it could continue) or even minimizing the degree of aggression (it was just a game, nothing serious happened). In 2019 Parlange et al. [5] have reported that subjects who claim to have performed acts of cyberbullying differ from those who have never done so with reference to all moral disengagement mechanisms. In a former study Meter and Bauman [16] found that moral disengagement is related to cyberbullying and traditional bullying. This result is not surprising since other studies [17] have shown an overlap between these two behaviours. However, while cyberbullying is also related to cybervictimisation, i.e. many bullies are also victims [16] this does not happen in real bullying [18]. The use of moral disengagement mechanisms, in fact, can be a different phenomenon if we consider traditional moral disengagement in comparison with the online one. Paciello et al. [19] have put forward the hypothesis that traditional moral disengagement is context-independent while online moral disengagement is context-dependent. The results of their analysis confirm this hypothesis, highlighting how moral disengagement resulting from aggressive online behaviours is a separate theoretical construct, although correlated, with respect to the moral disengagement that manifests itself in contexts that can be defined traditional.

Recent studies highlight how mechanisms of moral disengagement are effective not only in bullies but also in bully/victims and victims, albeit in a different way. Bullies and bully/victims seem to use the same tendencies to blame the victim and to distort the consequences, even though the latter to a lesser extent. Differently, victims seem to be inclined to use the mechanisms of self-blaming and moral justification in order to justify being a victim of bullying [20, 21].

In the complex mechanism between moral disengagement, cyberbullying and cybervictimisation an additional component could be relative to the emotions experienced in perpetrating or suffering aggressive behaviour. Research highlights the importance of emotional skills in promoting positive individual and social youth development [22–27]. The research that has been conducted so far, however, has not produced truly conclusive results [28]. It is difficult to understand how, for example, there are no differences in emotions and levels of social competence between students who claim to have suffered acts of aggression online and those who claim never to have been victims [29]. More recently, Marìn-López et al. [30] reported that emotional competence can be a protective factor against the possibility of aggressive behaviour, perpetrated or suffered in virtual contexts. At the same time, however, again the same emotional and social skills seem to lead to greater exposure, more frequent and intense relationships also on the web. And this, in turn, leads to a greater risk of involvement in relationships that can be qualified as cyberbullying or cybervictimisation [30].

Also recently, Gül et al. [31] studied the relationship between cyberbullying and cybervictimisation in a sample of Turkish adolescents who had applied to the psychiatry outpatient units and therefore, as one might expect, their psychiatric symptom scores were high. Among these young people the declared percentages of cyberbullying (53.3%) and cybervictimisation (62.6%) were also particularly high. Their analysis showed that most of the boys and girls who claimed to be victims of cyberbullying were also cyberbullies. However, the variable that most clearly differentiates those who suffer and those who engage in violent behaviour on the net is the lack of emotional awareness. While having a good degree of emotional awareness seems to be about cyberbullies, victims

seem to be rather deficient in recognizing their emotional states. Gül et al. [31] have tried to explain this finding arguing that perhaps e-victims, due to their lack of emotional awareness, are more likely to be exposed to situations that may represent a threat. And also that these repeated exposures can lead to increased awareness, development of hostile attitudes, and thus to bullying. In this hypothesis, therefore, one can suppose a temporal sequence such that, first one is victimised and, subsequently and consequently to this, one can become a cyberbully.

## 2 The Study

Moral and emotional aspects together with personality traits constitute a complex interplay to be studied in order to shed some light on aggressive behaviours perpetrated or suffered by bullies and victims. In particular, this study aims to analyse the emotions and the moral mechanisms that come into play when committing or experiencing antisocial behaviour through the use of social media.

Both the aggressors and the victims, in fact, may activate some mechanisms that allow them to bridge the gap between the actions committed or suffered and the moral principles [32, 33], in particular through four kinds of cognitive mechanisms: justifying the behaviour, shifting responsibility, minimizing the harm caused, and moving the causal focus onto the victim [14]. Through these mechanisms, as well as the bullies try to reduce the emotional impact of their action, the victims endeavour to justify their inability to react and minimise the aggressiveness of persecutors [34, 35] perhaps even lowering the feeling of negative emotions.

Consequently, this study aims to explore the role of personality characteristics and mechanisms of moral disengagement, in connection with the emotions experienced, in reference to the behaviours perpetrated or suffered online.

### 2.1 Method

#### Participants and Procedure

The participants were 655 students. Most of respondents were in the 14–17 age range (68.7%), and the remaining ones had between 18 and 20 years (29.3%) or between 21 and 25 years (2%). Female participants were 62.3%.

The sample was recruited from 16 Italian high schools, technical and vocational Institutes scattered in the Italian territory from north to south, through direct contact with the schools and their headmasters, upon request for authorisation and information on data processing.

The data were gathered from a structured self-reported questionnaire, to be filled in online, consisting of 6 sections respectively devoted to collect data on personal information, social media use, personality traits [36], moral disengagement about cyberbullying, cyber victimization and cyberbullying behaviours [16], and the emotions felt by the cyber-bullies and the cyber-victims, collected through the Positive and Negative affective schedule (PANAS) [37].

The department, that in our case carried out the function of ethical committee, evaluated and approved the study (report n. 10/2019 of 13 March 2019).

The first section of the questionnaire, Personal information, was aimed at collecting information on age, gender and school attended by the participants.

The use of social network was investigated in the second section through questions on the social media most frequently used, the reason for their use, the time spent on the net.

## Measures

### *Big Five Inventory*

Personality traits have been investigated through the Big Five Inventory (BFI) [38]. In this study the shorter 10 items version of the BFI was used [36]. It measures five dimensions, with two items for each one, considered fundamental in the personality description: extraversion, agreeableness, conscientiousness, neuroticism and openness. For each item the individuals were asked to express their level of agreement on a Likert-type scale from 1 “do not agree” to 5 “completely agree”.

Extraversion refers to a dynamic, active and dominant mode of behaviour and it is associated to high positive affect. Agreeableness describes a friendly, cordial and selfless attitude, it was associated with higher positive and lower negative affect. Conscientiousness measures the presence of reflexivity, scrupulousness and accuracy that can become fussiness and excessive attention to detail. It predicted low affect variability and low reactivity to stressors [39]. Neuroticism concerns characteristics related to anxiety, vulnerability and control of one’s behavioural and emotional reactions and is associated to higher negative and lower positive affect. It seems to be the personality factor most strongly related to emotions and affect in general [36, 40–43]. Openness describes people with several interests and open to new experiences. It seems to be related to higher reactivity to stressful events but has no variability on affect [39].

### *Moral Disengagement About Cyberbullying*

Moral disengagement about cyberbullying is an eight items self-report scale [15, 16] that has been structured to analyse moral disengagement from cyberbullying behaviours through four kinds of mechanisms: justifying the behaviour, shifting responsibility, minimizing the harm caused, and moving the causal focus onto the victim [14].

The questionnaire includes both items directly addressing moral disengagement of aggressive online behaviour, such as “Cyberbullying annoying classmates is just teaching them a lesson” and more general questions, such as “It’s okay to treat someone badly if they behave like a jerk”.

For each item the participants were asked to express their level of agreement on a Likert-type scale from 1 “strongly disagree” to 5 “strongly agree” [16].

### *Cybervictimization and Cyberbullying*

The scale includes 12 items, 6 for cybervictimisation and 6 for cyberbullying.

All the items asked participants to report the frequency (1 = never, 2 = 1–2 times, 3 = 3–5 times, 4 = 5 + times) with which respondents were harmed by others’ cyber-aggressors or with which respondents harmed others by sending mean or threatening messages or emails.



By way of example, the following is an item that identify the adolescents who consider themselves to be victims of cyberbullying “How often have you received mean or scary emails?” [16].

*Positive and Negative Affect Schedule (PANAS Short Scale) About Cyberbullies and Cybervictims*

Positive and Negative Affect Schedule measures the two paramount and culturally dominant aspects of an emotional experience, the positive or negative valence. High positive affect indicates activation, energy, enthusiasm, while a low positive affect stands for inaction, sadness and inattentivity [44].

Positive affect includes mood states such as interested, active, attentive, enthusiastic, while negative affect can go from disgust to fear to anger to guilt. According to Tellegen [45] and Watson and Clark [44] positive affect and negative affect are related to the personality factors of extraversion and neuroticism, respectively.

The PANAS scale consists of 20 terms that describe emotions and feelings, 10 positive, such as active, alert, attentive, determined, enthusiastic, excited, inspired, interested, proud, strong and 10 negative, such as afraid, scared, nervous, jittery, irritable, hostile, guilty, ashamed, upset, distressed [37].

For each emotion a 5-point Likert scale is given from 1 = Very slightly or not at all and 5 = extremely. The positive emotion scores are added together, for a range from 10 to 50. In the same way, negative emotions are added up. In this study, The PANAS schedule was used both to evaluate the emotions felt by the attackers and to evaluate the emotions felt by the victims.

## 2.2 Results

### Social Media Use

The majority of the respondents declared that they spend between 1 and 3 h per day using social media (54.4%). About 30% reported more than 3 h per day spent on social media, and 13.3% less than one hour. Only 5 participants (0.8%) declared they do not use social media at all. Time spent on social media was not significantly associated with personality traits, or with age group, but it was strongly associated with gender ( $X^2(3) = 27.3$ ;  $p < .001$ ), and the results of the analyses showed that the main difference between male and females was in the group that reported to spend more than 3 h per day on social media that comprised 38.7% of the females respondent versus 19.8% of the males respondents (cfr Table 1).

### Moral Disengagement

We examined first the fit of the measurement model for moral disengagement using confirmatory factor analysis (CFA). The results showed that fit of the model to the data was good [CFI = 0.938; TLI = 0.913, SRMR = 0.045; RMSEA = 0.070], and for further analyses we computed and stored the predicted respondents' scores on moral disengagement. The distribution of these scores was highly skewed to the right, and the analysis of the correlations with the BFI personality trait scores, adjusting for multiple comparisons, showed a significant, weak correlation between agreeableness and Moral

**Table 1.** Contingency table confronting the distribution of responses about the time spend on social media in male and females.

| <i>Time on Social Media</i> | <i>Gender</i> |        | <i>Total</i> |
|-----------------------------|---------------|--------|--------------|
|                             | F             | M      |              |
| Never                       | 2             | 3      | 5            |
|                             | 40 %          | 60 %   | 100 %        |
|                             | 0.5 %         | 1.2 %  | 0.8 %        |
| Less than 1 hour/day        | 44            | 43     | 87           |
|                             | 50.6 %        | 49.4 % | 100 %        |
|                             | 10.8 %        | 17.4 % | 13.3 %       |
| 1 to 3 hours/day            | 204           | 152    | 356          |
|                             | 57.3 %        | 42.7 % | 100 %        |
|                             | 50 %          | 61.5 % | 54.4 %       |
| More than 3 hours/day       | 158           | 49     | 207          |
|                             | 76.3 %        | 23.7 % | 100 %        |
|                             | 38.7 %        | 19.8 % | 31.6 %       |
| <b>Total</b>                | 408           | 247    | 655          |
|                             | 62.3 %        | 37.7 % | 100 %        |
|                             | 100 %         | 100 %  | 100 %        |

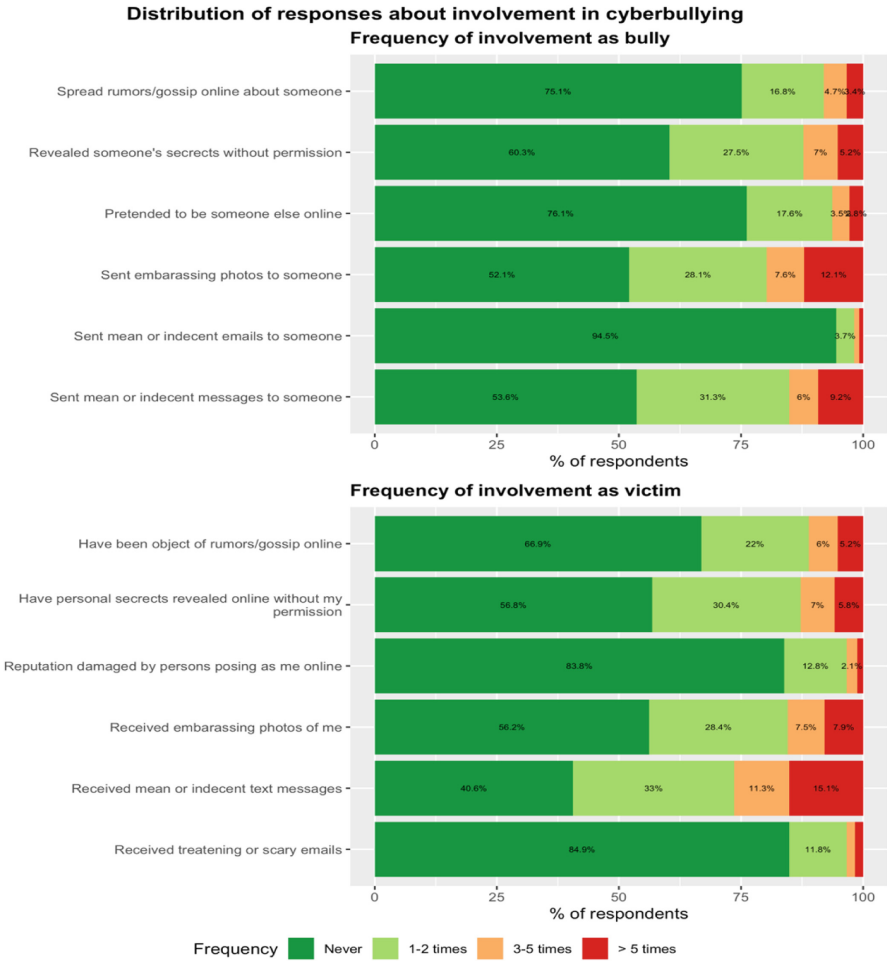
$$\chi^2=27.277 \cdot df=3 \cdot \text{Cramer's } V=0.204 \cdot \text{Fisher's } p=0.000$$

Disengagement ( $r = 0.12$ ;  $p < .05$ ). We also examined the association of the MD score with time spent on social media, age and gender using multiple linear regression. The analysis was conducted on the square root of the competed MD score, because the distribution of the original scores deviated significantly from normality. The results showed that moral disengagement was significantly associated with gender, with higher scores for males than females [ $F(1, 636) = 72.91$ ;  $p < .001$ ], but not with age or time on social media, and overall the model explained approximately 11% of the variability in the transformed MD scores.

**Involvement in Cyberbullying**

We initially examined the frequencies of the responses to the items of the cyberbullying involvement questionnaire [16]. The distribution of the responses is shown in the Fig. 1 as stacked frequencies bar charts.

We then conducted a confirmatory factor analysis to examine the measurement model for cyberbullying involvement, including two latent factors (cyberbullying and cyber-victimisation) each manifested by six ordinal indicators. The fit of the model was good [CFI = 0.956; TLI = 0.946; GFI = 0.982; SRMR = 0.089; RMSEA = 0.077]. Predicted scores for both factors were computed and saved for further analyses. For both factors the computed scores were significantly associated with time spent on social media [Cyberbullying:  $F(2, 638) = 11.74$ ;  $p < .001$ ; Cybervictimisation:  $F(2, 638) = 8.08$ ;  $p < .001$ ;], but not associated with Age and Gender, and the coefficients showed that the



**Fig. 1.** Stacked frequency bar charts showing the distribution of participants' responses to the 12 items of the questionnaire about involvement in cyberbullying, either as bully (above) or as victim (below).

more time the respondents spend on social media, the higher were the cyberbullying and cybervictimisation scores.

Sixty-four participants (9.8%) declared that they had been a bully (by posting photos, embarrassing images, offending comments/words, or revealing online someone else's secrets or personal/private information for others to read) more than once. One hundred-forty-four (22%) instead affirmed they had been victims of such acts of bullying. There was a significant association between having bullied and having been a victim of bullying [ $\chi^2(1) = 10.98; p < .001$ ]. Among those who declared to have been bullied, the frequency

of bullies (17.4%;  $n = 25$ ) was more than twice the one found among those who had not been bullied (7.6%;  $n = 39$ ).

### Emotions in Cyberbullying Involvement

We then examined the correlations between the positive and negative emotions felt by respondents in connections with either having bullied one of their school mates or having been the victim of such cyberbullying acts from school mates.

In the tables below are reported the mean scores for the negative and positive emotions as function of a) gender, b) having been a bully and c) having been victim of bullying.

#### *Emotions by Gender*

Female victims felt significantly stronger negative emotions than male victims [ $t(69.67) = 3.06$ ,  $p < .01$ ,  $d = 0.59$ ], but apart from this, there were no other significant differences between emotions by gender (see Table 2).

**Table 2.** Average scores for the intensity of positive and negative emotions after involvement in acts of cyberbullying either as bully or as victim of bullying, overall and as function of gender.

|                          | Sample |      | Females |      | Males |      | p-value |
|--------------------------|--------|------|---------|------|-------|------|---------|
|                          | Mean   | SD   | Mean    | SD   | Mean  | SD   |         |
| Pos. emotions bullying   | 2.80   | 0.94 | 2.66    | 0.95 | 2.97  | 0.93 | NS      |
| Neg. emotions bullying   | 2.17   | 1.01 | 2.23    | 1.05 | 2.1   | 0.97 | NS      |
| Positive emotions victim | 2.30   | 0.94 | 2.24    | 0.92 | 2.45  | 0.99 | NS      |
| Negative emotions victim | 3.01   | 1.06 | 3.18    | 1.02 | 2.59  | 1.05 | <.01    |

#### *Emotions in Bullies and Victims*

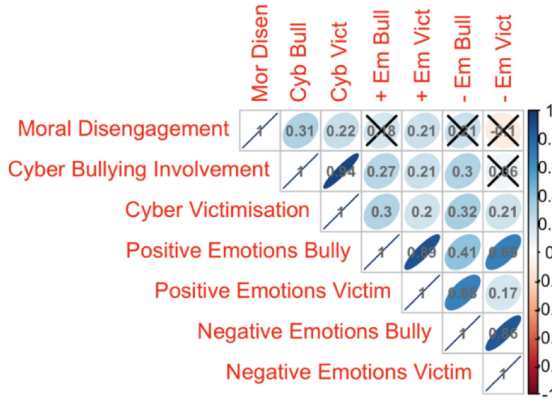
For those who reported to have performed acts of bullying, emotions scores were significantly higher for positive emotions than for negative ones [ $t(62) = 4.7$ ,  $p < .001$ ,  $d = 0.63$ ]. Conversely, for those who reported to have been victim of acts of bullying, emotions scores were significantly higher for negative emotions than for positive ones [ $t(142) = -6.59$ ,  $p < .001$ ,  $d = -0.71$ ] (see Table 3).

**Table 3.** Average scores for the intensity of positive and negative emotions after involvement in acts of cyberbullying either as bully or as victim of bullying.

| Role   | N   | Positive emotions |      | Negative emotions |      | p-value |
|--------|-----|-------------------|------|-------------------|------|---------|
|        |     | Mean              | SD   | Mean              | SD   |         |
| Bully  | 63  | 2.8               | 0.94 | 2.17              | 1.01 | <.001   |
| Victim | 143 | 2.3               | 0.94 | 3.01              | 1.06 | <.001   |

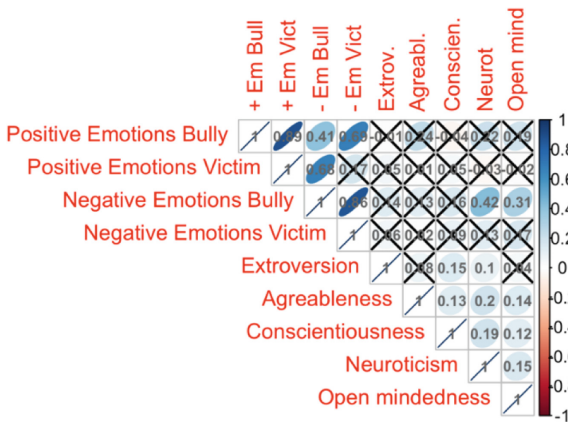
*Correlations Between Emotions and Other Variables*

For both groups (bullies and victims), a positive significant correlation was found between the PANAS negative and positive emotion scales (Fig. 2). For bullies, the magnitude of the correlation was moderate [ $r = 0.41$ ;  $t(61) = 3.48$ ;  $p < .001$ ], while for victims the emotions scales were weakly correlated [ $r = 0.17$ ;  $t(141) = 2.09$ ;  $p < .05$ ].



**Fig. 2.** Correlations between moral disengagement, involvement in cyberbullying as victims or bullies, and positive/negative emotions experienced after the acts. The Xs mark insignificant correlations after adjusting for multiple comparisons.

Negative emotions during acts of bullying were also significantly correlated with neuroticism ( $r = 0.41$ ) and openness ( $r = 0.31$ ) (Fig. 3).

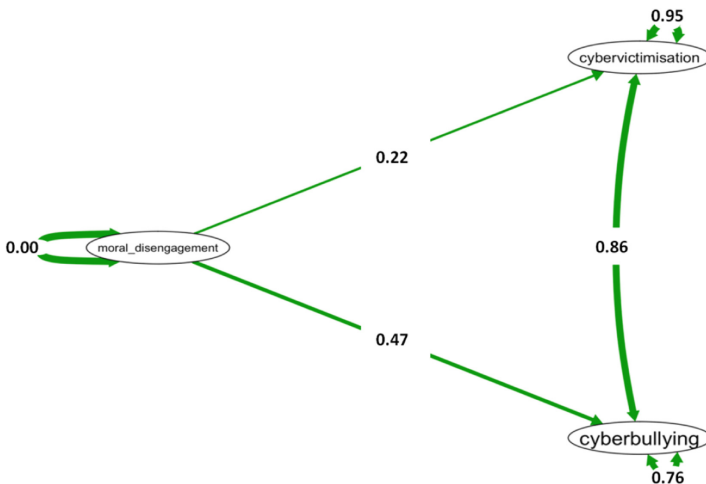


**Fig. 3.** Correlations between personality traits and positive/negative emotions in bullies and victims of cyberbullying acts. The Xs mark insignificant correlations after adjusting for multiple comparisons.

**Structural Models**

We used structural equation modelling to test hypothesis about the association between moral disengagement and involvement in cyberbullying. First, we tested a model derived by Meter and Bauman [16], in which moral disengagement is directly affecting cyberbullying involvement as a bully or as a victim. The model had a good fit to the data [CFI = 0.951; TLI = 0.945; GFI = 0.98; SRMR = 0.075; RMSEA = 0.058] and the results confirmed previous findings that moral disengagement was indeed significantly associated with both cyberbullying involvement factors (Fig. 4). We then tested a second model in which the binary endogenous variables coding whether participants had (vs had not) more than once posted offending/embarrassing pictures/words or personal information about a school mate so the others could read them, and whether they had been more than once the victim of such actions were included outcomes of moral disengagement. The results showed again a good fit of the model to the data [CFI = 0.959; TLI = 0.953; GFI = 0.98; SRMR = 0.071; RMSEA = 0.052], and both the direct effects of moral disengagement on the new binary outcomes were significant.

Finally, we conducted separate regression analyses on the positive and negative emotions felt by bullies and victims, to test the association of these emotions with moral disengagement, gender and personality traits. For bullies, positive emotions were not associated with any of the factors considered, while negative emotions were significantly associated to moral disengagement [B = 0.37; se = 0.15; p < .05], neuroticism [B = 0.59; se = 0.19; p < .01] and openness [B = 0.41; se = 0.16; p < .05]. For the victims, conversely, positive emotions were significantly associated with moral disengagement [B = 0.35; se = 0.14; p < .05], while negative emotions were not, but were associated with gender [B<sub>males</sub> = -0.65; se = 0.2; p < .01] and openness [B = 0.27; se = 0.14; p < .05].



**Fig. 4.** Structural equation model of the relationships between moral disengagement and involvement in cyberbullying, as a bully and as a victim.

### 3 Conclusion

The results of this study have provided additional knowledge useful for the understanding of complex phenomena such as those related to attacks perpetrated or suffered online.

The model of Meter and Bauman [16] was confirmed, highlighting how moral disengagement directly affects involvement in experiences as both bully and victim. In addition, a second model was tested which explored the hypothesis that moral disengagement may or may not lead to bully or victim involvement, and the results are significant in both cases. Clearly, it was possible to reaffirm that the mechanisms of moral disengagement can come into play as much in reference to being a bully as to being a victim, and probably can also be a precondition for those roles. Moreover, it was highlighted that this has to do with the gender of the respondents - males exercise more thought mechanisms aimed at moral disengagement - and not with time spent online. The latter variable, however, as a contextual factor is in any case related to the occurrence of episodes of cyberbullying or cybervictimisation and can lead to positive but also negative consequences [46].

With regard to emotional correlates, it should be noted that this study did not analyse specific emotions, but their direction (positive or negative) and their intensity. The results suggest that there is a subjective disposition to feel emotions with more or less intensity: both bullies and victims experience emotions that are more or less intense in a correlated way, i.e. if one feels positive intense emotions one also feels negative intense emotions.

Exploring the relationships between aggressive online behaviour and emotions, on the one hand we have a clear and predictable relationship between cyberbullying behaviour and positive emotions on the other hand between victimisation and negative emotions. But the use of moral disengagement thought mechanisms seems to have to do with a completely reversed emotional pattern. Bullies (regardless of gender) who show higher levels of moral disengagement feel negative emotions more intensely and are more characterized by the trait of neuroticism. In this regard, it can be hypothesized that the behaviour manifested by these subjects is likely to be ascribed to the inability to control aggressive impulses adequately (something that most clearly pertains to boys, see Parlangeli et al. [47]). This, in connection with a tendency to experience the world as hostile and threatening, may imply a greater need to resort to mechanisms of moral disengagement as a mean, in this case an ineffective mean, to reduce guilt and negative feelings following their actions.

Opposite results are found in relation to victims who have higher levels of moral disengagement in relation to positive emotions. Some authors [34, 35] have suggested that the use of mechanisms of moral disengagement by victims is due to reflective thought processes resulting from the events of aggression suffered. As if to say that cybervictimised young people could be looking for explanations that justify their inability to react, and therefore try to minimize emotions of self-blame. The results obtained with this study perhaps add a piece to this explanation, as they may suggest that victims must have already suffered aggressive behaviour in order to formulate thoughts of moral disengagement. This process must correspond to a psychological state that does not involve excessively negative emotions, so negative that they are annihilating and lead to a complete assumption of responsibility for what e-victims have suffered. Contrary to

this, recourse to mechanisms of moral disengagement can lead to the establishment of positive emotions for the victims.

Obviously, the explanations given are subsequent to the evidence of the data, since the study was carried out in an exploratory manner and no hypothesis had been made as to the direction of the relations between the variables considered. In the light of the results found, it seems however reasonable to think that their interpretation must take into account both the subjective dispositions to feel more or less intense emotions and the personality traits of the subjects. Thus, for both victims and bullies the relationship between negative emotions and neuroticism may appear quite reasonable. Less obvious, and certainly in need of further investigation, is the result that shows a relationship of the openness trait to the negative emotions of bullies.

In any case, both with regard to the relationship between moral disengagement and negative emotions in bullies, and between moral disengagement and positive emotions in victims, a more precise analysis of individual emotions would be necessary. PANAS [37] is not structured by contrasting negative emotions with positive corresponding emotions (e.g. sadness vs. joy). Moreover, it refers both to purely subjective emotions (e.g., active), and to others more of a social (e.g., hostile), or moral (i.e., ashamed) type. These differences between emotions, here instead considered and generically qualified as positive or negative, are probably not irrelevant if one thinks mechanisms such as moral disengagement in cyberbullying and cybervictimisation behaviours. Unfortunately, however, one limitation of the present study concerns the number of respondents to the questionnaire who claimed to be bullies or victims, a number that did not allow us to conduct analysis at a higher level of detail in relation to single emotions.

Another limitation is the lack of control of the level of sincerity of the respondents. The issues addressed are obviously characterized by a negative social value, and it is very likely that the answers obtained are affected by a desirability effect [48].

Further studies will have to explore with greater accuracy the different emotions in both cyberbullying and cybervictimisation behaviours, and their role as consequential, antecedent or more simply as correlates of the mechanisms of moral disengagement. From this could derive knowledge useful for a “design for reflection”, a social media design aimed both at a greater knowledge of the media themselves [49, 50] and a greater awareness of the relationships that are realized through them.

## References

1. Ybarra, M., Boyd, D., Korchmaros, J., Oppenheim, J.: Definition and measuring cyberbullying within the larger context of bullying victimization. *J. Adolesc. Health* **51**, 53–58 (2012). <https://doi.org/10.1016/j.jadohealth.2011.12.031>
2. Schneider, S.K., O'donnell, L., Stueve, A., Coulter, R.W.: Cyberbullying, school bullying, and psychological distress: a regional census of high school students. *Am. J. Public Health* **102**(1), 171–177 (2012). <https://doi.org/10.2105/AJPH.2011.300308>
3. Beran, T.N., Li, Q.: Cyber-Harassment: a study of a new method for an old behavior. *J Educ. Comput. Res.* **32**(3), 265–277 (2005). <https://doi.org/10.2190/8YQM-B04H-PG4D-BLLH>
4. Calvete, E., Orue, I., Estévez, A., Villardón, L., Padilla, P.: Cyberbullying in adolescents: modalities and aggressors' profile. *Comput. Hum. Behav.* **26**, 1128–1135 (2010). <https://doi.org/10.1016/j.chb.2010.03.017>



5. Parlangeli, O., Marchigiani, E., Bracci, M., Duguid, A.M., Palmitesta, P., Marti, P.: Offensive acts and helping behavior on the internet: an analysis of the relationships between moral disengagement, empathy and use of social media in a sample of Italian students. *Work* **63**(3), 469–477 (2019). <https://doi.org/10.3233/WOR-192935>
6. U.S. Department of Health and Human Services - Stopbullying.gov, Effects of bullying (2014). <https://www.stopbullying.gov/bullying/effects>
7. Menesini, E., Spiel, C.: Introduction: Cyberbullying: Development, consequences, risk and protective factors. *Eur. J. Dev. Psychol.* **9**, 163–167 (2012)
8. Dempsey, A.G., Sulkowski, M.L., Dempsey, J., Storch, E.A.: Has cybertechnology produced a new group of peer aggressors? *Cyberpsych. Beh. Soc. N.* **14**(5), 297–302 (2011)
9. Olweus, D.: Cyberbullying: an overrated phenomenon? *Eur. J. Dev. Psychol.* **9**, 520–538 (2012). <https://doi.org/10.1080/17405629.2012.682358>
10. Perren, S., Gutzwiller-Helfenfinger, E.: Cyberbullying and traditional bullying in adolescence: differential roles of moral disengagement, moral emotions and moral values. *Eur. J. Dev. Psychol.* **9**(2), 195–209 (2012). <https://doi.org/10.1080/17405629.2011.643168>
11. van Geel, M., Goemans, A., Toprak, F., Vedder, P.: Which personality traits are related to traditional bullying and cyberbullying. *Pers. Individ. Differ.* **106**, 231–235 (2017). <https://doi.org/10.1016/j.paid.2016.10.063>
12. Goodboy, A.K., Martin, M.M.: The personality profile of a cyberbully: examining the Dark Triad. *Comput. Hum. Behav.* **49**, 1–4 (2015). <https://doi.org/10.1016/j.chb.2015.02.052>
13. Bandura, A.: Moral disengagement in the perpetration of inhumanities. *J. Pers. Soc. Psychol.* **3**, 193–209 (1999)
14. Bandura, A., Barbaranelli, C., Caprara, G.V., Pastorelli, C.: Mechanism of moral disengagement in the exercise of moral agency. *J. Pers. Soc. Psychol.* **71**, 364–374 (1996)
15. Bussey, K., Fitzpatrick, S., Raman, A.: The role of moral disengagement and self-efficacy in cyberbullying. *J. Sch. Violence* **14**, 30–46 (2015). <https://doi.org/10.1080/15388220.2014.954045>
16. Meter, D.J., Bauman, S.: Moral disengagement about cyberbullying and parental monitoring: effects on traditional bullying and victimization via cyberbullying involvement. *J. Early Adolesc.* **38**, 303–326 (2018). <https://doi.org/10.1177/02724316166670752>
17. Monks, C.P., Robinson, S., Worlidge, P.: The emergence of cyberbullying: a survey of primary school pupils' perceptions and experiences. *School Psychol. Int.* **33**(5), 477–491 (2012). <https://doi.org/10.1177/0143034312445242>
18. Haynie, D.L., et al.: Bullies, victims, and bully/victims: distinct groups of at-risk youth. *J. Early Adolesc.* **21**, 29–49 (2001). <https://doi.org/10.1177/0272431601021001002>
19. Paciello, M., Tramontano, C., Nocentini, A., Fida, R., Menesini, E.: The role of traditional and online moral disengagement on cyberbullying: do externalising problems make any difference? *Comput. Hum. Behav.* **9**, 190–198 (2020)
20. Schacter, H.L., White, S.J., Chang, V.Y., Juvonen, J.: “Why me?": characterological self-blame and continued victimization in the first year of middle school. *J. Clin. Child Adolesc.* **44**(3), 446–455 (2015). <https://doi.org/10.1080/15374416.2013.865194>
21. Runions, K.C., Shaw, T., Bussey, K., Thornberg, R., Salmivalli, C., Cross, D.S.: Moral disengagement of pure bullies and bully/victims: shared and distinct mechanisms. *J. Youth Adolesc.* **48**(9), 1835–1848 (2019). <https://doi.org/10.1007/s10964-019-01067-2>
22. Ttofi, M.M., Farrington, D.: Effectiveness of school-based programs to reduce bullying: a systematic and meta-analytic review. *J. Exp. Criminol.* **7**, 27–56 (2010)
23. Zych, I., Ortega-Ruiz, R., Del Rey, R.: Scientific research on bullying and cyberbullying: where have we been and where are we going. *Aggress. Violent Behav.* **24**, 188–198 (2015). <https://doi.org/10.1016/j.avb.2015.05.015>

24. Schokman, C., Downey, L.A., Lomas, J., Wellham, D., Wheaton, A., Simmons, N., et al.: Emotional intelligence, victimisation, bullying behaviours and attitudes. *Learn Individ. Differ.* **36**, 194–200 (2014)
25. Elipe, P., Mora-Merchán, J.A., Ortega-Ruiz, R., Casas, J.A.: Perceived emotional intelligence as a moderator variable between cybervictimization and its emotional impact. *Front. Psychol.* **6**, 486 (2015). <https://doi.org/10.3389/fpsyg.2015.00486>
26. Marikuttu, P.J., Joseph, M.I.: Effects of emotional intelligence on stress, psychological well-being, and academic achievement of adolescents. *Indian J. Health Wellbeing* **7**, 699–702 (2016)
27. Parlangei, O., Bracci, M., Duguid, A.M., Marchigiani, E., Palmitesta, P.: Cyberbullying and prosocial behavior on the net: the influence of some socio-cognitive and contextual factors in a sample of italian high school students. In: Proceedings of the 11th ICERI International Conference of Education Research and Innovation, Seville, Spain, pp. 8982–8987 (2018). <https://doi.org/10.21125/iceri.2018>
28. Zych, I., Farrington, D., Llorent, V.J., Ttofi, M.M.: *Protecting Children Against Bullying and its Consequences*. Springer, New York (2017). <https://doi.org/10.1007/978-3-319-53028-4>
29. Gomez-Ortiz, O., Romera-Felix, E.M., Ortega-Ruiz, R.: Multidimensionality of social competence: measurement of the construct and its relationship with bullying roles. *Revista de Psicodidactica (English ed.)* **22**(1), 37–44 (2017)
30. Marín-López, I., Zych, I., Ortega-Ruiz, R., Hunter, S.C., Llorent, V.J.: Relations among online emotional content use, social and emotional competencies and cyberbullying. *Child Youth Serv. Rev.* **108**, 104647 (2020). <https://doi.org/10.1016/j.childyouth.2019.104647>
31. Gül, H., Fırat, S., Sertçelik, M., Gül, A., Gürel, Y., Kılıç, B.G.: Cyberbullying among a clinical adolescent sample in Turkey: effects of problematic smartphone use, psychiatric symptoms, and emotion regulation difficulties. *Psychiat. Clin. Psych.* **29**(4), 1–11 (2018)
32. Cuadrado-Gordillo, I., Fernández-Antelo, I.: Analysis of moral disengagement as a moderating factor in adolescents' perception of cyberbullying. *Front. Psychol.* **10**, 1222 (2019). <https://doi.org/10.3389/fpsyg.2019.01222>
33. Schoeps, K., Villanueva, L., Prado-Gascó, V.J., Montoya-Castilla, I.: Development of emotional skills in adolescents to prevent cyberbullying and improve subjective well-being. *Front. Psychol.* **9**, 2050 (2018). <https://doi.org/10.3389/fpsyg.2018.02050>
34. Allison, K.R., Bussey, K.: Individual and collective moral influences on intervention in cyberbullying. *Comput. Hum. Behav.* **74**, 7–15 (2017). <https://doi.org/10.1016/j.chb.2017.04.019>
35. Luo, A., Bussey, K.: The selectivity of moral disengagement in defenders of cyberbullying: contextual moral disengagement. *Comput. Hum. Behav.* **93**, 318–325 (2019). <https://doi.org/10.1016/j.chb.2018.12.038>
36. Rammstedt, B., John, O.P.: Measuring personality in one minute or less: a 10 item short version of the big five inventory in English and German. *J. Res. Pers.* **41**, 203–212 (2007). <https://doi.org/10.1016/j.jrp.2006.02.001>
37. Watson, D., Clark, L.A., Tellegan, A.: Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* **54**(6), 1063–1070 (1988). <https://doi.org/10.1037//0022-3514.54.6.1063>
38. John, O.P., Donahue, E.M., Kentle, R.L.: *The Big Five Inventory-Versions 4a and 54*. University of California, Berkeley, Institute of Personality and Social Research, Berkeley (1991)
39. Komulainen, E., Meskanen, K., Lipsanen, J., Lahti, J.M., Jylhä, P., Melartin, T., et al.: The effect of personality on daily life emotional processes. *PLoS ONE* **9**(10), e110907 (2014). <https://doi.org/10.1371/journal.pone.0110907>

40. Terracciano, A., McCrae, R.R., Costa Jr., P.T.: Factorial and construct validity of the Italian positive and negative affect schedule (PANAS). *Eur. J. Psychol. Assess.* **19**(2), 131–141 (2003). <https://doi.org/10.1027//1015-5759.19.2.131>
41. Caprara, G., Barbaranelli, C., Borgogni, L.: *BFQ - Big Five Questionnaire. Manuale. Organizzazioni Speciali, Firenze* (1993)
42. Gosling, S.D., Rentfrow, P.J., Swann Jr., W.B.: A very brief measure of the Big-Five personality domains. *J. Res. Pers.* **37**, 504–528 (2003). [https://doi.org/10.1016/S0092-6566\(03\)00046-1](https://doi.org/10.1016/S0092-6566(03)00046-1)
43. Guido, G., Peluso, A.M., Capestro, M., Miglietta, M.: An Italian version of the 10-item Big Five Inventory: an application to hedonic and utilitarian shopping values. *Pers. Individ. Differ.* **76**, 135–140 (2015). <https://doi.org/10.1016/j.paid.2014.11.053>
44. Watson, D., Clark, L.A.: Negative affectivity: the disposition to experience aversive emotional states. *Psychol. Bull.* **96**(3), 465–490 (1984). <https://doi.org/10.1037/0033-2909.96.3.465>
45. Tellegen, A.: Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In: Tuma, A.H., Maser, J.D. (eds.) *Anxiety and the Anxiety Disorders*, pp. 681–706. Erlbaum, Hillsdale (1985)
46. Bracci, M., Duguid, A.M., Marchigiani, E., Palmitesta, P., Parlangeli, O.: Digital discrimination: an ergonomic approach to emotional education for the prevention of cyberbullying. In: Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y. (eds.) *IEA 2018. AISC*, vol. 826, pp. 723–731. Springer, Cham (2019). [https://doi.org/10.1007/978-3-319-96065-4\\_76](https://doi.org/10.1007/978-3-319-96065-4_76)
47. Parlangeli, O., Bracci, M., Guidi, S., Marchigiani, E., Duguid, A.M.: Risk perception and emotions regulation strategies in driving behaviour: an analysis of the self-reported data of adolescents and young adults. *Int. J. Hum. Fact. Erg.* **5**(2), 166–187 (2018). <https://doi.org/10.1504/IJHFE.2018.092242>
48. Dalton, D., Ortegren, M.: Gender differences in ethics research: the importance of controlling for the social desirability response bias. *J. Bus. Ethics* **103**(1), 73–93 (2011). <https://doi.org/10.1007/s10551-011-0843-8>
49. Parlangeli, O., Mengoni, G., Guidi, S.: The effect of system usability and multitasking activities in distance learning. In: *Proceedings of the CHIItaly Conference*, 13–16 September, pp. 59–64, Alghero. ACM Library (2011). <https://doi.org/10.1145/2037296.2037314>
50. Parlangeli, O., Guidi, S., Fiore Farina, R.: Overloading disks onto a mind: quantity effects in the attribution of mental states to technological systems. In: *Proceedings of the 18th International Congress, IEA – International Ergonomics Association, Recife*, 12–16 February 2012 (2012)



# The Effects of Thinking Styles and News Domain on Fake News Recognition by Social Media Users: Evidence from Russia

Alexander Porshnev<sup>1</sup>  and Alexandre Miltsov<sup>2</sup> 

<sup>1</sup> National Research University Higher School of Economics, St. Petersburg, Russia  
aporshnev@hse.ru

<sup>2</sup> Nazarbayev University, Nur-Sultan, Kazakhstan

**Abstract.** The development and deployment of new technologies have influenced the media environment by enabling quick and effective dissemination of false news via social networks. Several experimental studies have highlighted the role of thinking style, social influence, source credibility and other factors when it comes to fake news recognition. Our study makes several contributions to existing knowledge. We introduce a measure of conspiracy thinking, a comparison between politics and business news recognition, and we investigate the effects of sensationalist headlines on users' abilities to differentiate between false and true news. 228 university students (203 completed the entire survey) from three departments (Humanities, Management, and Economics) took part in an online experiment. The results of a regression analysis demonstrate that double-checking of news online has a significant effect on individuals' overall ability of differentiating between true and false news. Thinking styles, prior experience, and such control variables as age and gender have no significant effect on the overall level of accuracy. We also discuss the effects of different factors responsible for the accuracy of fake news recognition in business and political news, as well as several limitations of the study.

**Keywords:** Fake news · Conspiracy thinking · Rational mentality · Magical beliefs · Trust

## 1 Introduction

The development and deployment of new technologies have changed the media environment by enabling quick and effective dissemination of false news via social networks. The rate of false news propagation depends on individuals' ability to differentiate between lies and truth in the information stream. Recent studies have identified several factors affecting the accuracy of fake news recognition by social network users. Among these are the credibility of the source [1, 2], the readers' thinking style [3–6], and social influence [7]. Exploring the influence of the style of thinking, researchers differentiated between two dimensions: rational thinking (as the ability to solve logical problems) and magical thinking (associated with beliefs in the extraordinary [8]). The belief in conspiracy

theories and the tendency to use implausible explanations when interpreting significant social or political events [9] may also influence individuals' ability of fake news detection. Recent research has also shown a relationship between the level of conspiracy thinking and the degree to which people's trust the police, their neighbors, or their relatives [9]. In addition, a number of studies looked at the interactions between conspiracy thinking and people's trust in existing medical paradigms [10], climate change [11], and right-wing authoritarianism [12]. At the same time, the effects of conspiracy theological thinking on people's ability to recognize fake news has not been investigated.

Furthermore, previous studies have identified such linguistics factors and stylistic features as the title [13] or the number of typographical errors [2, 14] as important predictors in fake news recognition.

It is worth mentioning that, to the date, experimental studies on fake news recognition have been focused predominantly on the assessment of political news. In our study, we are broadening this focus by including news from the sphere of business.

Thus, this study is guided by the following research questions: 1) What are the factors that influence the accuracy of fake news recognition; 2) To what extent are there differences in fake news recognition between news from different spheres? To answer our research questions, we will analyze the effects of such factors as rational thinking, conspiracy thinking, generalized trust, and role of the sensationalist headlines on the accuracy of fake news recognition when looking at a selection of political and business news.

## 2 The Role of Cognitive Factors and Personal Traits in News Assessment

### 2.1 Related Studies

**Thinking Styles.** An analysis of the literature shows that one of the most important factors in determining the quality of fake news detection is the style of thinking. Studies by Bronstein and colleagues have shown that rational thinking largely determines the ability of fake news detection. At the same time, dogmatism and belief in extrasensory phenomena reduce people's ability to differentiate between truth and lies [4–6]. Coe found that magical thinking heightens the susceptibility of a person towards fake news [8].

Considering the literature related to thinking patterns and people's abilities to evaluate the trustfulness of information, we propose to include the concept of conspiracy thinking, which has recently attracted considerable attention of political scientists, psychologists and sociologists. In 2018, an entire issue of the *European Journal of Social Psychology* was devoted to the concept of conspiracy theory and its impact on decision-making processes [15]. In their studies, Edelson and coauthors showed the influence of conspiracy thinking on decision-making mechanisms during elections [16]. Several studies have demonstrated a negative relationship between belief in conspiracies and interpersonal trust or trust in the police [9, 17, 18]. Uscinski and Olivella revealed a connection between conspiracy and authoritarian thinking [11].

Thus, we decided to investigate the effects of different thinking styles (rational, magical, and conspiratorial) on the accuracy of fake news recognition. Consequently, we have formulated the following hypotheses:

*Hypothesis 1. Rational thinking increases the accuracy of news recognition in any domain.*

*Hypothesis 2. Magical thinking decreases the accuracy of news recognition in any domain*

*Hypothesis 3. Conspiracy thinking decreases the accuracy of news recognition for the political domain only.*

### **General Trust**

The accuracy of differentiation between true and false news is usually associated with the credibility of the source [2] or with the degree of trust in a particular media outlet [8]. Even though, given the fact that on social media platforms, a news item can be presented without a link to the original source or with a misleading link, we decided to exclude this factor from our analysis and to provide all news without mentioning where they come from. Also, we assume that general trust in people may influence people's ability to recognize fake news. Consequently, we formulated the following hypothesis:

*Hypothesis 4. General trust in people decreases the accuracy of fake news recognition.*

**User Experience and Media Literacy.** Students from different departments have different literacy skills, interests, and patterns of news consumption. These three factors may influence their ability to recognize fake news [19]. This leads to the following two hypotheses:

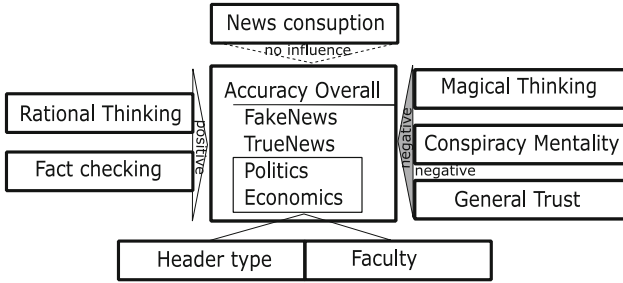
*Hypothesis 5. Interest in a particular type of news (e.g. politics) increases the accuracy of news recognition in this particular domain.*

*Hypothesis 6. Studying in a particular field (e.g. business or humanities) increases the accuracy of news recognition in this particular domain (business or humanities respectively).*

We also expect that people's ability to remember certain news (their prior experience) may influence the accuracy of fake news recognition. At the same time, some experiments have shown that people's memory may be inaccurate when it comes to news recollection and the very fact that certain news items (fake or real) may be repeated multiple times make them more believable [3]. Thus, we formulated the following hypothesis:

*Hypothesis 7. Ability to recall the news will have no influence on the accuracy of news recognition.*

**Fact Checking Behavior.** Fact-checking procedures are some the most powerful tools that individuals have at their disposal when it comes to fake news recognition. In our online experiment, participants operate in natural conditions and, thus, we cannot prevent them from double-checking certain news for accuracy. When it comes to fact checking, we have formulated the following hypothesis:



**Fig. 1.** News accuracy recognition model.

*Hypothesis 7. Fact checking increases the accuracy of news recognition in each domain.*

Thus, by studying the factors affecting the accuracy of news recognition (Fig. 1), we assume that the nature of their influence depends on the news section and the nature of its presentation (sensationalist headlines or not).

### 3 Methodology

#### 3.1 Experiment Design

We used a  $2 \times 2 \times 2$  design in our experiment. Students were offered 8 news items that varied by the following conditions: section - politics or business; type - false or true; with sensationalist (true) or neutral (false) headlines. Respondents were asked to indicate the degree of their trust in each news item on a 6-level Likert scale (from “absolutely false” to “absolutely true”). Control variables included: gender, age, level of education, university faculty, news consumption patterns, generalized trust, and political interests.

**News Selection.** The selection of news for this experimental study turned to be a challenge as fake and real news stories can differ in style, source, domain, and in their attractiveness to readers. This is why we decided against creating fake news stories ourselves and, instead, we used a series of existing fake news from a known fact-checking site – snopes.org. We selected four recent stories covering well-known persons (for politics) or international companies (for business) respectively. Our condition was that both companies and politicians should be widely known in Russia. Political news were about Donald Trump (2), Arnold Schwarzenegger (1), and climate change (1). Business news were about IKEA, Starbucks, Boston Dynamics, and the global bottled water industry. The full list of news items is presented in Appendix 1. Table A.1.

All news items were accompanied by two manually created titles. In all cases, one title was neutral and the other one – sensationalized. For example, the two titles for information about an IKEA advertisement campaign were: 1) “Test from IKEA” (“Тест от ИКЕА”) and 2) “IKEA shock” (“Шок от ИКЕА”). All news items was equalized in terms of their length and style to reduce the effect of wording and complexity. We also checked all news items for typographical errors.

### 3.2 Sample

We sent out our survey invitations by e-mail to students in three departments (Economics, Management, and Humanities) situated on a single campus of one of Russia's leading universities. 228 students (Response Rate: 12.25%) accessed the link to the study, and 203 completed both experimental and survey parts. 91 (44.82%) students were from the Faculty of Humanities, 37 (18.22%) from Management, and 75 (36.94%) from Economics. The majority of the sample were female students (female - 76.35%, male - 20.19%, other - 3.44%). The average age of participants was 20.3 years (Median = 20, SD = 2.63). In all faculties, the ratio of male to female respondents was similar. The students in Russia are characterized by low age dispersion, and we observe a similar trend in our sample - the mean age of respondent was 20 years (the minimum was 18 years and the maximum - 36). Two participants indicated that they were 51 and 109 years old respectively and were considered to be outliers and, thus, removed from further analysis.

By analyzing the time spent by our respondents on recognizing fake news, we were able to determine a group of participants who were either extremely quick or extremely slow in their responses. It is worth noting that the respondents, who spent a lot of time on this task, did not spend it on news checking. Instead, they used these time on other tasks unrelated to this study. Hence, this may be related to these respondents' overall distractedness and low motivation, which means that their results may have not been reliable. Similarly, there were students who we were extremely fast in answering questions.

Following a standard procedure, we excluded from further analysis all participants who did not answer our control questions and 5% from extremely fast and 5% from extremely slow respondents.

### 3.3 Control Variables

*Double-Checking of Information.* By administering the questionnaire online, we could not prevent our participants from using the Internet to double-check the news, so we included an information check question. For each news item, we asked participants whether they had checked this news during the experiment. Based on this information, we created the "News\_checked" variable. It is worth mentioning that the majority of participants (167-93.82%) did not check a single news item. One news was checked by 2 respondents, 2 by 5, 3 by 2, 4 by 1 and 5 news items by 1 participant (Table 1). In total participants performed 27 double-checked acts (for political news - 15 (55.55%), business - 12 (44.44%). News with sensationalist headlines were checked almost the same time (14, 51.85%), as with neutral titles (13, 48.14%).

*News Seen.* All news items selected for the experiment originated from publicly available sources and these news have been in circulation for a while, so it was possible that they had already been seen by our respondents. For each news item, we asked participants whether they had already seen this information. Based on these data, we created the "News\_seen" variable. It is worth mentioning that the majority of participants (91-51.12%) never saw a single news item prior to the experiment (Table 1).



**Table 1.** Frequency distribution for “News checked” and “News seen” variables.

| Number of news | 0   | 1  | 2  | 3  | 4 | 5 | Total |
|----------------|-----|----|----|----|---|---|-------|
| News seen      | 91  | 41 | 25 | 16 | 2 | 3 | 178   |
| News checked   | 167 | 2  | 5  | 2  | 1 | 1 | 178   |

*News Consumption.* Questions about news consumption included the following options: TV, Paper newspapers and journals, Radio, SNS and Forums, Bloggers, News sites and News aggregators, Friends, Other. Participants were asked to indicate their top three options All options were recoded as individual dummy variables. We present descriptive statistics in Appendix 1. Table A.2.

The top three sources of information for participants in our study were: SNS and forums – 164 (92.13%), news sites and aggregators – 101 (56.74%), and friends – 89 (50%).

*Accuracy.* We created seven accuracy coefficients. First, we calculated the overall accuracy (Acc) based on all news items included in the study (see Table 2). Next, we created one accuracy index for true (Acc<sub>truth</sub>) and one for false news (Acc<sub>fake</sub>). Then, we calculated accuracy indices for the two news sections (Acc<sub>bussn</sub>, Acc<sub>polit</sub>) and for news with sensationalist (Acc<sub>sens</sub>) and neutral headlines (Acc<sub>neutr</sub>). Descriptive statistics for the different dimensions of accuracy presented in Appendix 1. Table A.3.

**Table 2.** Descriptive statistics for overall accuracy

| Levels                 | 1 | 2  | 3  | 4  | 5  | 6  | 7 | 8 | Mean accuracy |
|------------------------|---|----|----|----|----|----|---|---|---------------|
| Number of participants | 7 | 20 | 33 | 40 | 50 | 24 | 2 | 2 | 4.10          |

The majority of participants were able to determine fewer than five news items correctly (84.26%). Six news were accurate recognized by 24 (13.48%) and only four participants were able to achieve higher accuracy (2.24%).

The accuracy of recognition for each news section was similar. At the same time, accuracy for true stories was lower than for fake ones, and participants, on average, were less accurate in deciphering political news compared to business news (see Appendix 1. Table A.3).

*Rational Thinking.* The standard procedure for measuring the rationality of thinking is the Cognitive Reflection Test created by Shane Frederick, however, as it was shown that women demonstrate significantly lower scores [20]. CRT is also routinely used as an pop-psychology test on the Internet and it was shown that participants who had already encountered these tasks obtained higher CRT scores [21]. This is why instead of CRT we decided to use the CRT-2 questionnaire created by Thompson and Openheim [22]. This questionnaire had only four questions, that we used in our study.

After checking the psychometric properties of the scale, we determined that this set of questions could not be considered as a single scale (Cronbach alpha = 0.37). For questions 1–3, most respondents were able to recognize the right answer, but for question 4 majority of participants selected the first intuitive (wrong) answer. It is worth mentioning that there are weak positive correlations between questions 1, 2, and 4 (min = 0.112, max = 22.3). Questions 3 and 4 are not correlated ( $-0.004$ ). It is evident that there are significant differences in answers to these questions. Hence, these questions were used as separate variables (Descriptive statistics presented in Appendix 1. Table A.4).

*Conspiracy Thinking.* Conspiracy thinking was measured by using four questions: three were adapted from the Conspiracy Mentality Scale [9] and one from the Conspiracy Mentality Questionnaire [23]. Even though these questions showed acceptable Cronbach alpha (0.69) in our sample, since they represented different beliefs of participants, we decided to use them separately (Descriptive statistics presented in Appendix 1. Table A.5).

*Magical Thinking.* To measure magical thinking, we selected top four items with the highest factor loading [24] adapted from the Magical Ideation Scale proposed by Eckblad and Chapman [25]. Selected items showed low reliability (Cronbach alpha = 0.59) and we decided to use these questions separately in our analysis. Descriptive statistics are presented in Appendix 1. A.7.

*General Trust.* News perception may be related to the general trust in people. To measure general trust, we used three questions adapted from the World Values Survey [26]. The results of combining three questions into a single scale showed low psychometric characteristics (Cronbach alpha = 0.55). Therefore, during the analysis, we did not combine these questions into a scale and, instead, considered them separately. Descriptive statistics are presented in Appendix 1. A.6.

### 3.4 Strategy of Analysis

The questions that we have selected to evaluate different thinking styles were not suitable for the creation of a reliable scale, so during the analysis we treated each question separately. First, we ran a correlation analysis of the variables, and then we conducted a dominance analysis [27] based on generalized linear model regressions (GLM). This allowed us to define the most influential items. Then, we conducted a dominance analysis for each factor (general trust, rational thinking, etc.), and after that the most influential factors from each dimension were used in a GLM regression to model the overall accuracy of fake news recognition (R core package stats). Next, we reproduce the same approach to analyze all the dimensions of accuracy. If several items from one dimension contributed the most to the different aspects of accuracy, we selected the one with the highest impact (average contribution). Finally, we performed regression modeling using GLM regressions (R core package stats).

## 4 Results

The results of the GLM regression modeling for general accuracy showed (Table 3) that the only variable significantly affecting the accuracy of fake news recognition is “News\_check” (the number of news checked by respondents).

Table 3 shows that the constructed model does not have high accuracy, which, on the one hand, may indicate that factors are weak, and on the other hand, may be associated with the difficulty of the task (as we selected plausible false news and implausible real stories).

**Table 3.** Regression model for overall accuracy (main variables)

| <i>Predictors</i>  | Accuracy         |                  |
|--|------------------|------------------|
|  | <i>Estimates</i> | <i>p</i>         |
| (Intercept)  | 5.04             | <b>&lt;0.001</b> |
| Conspiracy thinking: C.2 “Government agencies closely monitor all citizens”                                    | −0.12            | 0.171            |
| Magical thinking: “MT.4 I have felt that I might cause something to happen just by thinking too much about it” | −0.11            | 0.167            |
| General trust: Trust.1 “Most of the time people try to be helpful”   | −0.08            | 0.447            |
| CRT2.4 “How many cubic feet of dirt are there in a hole that is 3” deep × 3” wide × 3” long?”                  | −0.43            | 0.125            |
| who [humanities]   | 0.35             | 0.149            |
| who [managers]   | 0.08             | 0.789            |
| TV   | 0.36             | 0.180            |
| News_checked   | 0.46             | <b>0.006</b>     |
| Observations   | 178              |                  |
| R <sup>2</sup> Nagelkerke  | 0.220            |                  |

Next, we conducted a dominance analysis of variables for each dimension of accuracy (See Appendix 1. Table A.8). An analysis of the variables showed that “News seen” dominate among socio-demographic control variables (gender, age, faculty, news seen), and News check among behavioral control variables (news checked, interest in politics, interest in business) (Appendix 1. Table A.8). Also the analysis reveal existence of the differences in factors influences business and politics news recognition (Table 4). For example, conspiracy thinking is especially significant (Est. = −0.17,  $p < 0.05$ ) for a business news accuracy model. It can be noted that some variables (Trust.3, News\_checked) turned out to be significant for all dimensions of accuracy.

**Table 4.** Regression models for accuracy (Business and politics domains, neutral and sensationalist titles)/

| <i>Predictors</i>         | Business   |                | Politics    |                | Neutral     |                | Sensationalist |                |
|---------------------------|--|----------------|-------------|----------------|-------------|----------------|----------------|----------------|
|                           | <i>Est.</i>  | <i>p</i>       | <i>Est.</i> | <i>p</i>       | <i>Est.</i> | <i>P</i>       | <i>Est.</i>    | <i>p</i>       |
| (Intercept)               | 4.59   | < <b>0.001</b> | 2.63        | < <b>0.001</b> | 2.77        | < <b>0.001</b> | 4.45           | < <b>0.001</b> |
| Conspiracy mentality:     | -0.17  | <b>0.018</b>   | -0.00       | 0.989          | -0.10       | 0.159          | -0.07          | 0.360          |
|                           | C.1 “Many so called “coincidences” are in fact clues to how things really happened”                        |                |             |                |             |                |                |                |
| Magical thinking:         | -0.07  | 0.117          | -0.03       | 0.565          | 0.05        | 0.297          | -0.15          | <b>0.003</b>   |
|                           | MT.1 “I sometimes have a feeling of gaining or losing energy when certain people look at me or touch me”). |                |             |                |             |                |                |                |
| General trust:            | -0.06  | 0.307          | 0.07        | 0.329          | 0.10        | 0.078          | -0.10          | 0.119          |
|                           | Trust.3 “Most of the time people try to be helpful”  |                |             |                |             |                |                |                |
| Rational Thinking:        | -0.34  | <b>0.038</b>   | 0.26        | 0.158          | -0.02       | 0.895          | -0.06          | 0.745          |
|                           | CRT2.1 “If you’re running a race and you pass the person in second place, what place are you in?”          |                |             |                |             |                |                |                |
| News_seen                 | -0.02  | 0.767          | -0.13       | 0.063          | 0.07        | 0.262          | -0.22          | <b>0.001</b>   |
| TV                        | 0.16   | 0.355          | 0.18        | 0.364          | 0.39        | <b>0.030</b>   | -0.04          | 0.830          |
| News_check                | 0.28   | <b>0.010</b>   | 0.18        | 0.159          | 0.17        | 0.115          | 0.29           | <b>0.016</b>   |
| Observations              | 178  |                | 178         |                | 178         |                | 178            |                |
| R <sup>2</sup> Nagelkerke | 0.163  |                | 0.078       |                | 0.133       |                | 0.221          |                |

Regression modeling for each accuracy dimension showed that Magical Thinking, as expected, had a negatively effect on the accuracy of news perception. At the same time, this effect was significant only for news with sensationalist headlines (Est. = -0.15, p = 0.003).

There is a significant negative correlation between conspiracy theological thinking and the accuracy of recognition of business news (Est. = -0.17, p = 0.0018).

It is worth noting that if the respondents believed that they had already seen a particular news, that reduced their accuracy in determining the accuracy of news items with sensationalist headlines (Est. = -0.22, p = 0.001).

Surprisingly, respondents who indicated “TV” as one of the main sources of information were more accurate in determining the accuracy of news with neutral headlines (Est. = 0.39, p = 0.030). At this stage of our research, we do not know how to explain this pattern, which seems counterintuitive to the prevailing perception of Russian TV.

Another surprising finding the negative correlation between rational thinking and the accuracy of recognition of business news (Est. =  $-0.34$ ,  $p = 0.038$ ), which, in our opinion, is associated with the particular strategy of news selection in our study (plausible false news and implausible truthful ones were selected). News checking has a significant effect only for business-related news and for news with sensationalist headings. We can not speculate that this is due to the fact that participants have a tendency of checking business news and news with sensationalist heading more often, as number of double-checking acts are almost the same both for business and political news, and for sensationalist and neutral headlines.

Thus, we can conclude that the new domain is a significant factor that influences the perception of news. We can not see any universal factors as we predicted in hypothesis H.1, H.2, H.3. At the same time, the results do not allow us to fully rejected the hypothesis that thinking style influences fake news recognition. We also found no influence of trust (H.4), interest in politics or business (H.5) or faculty (H.6), so we can safely reject these hypotheses. The “*ability to recall the news will have no influence on the accuracy of news recognition*” is the only hypothesis we can partially confirm in our study.

Our research has demonstrated that the domain of political news is the least predictable area, as none of tested variables contribute to accuracy in this sphere. The use of sensationalist headlines attracts the readers’ attention, but, at the same time, these headlines increase the likelihood of double-checking these news items in other information sources (Table 4).

## 5 Discussion

The results raise two important questions. First, can we interpret results as influence of a factors, as all developed and adapted scales (conspiracy mentality, trust and rational thinking) at Russian sample have low psychometric features and we used items separately. Hence, we could not conclude that these factors have any relationship with the accuracy of fake news recognition. Consequently, we need to conduct further research on the role of these factors.

The second question is about the criteria of news selection. This is an important matter because it may potentially affect the results. If the selected news cannot be distinguished by rational thinking, it is evident why the news checking behavior is the main significant factor.

It worth mentioning that in the experiments conducted by Pennycook and coauthors, the CRT and CRT2 demonstrated high correlation (0.57), but they didn’t report the Cronbach alpha- coefficient for CRT2, which may be a sign of low reliability of this construct. Moreover, Pennycook and coauthors did not use the CRT2 measurement in their regression analysis [5].

In comparison to the study of Pennycook and coauthors, our participants showed less accuracy in recognizing true news ( $Acc_{\text{true}} = 1.86$ , instead of 2.76 [5]) and they were more accurate in recognizing false news ( $Acc_{\text{fake}} = 2.24$ , instead of 1.83 [5]).

Our data analysis shows that conspiracy thinking has an effect only on the accuracy in business-related news. This goes contrary to our initial expectation that would have an effect on respondents' ability to differentiate between true and fake news in the political domain. However, once we took a deeper look at the composition of our news, we discovered that the business news that we selected (for example, the news about bottled water or IKEA) could have been explained from a conspiracy beliefs angle, while selected political news could not be explained by any conspiracy ideas.

Thus, we can conclude that certain methodological limitations have impacted our abilities to fully explore our initial hypotheses.

### ***Limitations***

Our research has three main limitations. The first limitation has to do with the type of news selected for the study, where the relatively low accuracy in news recognition may be associated with the difficult task of recognizing fake news that in fact look credible. Second, since our participants were university students who are probably more rational than the average in Russia, the results obtained from this sample may not be generalized for the whole population of the Russian Federation.

Third limitation arises from the low psychometric features of proposed thinking scales and do not allow us to make final conclusion about the role of thinking style in fake news recognition.

## **6 Conclusion**

The main questions of our study were the following: what are the main factors that influence the accuracy of fake news recognition and do they differ for news from different spheres?

Our analysis of the dominance and indices of accuracy shows that the main factor responsible for the quality of news recognition is the number of news checked from external sources. Other factors, such as thinking style, have a significant influence on the accuracy of business news recognition or on news with sensationalist headlines. It is interesting that such major factors as general trust and such control variables as gender and faculty proved to be insignificant when it comes to fake news recognition.

Our analysis raises several further questions, such as our limitation to generalize our findings beyond the walls of an advanced university in Russia. We would also like to research the role of experimental settings in investigating different factors responsible for fake news recognition.

**Acknowledgements.** The research was implemented in the framework of the Russian Scientific Fund Grant №19-18-00206 at the National Research University Higher School of Economics (HSE) in 2019.

## Appendix 1

**Table A.1.** Amount of participants indicated the information source as main source of information

| Domain | True | Title   | Text   |
|--------|------|---|--|
| P      | T    | Neutral: Пожары в Калифорнии<br>Sensationalist: Гори, Калифорния!         | Трампа подверг критике действия властей Калифорнии по ликвидации пожаров. «Каждый год, как только огонь начинает полыхать в Калифорнии, <...> он [губернатор Калифорнии] приходит к федеральному правительству за финансовой помощью. Хватит!» |
| P      | T    | Neutral: На Шварценеггера напали<br>Sensationalist: Терминатора избили    | На Арнольда Шварценеггера напали в ЮАР. И это попало на видео — на кадрах мужчина разбегается и в прыжке бьет 71-летнего актера ногой в спину. Железный Арни такого не ожидал и не удержал равновесие  |
| P      | F    | Neutral: Отчет о климате<br>Sensationalist: Глобальная подделка           | NASA и межправительственная группа экспертов по изменению климата (NOAA) подделали данные в глобальном температурном отчете GISTEMP, чтобы преувеличить проблему глобального потепления. В докладе были найдены ошибки                         |
| P      | F    | Neutral: Трамп и Маттарелла<br>Sensationalist: Трамп отжигает!            | Президент США Дональд Трамп назвал лидера Италии «президентом Мוצареллой». Это произошло на пресс-конференции после официальной встречи глав двух государств. В действительности итальянского коллегу Трампа зовут Серджо Маттарелла           |
| E      | T    | Neutral: Вода в бутылках<br>Sensationalist: Лохотрон в бутылке            | Исследователи американской Environmental Working Group подчеркивают: половина производителей бутилированной воды признают, что это та же самая водопроводная вода, «прошедшая дополнительную очистку»  |
| E      | T    | Neutral: Тест от IKEA<br>Sensationalist: Шок от IKEA!                     | Компания IKEA предложила женщинам пройти тест на беременность с помощью специальной рекламной страницы в шведском журнале Amelia. Если беременная женщина помочится на нее, то на странице появится новая цена                                 |
| E      | F    | Neutral: Робот угрожал человеку<br>Sensationalist: Робот взбесился!       | На испытаниях инженерной компании Boston Dynamics робот, управляемый искусственным интеллектом, угрожал человеку оружием. Он напал на сотрудников компании, проводивших тестирование устройства  |
| E      | F    | Neutral: Новые стаканы Starbucks<br>Sensationalist: «Сатанинские» стаканы | Акции компании Starbucks упали после того, как генеральный директор Говард Шульц представил новый дизайн стаканов. Пользователи сети сочли их облик «сатанинским»  |

**Table A.2.** Amount of participants indicated the information source as main source of information

| Sources of information          | Amount of participants indicated source among three top sources | %     |
|---------------------------------|---|-------|
| SNS and forums                  | 164   | 92.13 |
| News sites and news aggregators | 101   | 56.74 |
| Friends                         | 89  | 50.00 |
| Bloggers                        | 65  | 36.52 |
| TV                              | 36  | 20.22 |
| Radio                           | 7   | 3.93  |
| Other                           | 5   | 2.81  |
| Paper newspapers and journals   | 1   | 0.56  |

**Table A.3.** Descriptive statistics for accuracy dimensions

| Levels of accuracy | Amount of participants by levels |    |    |    |    | Mean |
|--------------------|----------------------------------|----|----|----|----|------|
|                    | 0                                | 1  | 2  | 3  | 4  |      |
| Fake               | 2                                | 36 | 73 | 51 | 16 | 2.24 |
| True               | 15                               | 43 | 77 | 38 | 5  | 1.86 |
| Business           | 5                                | 39 | 69 | 49 | 16 | 2.18 |
| Politics           | 19                               | 41 | 62 | 47 | 9  | 1.92 |
| Exclamation        | 13                               | 43 | 59 | 49 | 14 | 2.04 |
| Neutral            | 7                                | 45 | 67 | 49 | 10 | 2.04 |

**Table A.4.** Descriptive statistics for rational thinking questions

| D      | Question  | 0   | 1   |
|--------|---|-----|-----|
| CRT2.1 | If you're running a race and you pass the person in second place, what place are you in?<br>(intuitive answer: first; correct answer: second)                 | 48  | 130 |
| CRT2.2 | A farmer had 15 sheep and all but 8 died. How many are left? (intuitive answer: 7; correct answer: 8)   | 18  | 160 |
| CRT2.3 | Emily's father has three daughters. The first two are named April and May. What is the third daughter's name? (intuitive answer: June; correct answer: Emily) | 21  | 157 |
| CRT2.4 | How many cubic feet of dirt are there in a hole that is 3'' deep × 3'' wide × 3'' long? (intuitive answer: 27; correct answer: none)                          | 145 | 33  |



**Table A.5.** Descriptive statistics for conspiracy beliefs questions

| ID  | Question   | Mean  | SD    | Skew   | Kurtosis | SE    |
|-----|--|-------|-------|--------|----------|-------|
| C.1 | Many so called “coincidences” are in fact clues as to how things really happened                           | 4.241 | 1.026 | −0.647 | 0.394    | 0.076 |
| C.2 | Government agencies closely monitor all citizens   | 4.314 | 1.298 | −0.500 | −0.677   | 0.097 |
| C.3 | The government or covert organizations are responsible for events that are unusual or unexplained          | 3.606 | 1.311 | 0.000  | −0.681   | 0.098 |
| C.4 | The alternative explanations for important societal events are closer to the truth than the official story | 4.078 | 1.227 | −0.221 | −0.7637  | 0.092 |

**Table A.6.** Descriptive statistics for conspiracy beliefs questions

| ID      | Questions   | Mean  | sd    | Median | Skew   | Kurtosis | se    |
|---------|---|-------|-------|--------|--------|----------|-------|
| Trust.1 | Most of the time people try to be helpful                           | 3.769 | 1.093 | 4      | −0.209 | −0.49755 | 0.081 |
| Trust.2 | People can be trusted   | 2.915 | 1.225 | 3      | 0.159  | −0.943   | 0.091 |
| Trust.3 | Most people would try to take advantage of you if they got a chance | 3.938 | 1.276 | 4      | −0.160 | −0.73096 | 0.095 |

**Table A.7.** Descriptive statistics for Magical thinking questions

| ID   | Questions  | Mean  | sd       | Median | Skew   | Kurtosis | se    |
|------|--|-------|----------|--------|--------|----------|-------|
| MT.1 | I sometime have a feeling of gaining or losing energy when certain people look at me or touch me | 3.387 | 1.584    | 4      | −0.087 | −1.228   | 0.118 |
| MT.2 | Someone can put a jinx on me.  | 2.089 | 1.345    | 2      | 1.054  | 0.120    | 0.100 |
| MT.3 | I have wondered whether the spirits of the dead can influence the living                         | 1.882 | 1.146    | 2      | 1.416  | 1.527    | 0.085 |
| MT.4 | I have felt that I might cause something to happened just by thinking too much about it          | 3.449 | 1.445908 | 3.5    | −0.016 | −1.084   | 0.108 |

**Table A.8.** Dominant variables from each factor for dimensions of Accuracy (average contribution)

| Dimensions           | Business                 | Politics              | Neutral                | Exclamation              | Overall                     |
|----------------------|--------------------------|-----------------------|------------------------|--------------------------|-----------------------------|
| Conspiracy           | <b>C.1 (0.024)</b>       | consp_monitor (0.002) | consp_truth (0.001)    | consp_monitor (0.014)    | consp_monitor (0.01)        |
| Magic thinking       | MT.1 (0.01)              | magic_spirits (0.018) | magic_eye (0.02)       | <b>MT.1 (0.041)</b>      | magic_think (0.017)         |
| Trust                | Trust.3 (0.013)          | Trust.3 (0.004)       | <b>Trust.3 (0.017)</b> | Trust.3 (0.024)          | people_help (0.002)         |
| Rational thinking    | CRT2.1 (0.011)           | <b>CRT2.1 (0.012)</b> | CRT2.2 (0.007)         | CRT2.4 (0.004)           | CRT2.4 (0.01)               |
| Control              | Faculty (0.005)          | News_seen (0.014)     | News_seen (0.016)      | <b>News_seen (0.047)</b> | who (0.01)                  |
| News sources         | News aggregators (0.011) | TV (0.01)             | <b>TV (0.029)</b>      | Friends (0.019)          | TV (0.015)                  |
| Behavioral variables | News_checked (0.034)     | News_checked (0.008)  | News_checked (0.029)   | News_checked (0.011)     | <b>News_checked (0.036)</b> |

## References

1. Pornpitakpan, C.: The persuasiveness of source credibility: a critical review of five decades' evidence. *J. Appl. Soc. Psychol.* **34**, 243–281 (2004). <https://doi.org/10.1111/j.1559-1816.2004.tb02547.x>
2. Morris, M.R., Counts, S., Roseway, A., Hoff, A., Schwarz, J.: Tweeting is believing?: understanding microblog credibility perceptions. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work - CSCW 2012, Seattle, Washington, USA, p. 441. ACM Press (2012). <https://doi.org/10.1145/2145204.2145274>
3. Pennycook, G., Cannon, T.D., Rand, D.G.: Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol. Gen.* **147**, 1865–1880 (2018). <https://doi.org/10.1037/xge0000465>
4. Pennycook, G., Rand, D.G.: Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. <https://onlinelibrary.wiley.com/doi/abs/10.1111/jopy.12476>. <https://doi.org/10.1111/jopy.12476>. Accessed 04 Sep 2019
5. Pennycook, G., Rand, D.G.: Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019). <https://doi.org/10.1016/j.cognition.2018.06.011>
6. Bronstein, M.V., Pennycook, G., Bear, A., Rand, D.G., Cannon, T.D.: Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *J. Appl. Res. Mem. Cogn.* **8**, 108–117 (2019). <https://doi.org/10.1016/j.jarmac.2018.09.005>
7. Colliander, J.: “This is fake news”: investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Comput. Hum. Behav.* **97**, 202–215 (2019). <https://doi.org/10.1016/j.chb.2019.03.032>
8. Coe, C.M.: Tell me lies: fake news, source cues, and partisan motivated reasoning (2018)
9. Stojanov, A., Halberstadt, J.: The conspiracy mentality scale: distinguishing between irrational and rational suspicion. *Soc. Psychol.* **50**, 215–232 (2019). <https://doi.org/10.1027/1864-9335/a000381>

10. Lamberty, P., Imhoff, R.: Powerful pharma and its marginalized alternatives?: effects of individual differences in conspiracy mentality on attitudes toward medical approaches. *Soc. Psychol.* **49**, 255–270 (2018). <https://doi.org/10.1027/1864-9335/a000347>
11. Uscinski, J.E., Olivella, S.: The conditional effect of conspiracy thinking on attitudes toward climate change. *Res. Politics* **4**, 205316801774310 (2017). <https://doi.org/10.1177/2053168017743105>
12. Frischlich, L., Brinkschulte, F., Becker, M.: The moderating role of right-wing authoritarianism and conspiracy mentality for the perception and effects of distorted news articles, 11
13. Horne, B.D., Adali, S.: This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: *The Workshops of the Eleventh International AAAI Conference on Web and Social Media AAAI (ICWSM 2017)*, Technical Report WS-17-17: News and Public Opinion, pp. 759–766 (2017)
14. Sitaula, N., Mohan, C.K., Grygiel, J., Zhou, X., Zafarani, R.: Credibility-based fake news detection. [arXiv:1911.00643](https://arxiv.org/abs/1911.00643) [cs] (2019)
15. van Prooijen, J.-W., Douglas, K.M.: Belief in conspiracy theories: basic principles of an emerging research domain. <https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2530>. <https://doi.org/10.1002/ejsp.2530>. Accessed 04 Sep 2019
16. Edelson, J., Alduncin, A., Krewson, C., Sieja, J.A., Uscinski, J.E.: The effect of conspiratorial thinking and motivated reasoning on belief in election fraud. *Polit. Res. Q.* **70**, 933–946 (2017). <https://doi.org/10.1177/1065912917721061>
17. Abalakina-Paap, M., Stephan, W.G., Craig, T., Gregory, W.L.: Beliefs in conspiracies. *Polit. Psychol.* **20**, 637–647 (1999)
18. Goertzel, T.: Belief in conspiracy theories. *Polit. Psychol.* **15**, 731–742 (1994)
19. Herrero-Diz, P., Conde-Jiménez, J., Tapia-Frade, A., Varona-Aramburu, D.: The credibility of online news: an evaluation of the information by university students /La credibilidad de las noticias en Internet: una evaluación de la información por estudiantes universitarios. *Cultura y Educación* **31**, 407–435 (2019). <https://doi.org/10.1080/11356405.2019.1601937>
20. Frederick, S.: Cognitive reflection and decision making. *J. Econ. Perspect.* **19**, 25–42 (2005). <https://doi.org/10.1257/089533005775196732>
21. Stieger, S., Reips, U.-D.: A limitation of the cognitive reflection test: familiarity. *PeerJ* **4**, e2395 (2016). <https://doi.org/10.7717/peerj.2395>
22. Thomson, K.S., Oppenheimer, D.M.: Investigating an alternate form of the cognitive reflection test. *Judgm. Decis. Mak.* **11**, 15 (2016)
23. Bruder, M., Haffke, P., Neave, N., Nouripanah, N., Imhoff, R.: Measuring individual differences in generic beliefs in conspiracy theories across cultures: conspiracy mentality questionnaire. *Front. Psychol.* **4** (2013). <https://doi.org/10.3389/fpsyg.2013.00225>
24. Байрамова, Э.Э., Ениколопов, С.Н.: Адаптация методики определения уровня магического мышления М. Экблада и Л.Дж. Чапмана на русскоязычной выборке. *Психиатрия*, 40–46 (2016)
25. Eckblad, M., Chapman, L.J.: Magical ideation as an indicator of schizotypy. *J. Consult. Clin. Psychol.* **51**, 215–225 (1983). <https://doi.org/10.1037/0022-006X.51.2.215>
26. Inglehart, R., et al. (eds.): *World Values Survey: Round Six* (2014). <http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp>
27. Navarrete, C.B., Soares, F.C.: Dominance Analysis Package. <https://cran.r-project.org/web/packages/dominanceanalysis/dominanceanalysis.pdf>



# Using Deep Learning to Detect Rumors in Twitter

Eliana Providel<sup>1,2</sup> and Marcelo Mendoza<sup>1</sup>(✉)

<sup>1</sup> Departamento de Informática, Universidad Técnica Federico Santa María, Santiago, Chile

`eliana.providel@sansano.usm.cl`, `mmendoza@inf.utfsm.cl`

<sup>2</sup> Escuela de Ingeniería Civil Informática, Universidad de Valparaíso, Valparaíso, Chile

**Abstract.** The automatic detection of rumors in social networks is an important problem that would allow counteracting the effects that the propagation of false information produces. We study the performance of deep learning architectures in this problem, analyzing ten different machines on word2vec and BERT. Our results show that some architectures are more suitable for some particular classes, suggesting that the use of committee machines would offer advantages in this task.

**Keywords:** Rumor detection · Empirical factors · Deep learning

## 1 Introduction

Deep learning architectures have been successfully tested in different scenarios where it is necessary to evaluate the predictive capacity of the data that describes an event of interest. Its success is directly dependent on the complexity of these architectures, which are implemented over millions of parameters that allow detecting patterns of infrequent correlation in the data. One of its limitations is the need to train these networks with large volumes of data, preventing network over-fitting to training data.

Social networks are the medium through which more information is shared today. In their role as information mediators, social networks concentrate billions of users worldwide on their different platforms. His insight into people's lives has changed how we interact with our social environment. The new generations, besides, are reported mainly through social networks, which have displaced traditional media such as the written press, radio, and television.

Social networks users interact with information of different types. Many times, through social networks, we receive information about news events. However, the attitude of social network users is not restricted only to receive information. One of the highest potentials of social networks is to provide users with mechanisms through which it is possible to comment on the information, supporting and forwarding it to our followers, or questioning it. The ability that

users have to modify the content of the information is a crucial aspect to understand how the propagation of information in social networks finally occurs.

As social networks are the media through which new generations are informed, it is necessary to provide mechanisms to spread truthful and verifiable information. The democratization of access to information and the sharing of information on social networks has enormous potential, but at the same time, it has a significant threat. Networks allow information to be shared without the mediation of editorial lines, which are often coerced by interest groups. This fact that is potential at the same time is a threat since the non-mediation of an editorial line allows spreading unconfirmed information. This scenario has favored the proliferation of false and biased information, which seeks to influence public opinion, misinforming [2]. In short, what seemed a virtue of social networks, has also become its primary threat.

One of the ways to counteract the effect that the proliferation of fake news and rumors have on social networks is to provide information verification mechanisms. Journalists play a fundamental role in this aspect. Different verification initiatives have made it possible to establish levels of the truthfulness of news, which favor the distinction between confirmed news and fake news. For example, Politifact have allowed verifying information in the United States, detecting a growing proliferation of fake news during political electoral campaigns [1]. Verification by human experts is the most accurate, truthful, and reliable way to determine the veracity of the information. However, it requires a remarkable check effort, limited by the availability of human resources. Consequently, these valuable efforts allow for an ex-post verification, which is after the information has been propagated.

Artificial intelligence can help fight fake news in different ways. One of which we take charge in this work is the automatic detection of rumors using the comments that users make. Users of social networks make use of their common sense when interacting with information, and many of them question or deny them. For the methods of automatic detection of fake news, the use of the information that the users themselves provide is a valuable source that would allow verification during the propagation of fake news. The scalability of computational methods would also favor the processing of a higher volume of information, helping journalists to have verification information during the propagation of a rumor.

In this work, we study different deep learning architectures that process social network conversational threads in order to establish the level of veracity of a rumor. To do this, we tested different encodings of words along with sequential learning architectures based on extensions and variants of recurrent neural networks. These architectures process the threads of conversation in the same order in which they occur on social networks. The objective of this work is to determine which word encodings are more suitable for this type of tasks, and which recurrent network architectures have better predictive capabilities in this scenario. To meet this objective, we report extensive experimentation evaluating the performance of different deep network configurations in this task, identifying under which experimental configurations the methods acquire better predictive capabilities.

The work is organized as follows. In Sect. 2, we review related work. The description of the data used in this study is provided in Sect. 3. Section 4 presents the experimental and methodological design of this study. Experiments and results are shown in Sect. 5. Section 6 presents the discussion of results. Finally, we conclude in Sect. 7, providing remarks and outlining future work.

## 2 Related Work

Social networks have had enormous value in the analysis of rumor veracity. It has been shown that social network users tend to question or deny false rumors [18], in a kind of collective phenomenon that allows the network to self-regulate against the proliferation of fake news. One of the first methods that followed this perspective was proposed by Castillo *et al.* [5], who extracted aggregated features from Twitter to classify news according to their level of veracity. The results obtained by the authors showed that social networks provided data with high predictive capacity that would allow the early detection of fake news [6]. These works focused their efforts on providing white-box models that favored the explainability of the results achieved by the classifiers. For this reason, many of the reported results correspond to models based on decision trees or Bayesian networks.

In a sequel, many improvements and extensions were proposed to the methods introduced by Castillo *et al.* [5]. For example, Zhao *et al.* [22] showed that the use of lexical features of the news comments were crucial data elements for the analysis of the veracity of the information. The lexical features placed in relevance in the work of Zhao *et al.* are related to the stance that users adopt in front of new information, something that had already been consigned in the work of Mendoza *et al.* [18]. This approach was systematized later, denoting this task as automatic stance detection. It has been shown that automatic stance detection can provide essential insights into fake news detection methods. Buguño *et al.* [3] recently showed that automatic stance detection based on lexical characteristics is a challenging problem that requires sophisticated deep learning architectures in order to be successfully addressed.

Later work showed that the dynamics of the propagation phenomenon in the social network could also reveal valuable information for the detection of rumors. For example, Kwon *et al.* [11] studied the volume of tweets along time, showing that fake news spreads faster than the real ones. This finding is due to the emotional burden that fake news produces on users, since its emotional impact triggers a compulsion towards the sharing of this type of information, accelerating its propagation in the network. A large-scale longitudinal study was conducted by Friggeri [8], who observed differences between true and fake news in terms of news sharing rates. Hannak *et al.* [9] determined that fact checking affects the dynamics observed in terms of volume and news sharing. The predictability of this data was studied by Ma *et al.* [13], who used kernel-based learning to model the patterns of spreading fake news on social networks. Subsequently, this study was replicated in Sina Weibo [21], showing that in this

network, there are also measurable differences in temporal propagation patterns between true and fake news.

One way to integrate temporal and lexical characteristics into a coherent representation is to use sequential learning machines [10]. Deep learning architectures offer good properties for this purpose, with recurrent neural networks being the most commonly used machines in sequential learning. Ma *et al.* [12] showed that recurrent neural networks offer functional predictive capabilities in rumor detection by being fed with comment sequences. Bugueño and Mendoza [4] tested several recurrent neural network architectures suggesting that bidirectional networks would offer better rumor detection results than unidirectional networks. The reason why sequential machines outperform other types of architectures in performance is that the conversational thread simultaneously accounts for the volume of comments associated with a story and the users' stance. These models have also been used to determine the most influential users in spreading fake news, using soft-attention mechanisms [20]. The use of multi-task learning techniques has also been explored in this domain [15], showing that the joint learning of stance and rumor detection offers improvements in the performance of rumor detection methods. A variant of the recurrent neural network, called the recursive network, allows ingesting the conversational trees in a neural network according to an in-order traversal path of the thread. It has been shown that methods based on recursive networks offer good results in the task of detecting rumors [16]. Recently, generative adversarial networks have also been studied in this problem, showing interesting results [17].

In summary, what the previous works show is that different deep learning architectures offer advantages to the methods of rumor detection. However, it is not clear which of these architectures is best suited to address this problem. The objective of this work is to clarify this question.

### 3 Data Description

The task of automatic rumor detection requires tagged datasets from which to conduct the training of a learning machine. The consolidation of these datasets is an expensive process that requires the effort of experts who tag news for these purposes. Initially, the detection of rumors was addressed using two classes, true rumor, and false rumor [5]. Subsequently, Zubiaga *et al.* [23] refined the rumor typology by defining four classes, true rumor, false rumor, non-rumor, and unverified. In this work, we will use this last typology to conduct our experiments.

A true rumor corresponds to news that are propagated without being confirmed, for which ratifying evidence is available after a while. A false rumor is spread without being confirmed, for which after a while, there is enough evidence that allows rejecting its truthfulness. A non-rumor corresponds to a verifiable news from the moment of propagation. Finally, an unverified news is a rumor of which there is no evidence available to support or deny its truthfulness.

For experimental evaluation, we use a publicly available Twitter dataset released by Ma *et al.* [14], named Twitter 16. This version of the dataset comprises a collection of wide spread tweets along with their propagation threads

(i.e. replies, comments and retweets) provided as a tree structure. Each propagation tree was annotated into one of the four classes of the typology proposed by Zubiaga *et al.* [23] by human experts. The veracity labels correspond to tags provided by verification sites administered by expert journalists, as is the case of snopes.com and emergent.info. These labels are considered as the gold standard of this task.

We note that the proportion between tweets and retweets is highly unbalanced (almost 1 to 10). Then, to prevent over-fitting we provided only distinct tweets to each model. The dataset comprises 205 non rumors, 205 false rumors, 205 true rumors, and 203 unverified rumors, with a total amount of 21741 posts. The text was processed dropping out stopwords.

Each conversational thread was serialized, creating a sequence of tweets sorted according to timestamp. The dataset comprises variable-length sequences, the length of the shorter sequence being two tweets, and the longest being 599 tweets. The average length of the sequences is 36 tweets, which was used as the length of the input sequence. The longer sequences were truncated and the shorter ones were adjusted to this length using padding.

Each sequence of tweets was ingested in each model by vectorizing the tweets. Two text encoders were used for this purpose. The first, word2vec [19], provides an encoding at the level of each word. To construct an encoding of each tweet, the vectors of each word were averaged. This technique is known as average word embeddings. A second encoding was obtained using BERT [7], which has a variant for which a vector representing the sentence is provided. This variant, known as BERT as service, was used to encode each tweet.

In the case of word2vec, the word embeddings were trained by processing tweets from the Twitter 16 dataset. This decision allows having vectorized representations of words appropriate to the dataset used to train the models. The training of word2vec was done using a 300-dimensional embedding size, with sliding windows of five tokens and padding to handle variable-length sentences. The continuous bag-of-words (CBOW) variant of word2vec was used to construct these word embeddings. The training of these embeddings considered 200 epochs.

In the case of BERT, no specific training was done on the Twitter 16 dataset. Instead, the BERT as service variant<sup>1</sup> trained on Wikipedia was used, getting the vectors for each tweet of Twitter 16 using the pre-trained Wikipedia model. The embedding size of BERT has 768 dimensions and was trained using a transformer with 12 layers and 12 heads, with a total amount of 110 millions of parameters.

## 4 Experimental Methodology

In this study, different deep network architectures were tested using the Twitter 16 dataset. To conduct the validation process, the cross-validation methodology was applied using five folds. Because the data is balanced between the four classes, the measure of accuracy was used, which allows conducting a validation

<sup>1</sup> <https://github.com/hanxiao/bert-as-service>.



in problems with balanced classes. All architectures were trained considering 200 epochs, with 256 and 512 dimensionality units.

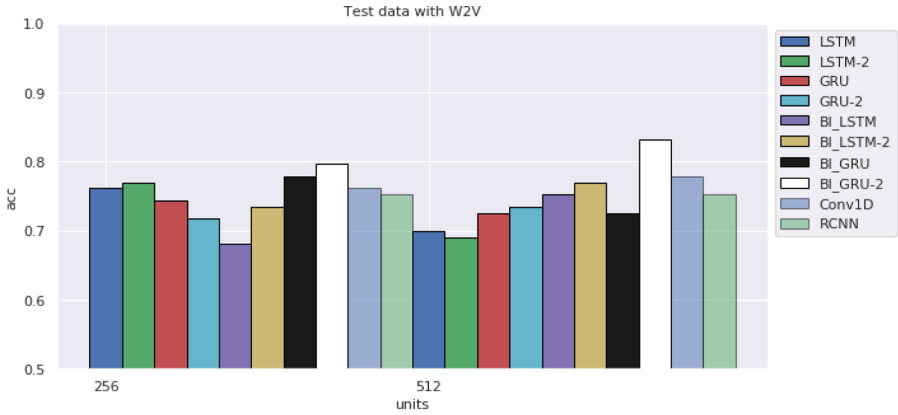
Ten different architectures were tested, these are LSTM, Stacked LSTM, GRU, Stacked GRU, bidirectional LSTM (bi-LSTM), bidirectional stacked LSTM, bidirectional GRU (bi-GRU), bidirectional stacked GRU, convolutional 1D, and recurrent convolutional neural network (RCNN). At the output of each architecture, a fully connected layer was added with four output neurons that implement a softmax output function. For all these networks it was considered a cross-entropy categorical loss function, an Adam adaptive optimizer and a dropout factor of 0.3 applied at the output layer applied to all the architectures, except in RCNN where the dropout factor was used at 0.2.

Details of each architecture are shown in the following list.

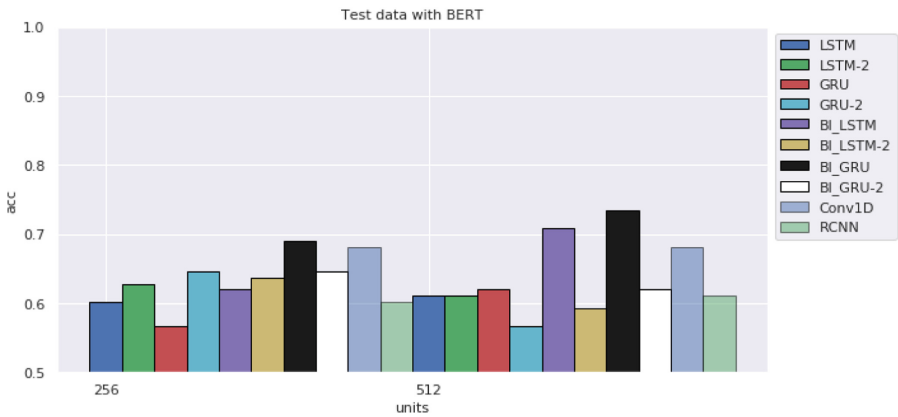
- LSTM: The LSTM considers two variants with 256 and 512 dimensional units, a dropout at the output, and a fully connected layer connected to four output neurons that implement a softmax.
- Stacked LSTM: It has an architecture similar to the LSTM but with two stacked networks, dropout at the output, and a fully connected layer connected to four output neurons that implement a softmax.
- GRU: The GRU considers two variants with units of 256 and 512 dimensions, a dropout at the output, and a fully connected layer connected to four output neurons that implement a softmax.
- Stacked GRU: It has an architecture similar to the GRU but with two stacked networks, dropout at the output, and a fully connected layer connected to four output neurons that implement a softmax.
- Bi-LSTM: The bidirectional LSTM considers two variants with 256 and 512 dimensional units, a dropout at the output, and a fully connected layer connected to four output neurons that implement a softmax.
- Stacked Bi-LSTM: It has an architecture similar to bi-LSTM but with two stacked networks, dropout at the output, and a fully connected layer connected to four output neurons that implement a softmax.
- Bi-GRU: The bidirectional GRU considers two variants with 256 and 512 dimensional units, a dropout at the output, and a fully connected layer connected to four output neurons that implement a softmax.
- Stacked Bi-GRU: It has an architecture similar to bi-GRU but with two stacked networks, a dropout at the output, and a fully connected layer connected to four output neurons that implement a softmax.
- Conv 1D: The 1D convolutional network considers a convolutional layer, followed by a max-pooling type layer, followed by another 1D convolutional layer and another max-pooling layer. At the exit of these four layers, a flatten type layer is connected, a dropout is applied, and a fully connected layer connected to four output neurons that implement a softmax is connected.
- RCNN: The convolutional recurrent neural network considers a convolutional layer followed by a max-pooling type layer, followed by another 1D convolutional layer and another max-pooling layer. The recurrent layer considers LSTM type units and, at the output, a fully connected layer connected to four output neurons that implement a softmax.

## 5 Experimental Results

In this section, we show the results obtained in the experiments with the ten architectures studied on Twitter 16 for rumor detection. The accuracy results on the testing partitions, averaged over the five-folds, are panned according to the text encodings used into two different plots. Each plot shows each of the accuracy bars indicating the dimensionality of the units considered in each model. These results are shown with bars in Fig. 1.



(a) W2V



(b) BERT

**Fig. 1.** Accuracy for test data.

As the results show, the best performance is obtained using word2vec trained on Twitter 16. This result may seem surprising since BERT was pre-trained on a much larger corpus. What this result indicates is that for a specific predictive

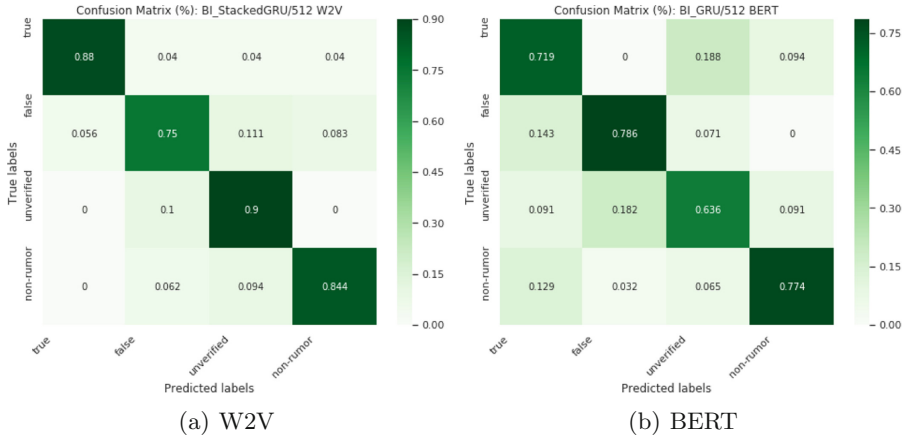
task, training an encoding on the same corpus to be used in the predictive task reduces the vocabulary to the target vocabulary used in the dataset, allowing to define encodings that achieve a better separation and characterization of the dataset texts in particular. This representation will have limitations when compared to BERT, such as the out-of-vocabulary word limitation. Along the same lines, likely, a model trained on a specific dataset instead of a generic corpus has lower generalization capabilities. However, these limitations do not affect the predictive capacity of the architectures in the testing partitions, managing to generalize well in most of the configurations tested on Twitter 16.

Concerning the dimensionality of the internal units of each network, it can be seen that by increasing from 256 to 512 dimensions, some improvements are achieved. It is notable that when using BERT, the increase from 256 to 512 dimensions does not necessarily imply an improvement and in some architectures there is a deterioration in performance. As for the architectures, the results show that more complex architectures, with a higher number of parameters, offer better results. Among these, when considering word2vec, it is observed that the best performance is achieved using a stacked bidirectional GRU with 512-dimensional units. The bidirectional GRU also shows good results using 256-dimensional units. Surprisingly, the performance of the bi-GRU deteriorates when using units with 512 dimensions. As for BERT encodings, the best results are obtained using bidirectional networks, the best result being obtained by a Bi-GRU, followed closely by a Bi-LSTM. These results reinforce the findings of Bugeño *et al.* [4], who had already shown that recurrent bidirectional architectures offered better results than unidirectional ones. Our experiments consistently show that the consideration of left and right contexts around each tweet offers good results in this task.

## 6 Discussion of Results

To understand the performance of the models in each category, the matrices of confusion are shown disaggregated by class for the best models of each encoding. The results are shown in Fig. 2.

The results of the stacked Bi-GRU with 512-dimensional units on word2vec show good performance in all classes. It can be seen that the most challenging class is false rumor, for which the model reaches an accuracy of 0.75. Surprisingly the results in the unverified class are very good, with an accuracy of 0.9. The only class with which the model is confused in this scenario is false rumor. The results of the Bi-GRU with 512-dimensional units on BERT show a deterioration in performance. The most challenging class for this model is unverified, just the opposite of what is got using word2vec. By using BERT, the best performance is obtained in the false rumor class. The performances of both models seem to be complementary, suggesting that the most accessible classes for one model are the most difficult for the other, and vice versa.



**Fig. 2.** Confusion matrix for the best model of each encoding.

To understand these results, we will show the performance of each architecture in each class separately. Our intuition indicates that there should be some architectures that allow some classes to be better detected than others. The results of each architecture disaggregated by class for word2vec and BERT are shown in Figs. 3 and 4, respectively.

The results in word2vec show that the most challenging class is false rumor, where all architectures get lower results. The stacked Bi-GRU is especially useful in detecting unverified rumors, while the RCNN works well in true rumors. It is striking that the RCNN has deficient performance in the non-rumor class. In this last class, the best architectures are the stacked Bi-GRU and the stacked LSTM.

The results using BERT show that the most challenging class is that of unverified rumors. Although in BERT, the results in the false rumor class are poor, the Bi-GRU achieves good results, surpassing those obtained in word2vec. The convolutional neural network obtains good results in the non-rumor class, and in general, all the machines obtain competitive results in the true rumor class.

The previous results show that the results per class are diverse, indicating that some specific architectures are better for each class. It is widely known that when different machines specialize in solving specific classes of a problem, it is suggested to assemble them to obtain a better overall result. Our findings indicate that the use of committee machines would offer good results in this problem.



Fig. 3. Performance per class using word2vec.



Fig. 4. Performance per class using BERT.

To illustrate the variety of results by class, in Fig. 5, we highlight the best performances recorded in our experiments. The figure separates the results for both text encodings (word2vec and BERT), showing in each column the results obtained in a particular class. Each row shows the results obtained by a particular

architecture. The best results are highlighted by class (one per column), which shows that practically all architectures achieve one of the best results in some class.

#### word2vec

|           | true-256 | false-256 | unverified-256 | nonrumor-256 | true-512 | false-512 | unverified-512 | nonrumor-512 |
|-----------|----------|-----------|----------------|--------------|----------|-----------|----------------|--------------|
| LSTM      | 0.80     | 0.722     | 0.80           | 0.750        | 0.80     | 0.639     | 0.60           | 0.750        |
| LSTM-2    | 0.80     | 0.722     | 0.70           | 0.844        | 0.80     | 0.528     | 0.75           | 0.750        |
| GRU       | 0.80     | 0.639     | 0.80           | 0.781        | 0.80     | 0.556     | 0.80           | 0.812        |
| GRU-2     | 0.76     | 0.694     | 0.65           | 0.750        | 0.80     | 0.667     | 0.70           | 0.781        |
| BI LSTM   | 0.80     | 0.583     | 0.55           | 0.781        | 0.76     | 0.722     | 0.70           | 0.812        |
| BI LSTM-2 | 0.80     | 0.639     | 0.80           | 0.750        | 0.84     | 0.667     | 0.80           | 0.812        |
| BI GRU    | 0.88     | 0.639     | 0.90           | 0.781        | 0.84     | 0.500     | 0.80           | 0.844        |
| BI GRU-2  | 0.84     | 0.694     | 0.95           | 0.781        | 0.88     | 0.750     | 0.90           | 0.844        |
| CONV1D    | 0.84     | 0.639     | 0.80           | 0.812        | 0.84     | 0.667     | 0.85           | 0.812        |
| RCNN      | 0.84     | 0.722     | 0.75           | 0.719        | 0.92     | 0.722     | 0.80           | 0.625        |

#### BERT

|           | true-256 | false-256 | unverified-256 | nonrumor-256 | true-512 | false-512 | unverified-512 | nonrumor-512 |
|-----------|----------|-----------|----------------|--------------|----------|-----------|----------------|--------------|
| LSTM      | 0.688    | 0.500     | 0.545          | 0.645        | 0.719    | 0.464     | 0.545          | 0.677        |
| LSTM-2    | 0.781    | 0.536     | 0.409          | 0.710        | 0.719    | 0.607     | 0.364          | 0.677        |
| GRU       | 0.594    | 0.571     | 0.409          | 0.645        | 0.781    | 0.536     | 0.545          | 0.581        |
| GRU-2     | 0.688    | 0.643     | 0.591          | 0.645        | 0.656    | 0.607     | 0.455          | 0.516        |
| BI LSTM   | 0.719    | 0.571     | 0.500          | 0.645        | 0.750    | 0.679     | 0.727          | 0.677        |
| BI LSTM-2 | 0.688    | 0.643     | 0.455          | 0.710        | 0.688    | 0.536     | 0.409          | 0.677        |
| BI GRU    | 0.750    | 0.679     | 0.545          | 0.742        | 0.719    | 0.786     | 0.636          | 0.774        |
| BI GRU-2  | 0.750    | 0.643     | 0.455          | 0.677        | 0.781    | 0.571     | 0.455          | 0.613        |
| CONV1D    | 0.750    | 0.571     | 0.591          | 0.774        | 0.688    | 0.750     | 0.591          | 0.677        |
| RCNN      | 0.594    | 0.714     | 0.409          | 0.645        | 0.688    | 0.607     | 0.409          | 0.677        |

Fig. 5. Best results per class. The results are depicted per text encoding.

The results clearly show that some classes are more complicated than others in terms of predictability. The false rumor class is the most difficult to predict, consistently obtaining the least promising results in all experimental configurations tested in this study. The best overall results in this class are obtained using BERT with a BI-GRU architecture. The results also show that in some classes, several architectures achieve similar performances. This is the case of the false rumor class with word2vec, where three architectures achieve the same performance. Although the results in the unverified class are surprisingly good using word2vec, these are very low when using BERT, which shows that there

is a strong dependence between the encoding and the specific class in which the forecast is made. In the case of word2vec, the best results are achieved by a wide variety of architectures. The only architectures that do not achieve a global optimum in a class are GRU, stacked GRU, BI-LSTM, and stacked BI-LSTM. The other six architectures have some optimal results for any of the classes studied. When using BERT, the performances come down, but more architectures reach a global optimum in some particular class. In BERT, all architectures achieve an optimum in some classes. These results indicate that no architecture manages to solve the problem in all classes, but rather some architectures are more suitable for specific instances of the problem. This finding reinforces the idea that the use of committee machines should offer better results across classes in this problem.

## 7 Conclusions

In this work, we have reported experimental results in the detection of rumors using deep learning. We have tested ten different architectures, on two different text encodings (word2vec and BERT). The results show two interesting findings. The first is related to the representation of the text. Our experiments show that performance in a specific dataset improves as text encodings are tuned to the specific dataset in which each machine is trained. Therefore, word2vec outperforms BERT, at the cost of limiting the generalization of the model to other datasets. The second finding has to do with the variety of results between classes. This work shows that some architectures work better in some classes than in others, suggesting that the use of committee machines in this problem should help to obtain better overall results.

We are currently working on extending our methods to Spanish. In addition, we are exploring the use of ensemble learning in these tasks. The use of committee machines in this field should offer advantages over machines that work on a single architecture.

**Acknowledgement.** Mr. Mendoza acknowledge funding from the Millennium Institute for Foundational Research on Data. Mr. Mendoza was also funded by ANID PIA/APOYO AFB180002 and ANID FONDECYT REGULAR 1200211.

## References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**(2), 211–36 (2017)
2. Allport, G.W., Postman, L.: *The Psychology of Rumor* (1947)
3. Bugueño, M., Mendoza, M.: Applying self-attention for stance classification. In: Nyström, I., Hernández Heredia, Y., Milián Núñez, V. (eds.) *CIARP 2019. LNCS*, vol. 11896, pp. 51–61. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-33904-3\\_5](https://doi.org/10.1007/978-3-030-33904-3_5)

4. Bugueño, M., Sepulveda, G., Mendoza, M.: An empirical analysis of rumor detection on microblogs with recurrent neural networks. In: *Social Computing and Social Media. Design, Human Behavior and Analytics - 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, 26–31 July 2019, Proceedings, Part I*, pp. 293–310 (2019)
5. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, 28 March–1 April 2011*, pp. 675–684. ACM (2011)
6. Castillo, C., Mendoza, M., Poblete, B.: Predicting information credibility in time-sensitive social media. *Internet Res.* **23**(5), 560–588 (2013)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186 (2019)
8. Friggeri, A., Adamic, L.A., Eckles, D., Cheng, J.: Rumor cascades. In: *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, 1–4 June 2014*
9. Hannak, A., Margolin, D., Keegan, B., Weber, I.: Get back! you don't know me like that: The social mediation of fact checking interventions in Twitter conversations. In: *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, 1–4 June 2014*
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
11. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y.: Prominent features of rumor propagation in online social media. In: *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, 7–10 December*, pp. 1103–1108. IEEE (2013)
12. Ma, J., et al.: Detecting rumors from microblogs with recurrent neural networks. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016*, pp. 3818–3824 (2016)
13. Ma, J., Gao, W., Wei, Z., Lu, Y., Wong, K.-F.: Detect rumors using time series of social context information on microblogging websites. In: *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, 19–23 October*, pp. 1751–1754. ACM (2015)
14. Ma, J., Gao, W., Wong, K.-F.: Detect rumors in microblog posts using propagation structure via kernel learning. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, 30 July–4 August 2017, Volume 1: Long Papers, vol. 1*, pp. 708–717 (2017)
15. Ma, J., Gao, W., Wong, K.-F.: Detect rumor and stance jointly by neural multi-task learning. In: *Companion of the The Web Conference WWW 2018, Lyon, France, 23–27 April 2018*, pp. 585–593. *International World Wide Web Conferences Steering Committee* (2018)
16. Ma, J., Gao, W., Wong, K.-F.: Rumor detection on twitter with tree-structured recursive neural networks. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, 15–20 July 2018, Volume 1: Long Papers, vol. 1*, pp. 1980–1989 (2018)



17. Ma, J., Gao, W., Wong, K.-F.: Detect rumors on Twitter by promoting information campaigns with generative adversarial learning. In: The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, 13–17 May 2019, pp. 3049–3055 (2019)
18. Mendoza, M., Poblete, B., Castillo, C.: Twitter under crisis: can we trust what we RT? In: Proceedings of the 1st Workshop on Social Network Mining and Analysis, SOMA, Washington, USA, 2010, pp. 71–79 (2010)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held 5–8 December 2013, Lake Tahoe, Nevada, United States, pp. 3111–3119 (2013)
20. Rath, B., Gao, W., Ma, J., Srivastava, J.: From retweet to believability: Utilizing trust to identify rumor spreaders on Twitter. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, 31 July–03 August 2017, pp. 179–186. ACM (2017)
21. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on Sina Weibo by propagation structures. In: 31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, 13–17 April 2015, pp. 651–662. IEEE (2015)
22. Zhao, Z., Resnick, P., Mei, Q.: Enquiring minds: Early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, 18–22 May 2015, pp. 1395–1405. International World Wide Web Conferences Steering Committee (2015)
23. Zubiaga, A., Liakata, M., Procter, R., Hoi, G.W.S., Tolmie, P.: Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS ONE **11**(3), e0150989 (2016)



# The Role of Moral Receptors and Moral Disengagement in the Conduct of Unethical Behaviors on Social Media

Christian W. Scheiner<sup>1,2</sup>(✉)

<sup>1</sup> Institute of Entrepreneurship, and Business Development,  
Universität zu Lübeck, Lübeck, Germany  
christian.scheiner@uni-luebeck.de

<sup>2</sup> Christian-Albrechts-Universität zu Kiel, Kiel, Germany

**Abstract.** Within the last years, the perception of social media has dramatically changed in public opinion as the dark side of social media has become more and more visible. Previous research has tried to explain unethical behavior on social media with intrinsic and extrinsic motives. The main goal of this study is to develop and provide a conceptual model that links moral receptors with moral disengagement in the context of unethical behavior on social media. For that reason, a set of propositions are developed linking the moral receptors “harm/care,” “fairness/reciprocity,” “in-group/loyalty,” “authority/respect,” and “purity/sanctity” with moral disengagement in order to explain immoral conduct on social media.

**Keywords:** Moral receptors · Moral foundations · Moral foundation theory · Unethical behavior · Moral disengagement · Social media

## 1 Introduction

Within the last years, the perception of social media has dramatically changed. While social media seemed to be a promising instrument to bring people together, it is nowadays often seen in the general public as an instrument to divide people from each other, to increase differences between groups, and as a place where people tend to show unethical behavior to a higher degree than in real life.

Given the acceptance and diffusion of social media within the general population, it is safe to assume that there is no bias in the composition of social media users with respect to subclinical dark personality traits. Hence, personality cannot serve as a rationale in order to explain the tendency to unethical conduct.

Previous research has suggested and examined the role of moral disengagement in order to explain why decent people conduct malign behavior in general (e.g. [1–5]) and in social media in specific (e.g. [6]). Hence, researchers have started to shed light on the mechanisms which enable people to disengage from their own moral standards without feeling any pain or regret.

Often, extrinsic and intrinsic motives served in this context as a starting point to trigger unethical behavior. However, even if these motives illustrate a sound base for examining unethical behavior, they cannot fully explain what triggered this behavior. Therefore, a conceptual framework is developed within this paper as an attempt to explain this phenomenon. This paper shall also serve as a basis for discussion, which is shifting attention towards the role of moral receptors in the conduct of unethical behavior on social media.

## 2 Theoretical Background

### 2.1 Social Media

Statista [7] estimated that 2.48 billion people were using social networks in 2017. The most popular social network in 2019 was Facebook, with more than 2.414 billion active users, followed by YouTube with 2 billion active users, WhatsApp 1.6 billion active users, Facebook Messenger 1.300 billion active users, and WeChat with 1.133 billion active users [8]. The penetration rate in January 2019 was 45% on average globally and ranged from 70% in Eastern Asia to 7% in Middle Africa [9]. The average daily social media usage of users worldwide amounted to 136 min per day [10]. Given these figures, social media is a global phenomenon that has become an important element of the daily lives of almost a third of the total global population.

Despite its widespread use, different understandings exist what social media is. Following Kaplan and Haenlein [11], social media “is a group of internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the exchange of User Generated Content” (p. 61). According to the Organisation for Economic Cooperation and Development (OECD), [12] user-generated content or user-created content possess three decisive characteristics: First, the work is published (e.g., on a publicly accessible online medium). Second, a person has also invested a certain amount of creative effort in order to create content or adapt existing work to create something new. Third, the content is created outside of professional practices and routines.

### 2.2 Social Media and Unethical Behavior

Social media is often associated with democratizing information access and exchange, offering almost unlimited interactivity, and being open to anyone regardless of gender, race, social status, educational level, age, or other characteristics [6, 13, 14]. Respondents stated in a global survey that social media had eased communication, increased freedom of expression, and offered better access to information. At the same time, respondents pointed out that social media had impacted personal privacy negatively, has led to a polarization in politics, and has been perceived as a distraction (Statista 2020c). Hence, social media blend good and bad elements within themselves. The distinction between moral, morally questionable, and immoral cannot easily be made as its assessment is often based on subjective opinion and resides therewith in the perception and judgment of the individual. The evaluation is further aggravated as its impact is not limited to online activities but influences life in general, and the behavior of people offline [10].

Based on the honeycomb framework of Kietzmann et al. [15] that describes the functionalities of social media according to the building blocks (1) conversations, (2) sharing, (3) presence, (4) relationships, (5) reputation, (6) groups, and (7) identity, Baccarella et al. [16] developed a framework in order to describe the dark side and morally problematic side of social media. They argue and provide evidence that each single building block possesses a negative flipside.

With respect to the first building block, users can show aggressive, misinforming, and misleading behavior in conversations (conversations). It is furthermore possible to share and distribute inappropriate content (sharing). The availability and location of users can be tracked without their awareness and used without their consent (presence). The relationship functionality on social media can be used in a detrimental way (e.g., cyberbullying, online harassment, or stalking) (relationships). It is furthermore possible to harm the reputation of others or to be affected personally by detrimental activities (reputation). The possibility to create and become a member of a group on social media can lead to in-group/out-group biases (groups). All previously mentioned functionalities on social have a direct effect on the identity of users. Baccarella et al. [16] argue, therefore, that “social media users are not in control of their own identity any longer” (p. 435), which can lead to a multitude of safety and privacy risks. Figure 1 summarizes and displays

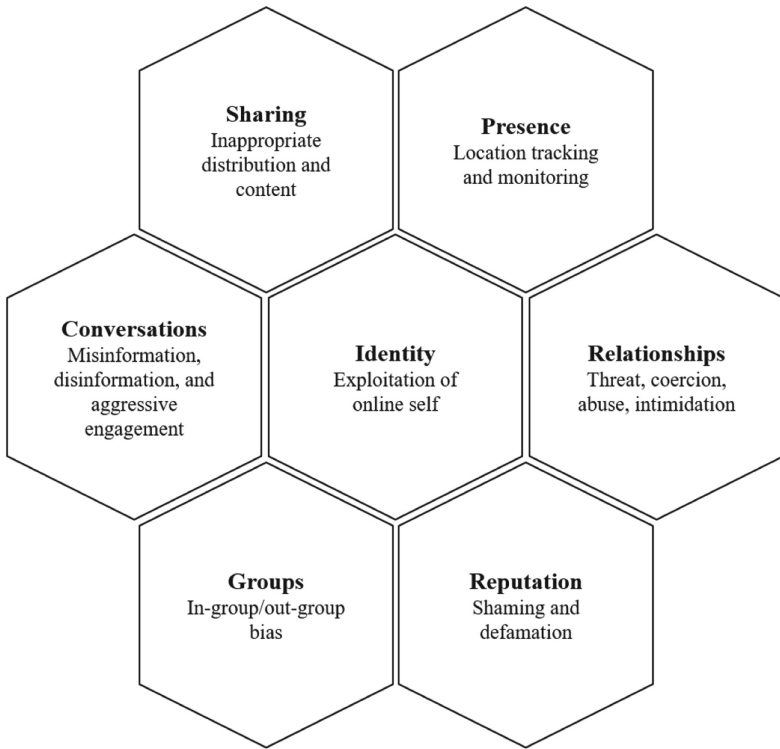


Fig. 1. The dark side of social media [16].

the social media functionalities as a source for unethical behavior or, as Baccarella et al. [16] coined it “the dark side of social media.”

### 2.3 Approaches to Morality in Moral Psychology

In order to understand individual ethics, different approaches and research streams can be identified. The most popular and relevant of which will be described in the following.

**Cognitive Development Theory.** The most popular and influential theory is the cognitive moral development theory of Kohlberg [17–19]. Based on a longitudinal study on children and young adults, Kohlberg [17] proposed that the relative moral development is influencing the ethical justification and moral reasoning of individuals.

The cognitive development theory sees three levels of cognitive development with two stages within each level [17]. In the pre-conventional level, individuals try to maximize their gains and to minimize their losses. As a consequence, the question of right or wrong is answered by the expectations of rewards or punishments, respectively. Individuals on stage one are mainly guided by obedience without any critical, reflective consideration. At stage two, reciprocity and fairness considerations play a major role in the moral evaluation of behavior. In the conventional level, moral judgement is first centered on the acceptance by socially important people (esp. family) and afterward replaced by the intention to preserve social order. Hence, the evaluation of being moral is determined on that stage by the degree of following rules, regulations, and guidelines. In the post-conventional level, individuals broaden their perspective and look beyond the most direct social relationships. Self-chosen principles serve as a basis for defining what is right or wrong. The role of conventional authorities loose subsequently on importance. The cognitive development theory distinguishes on that level between the stage where individuals are aware of the relative nature of personal values and norms and the stage where individuals apply self-chosen ethical principles of justice and human rights. Stage six illustrates the terminal stage of moral development. Kohlberg believed that this stage was achieved only by a small group of individuals. In contrast to the first five stages, the sixth stage was not supported by his data.

In the cognitive development theory, stages are understood as “structured wholes.” Individuals are expected to behave within these “structured wholes” in a consistent way. The stages are, in addition, hierarchically integrated and built upon each other. As a consequence, individuals apply intellectual tools in later stages of moral development, which have been developed in previous stages [20, 21]. Changes in moral reasoning are triggered by perceived discrepancies between the actual level and the next higher one [22, 23].

**Moral Foundation Theory.** Based on the premise that morality is partly inborn and partly learned, Haidt and Josephs [24, 25] and Haidt and Graham [26] developed the moral foundation theory. The moral foundation theory sees innate but modifiable moral foundations as the base for “parents and other socializing agents [...] to build on as they teach children their local virtues, vices, and moral practices” ([27], p. 1030). It is important to highlight that the moral foundation theory speaks explicitly of virtues instead of values “because of its narrower focus on morality and because it more strongly suggests cultural learning and construction” ([27], p. 1030).

In line with the virtue theory, virtues are traits and skills [25]. They are traits in the form of “dynamic patternings of perception, emotion judgement, and action” ([25], p. 386). They are skills - or, more precisely, social skills - as possessing a virtue means “to have extended and refined one’s abilities to perceive morally relevant information so that one is fully responsive to the local sociomoral context” ([25], p. 386).

In the moral foundation theory, morality has its origin in human evolution and is then refined over time [24]. Haidt and colleagues adopt in this argument the innateness view of Gary Marcus [28] as genes provide the first draft for morality, and experience later edits this draft [25]. Individuals are subsequently born or at least prepared with a certain intuition of acceptable and unacceptable behavior towards other human beings. Haidt and Joseph [24] use the term intuitions to particularly emphasize that moral judgment occurs mostly in an unconscious, rapid, automatic, and effortless way. Moral judgment resembles therewith aesthetic judgment as “an evaluate feeling (like-dislike, good-bad) [...] without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion” ([29], p. 188).

While this view contradicts the basic assumptions of cognitive-developmental research, supporting evidence for this view can be found in neuroscience [30], social psychology [31], and primatology [19, 32]. The moral foundation theory takes therewith a cross-disciplinary approach, which is also mirrored in its underlying, functionalist definition of morality [19]. Morality is understood in the moral foundation theory as moral systems that “are interlocking sets of values, practices, institutions and evolved psychological mechanisms that work together to suppress or regulate selfishness and make social life possible” ([19], p. 70).

In order to identify the moral foundations, Haidt and Joseph did not choose an empiricist approach but took a modified nativist approach instead where they were searching for universal, cultural variable building blocks of morality as well as blocks that could be found in other primates [24].

They used a scoring system to identify core building blocks and identified four initial moral foundations: Suffering/compassion, hierarchy/respect, reciprocity/fairness, and purity [24]. Suffering and compassion can be traced back to the relatively long time span until humans are self-sufficient and not fully dependent on others (esp. the mother). The ability to detect signs of suffering and distress early in one’s own child improved the likelihood of survival and secured an evolutionary advantage. Hierarchy as a second moral foundation evolved to allow humans the formation and stabilization of social groups. The better an individual could handle the power dynamics within a group, the higher the chance to climb up the hierarchical ladder or to be an accepted part of the group—reciprocity as a third moral foundation was essential to establish cooperation and collaboration with others; especially with non-kin. Being able to engage with others in a cooperative way offered the possibility to escape zero-sum games and to generate better outcomes. Suffering, hierarchy, and reciprocity have in common that they are all answers to social challenges. Purity, however, as a fourth moral foundation, is concerned with the body and bodily activities (e.g., eating, sex, menstruation). Avoiding related, dangerous microbes and parasites improved the likelihood of survival and acceptance within social groups. Haidt and Joseph [24] point out that the moral law of Judaism, Hinduism, Islam, and traditional societies focus to a great extent on regulating purity

and set the foundations for developing an elaborated set of rules and norms for bodily functions and practices. Violations against those rules and norms triggered, therefore, over time self-sanctioning as well as social sanctioning. Table 1 summarizes the initial four moral foundations.

**Table 1.** Four moral foundations and the emotions and virtues associated with them [24].

|                                   | Suffering                                     | Hierarchy  | Reciprocity  | Purity  |
|-----------------------------------|---|--|--|---|
| Proper domain (original triggers) | Suffering and vulnerability of one's children | Physical size and strength, domination, and protection | Cheating vs. cooperation in joint ventures, food sharing | People with diseases or parasites, waste products |
| Actual domain (modern examples)   | Baby seals, cartoon characters                | Bosses, gods   | Marital fidelity, broken vending machines                | Taboo ideas (communism, racism)                   |
| Characteristic emotions           | Compassion                                    | Resentment vs. respect/awe                             | Anger/guilt vs. gratitude                                | Disgust   |
| Relevant virtues                  | Kindness, compassion                          | Obedience, deference, loyalty                          | Fairness, justice, trustworthiness                       | Cleanliness, purity, chastity                     |

Haidt and Joseph [25] further developed the moral foundation theory and refined the moral foundations. Table 2 summarizes the five moral foundations “harm/care”, “fairness/reciprocity”, “in-group/loyalty”, “authority/respect”, and “purity/sanctity”. The first foundation, “harm/care,” resembles the former “suffering” and stands for the need to care for and protect the vulnerable (esp. own children) and the weak. In actual practice, this foundation is triggered by a multitude of things connected to (potential) harm. With respect to reciprocity as second moral foundation, Haidt and Joseph [25] added that it comes with “a suite of cultural products, such as virtue and vice words related to fairness, religious injunctions about reciprocity, cultural constructs such as rights, and social institutions related to justice” (p. 383). “In-group/loyalty” as a third moral foundation is similar to the former element “hierarchy” but takes a broader perspective. Instead of focusing solely on the power structure of a social group, the tendencies to form social groups and to compete with other groups are emphasized. At the same time, the importance of in-group/out-group effects are highlighted. “Authority/respect” as the fourth moral foundation has its origin also in the former foundation “hierarchy”.

It concerns the social and psychological side effects of life in a social group with dominance hierarchies. While superiors have an obligation to protect their subordinates from external threats, subordinates have to show respect and deference to their superiors. Concerning “purity/sanctity” as a fifth moral foundation, Haidt and Joseph [25] kept the core unchanged but further elaborated that purity contributed in some societies to the ideas about sacredness, about overcoming carnal desires, and about the perception that the body is to be treated as a temple. Subsequent virtues are temperance, chastity, or piety, and lust, or intemperance as vices. Haidt and Joseph [25] highlight furthermore

**Table 2.** The five foundations of intuitive ethics [25].

|  | Harm/Care  | Fairness/Reciprocity                                     | In-group/Loyalty   | Authority/Respect                               | Purity/Sanctity   |
|--|--|--|--|---|---|
| Adaptive challenge                                   | Protect and care for young, vulnerable, or injured kin | Reap benefits of dyadic cooperation with non-kin         | Reap benefits of group cooperation                       | Negotiate hierarchy, defer selectively          | Avoid microbes and parasites                                  |
| Proper domain (adaptive triggers)                    | Suffering, distress, or threat to one's kin            | Cheating, cooperation, deception                         | Threat or challenge to the group                         | Signs of dominance and submission               | Waste products, diseased people                               |
| Actual domain (examples of modern triggers)          | Baby seals, cartoon characters                         | Marital fidelity, broken vending machines                | Sports teams one roots for                               | Bosses, respected professional                  | Taboo ideas (communism, racism)                               |
| Characteristic emotions relevant virtues (and vices) | Compassion   | Anger, gratitude, guilt                                  | Group pride, belongingness, rage at traitors             | Respect, fear                                   | Disgust   |
| Relevant virtues (and vices)                         | Caring, kindness, (cruelty)                            | Fairness, justice, honesty, trustworthiness (dishonesty) | Loyalty, patriotism, self-sacrifice (treason, cowardice) | Obedience, deference (disobedience, uppitiness) | Temperance, chastity, piety, cleanliness (lust, intemperance) |



that purity “is often deeply moralized, not only as a concern about the self but also in the form of beliefs and feelings about groups and the word as a whole” (p. 384).

The moral foundations act not independent from each other but are often triggered simultaneously and cause together certain responses. When purity is, for instance, prominent in a social group to the extent that it is not only a concern but rather an obsession, it can lead to extreme violence, especially when it occurs in combination with the in-group and loyalty foundation [25]. Historical examples of ethnic cleansing serve as a deterrent and horrific examples.

## 2.4 Moral Disengagement

According to the social cognitive theory, the moral agency operates as a self-regulatory system through the self-monitoring subfunction, the judgmental subfunction, and the self-reactive subfunction [1]. People subsequently monitor their (potential) behavior constantly and evaluate it against their own moral standards and situational circumstances. In case the (anticipated) behavior is in accordance with own moral standards, a positive self-reaction occurs. If anticipated behavior contradicts its own moral standards, a negative self-reaction in the form of self-sanctioning is triggered.

It is, however, important to point out that this self-regulatory function does not create a fixed control mechanism [33]. It operates only when it is activated. As a consequence, the same behavior can sometimes be followed by self-reward or self-punishment [33]. This allows individuals to perform a behavior in violation of their own moral standards without any feeling of guilt [1]. Moral disengagement describes the condition when the control mechanism is deactivated.

It is important to emphasize that moral disengagement is not restricted to exceptional incidents but can occur in ordinary situations where ordinary people show self-serving conduct, which causes harm to others [34].

Bandura [33] distinguishes eight moral disengagement practices. These practices are located at different points in the self-regulatory process and focus on the reprehensible conduct, the cause-effect relationship, the detrimental effects, or the victim. Moral justification, palliative comparison, and euphemistic labeling focus on reprehensible conduct. Immoral conduct is subsequently not evaluated as such and shown conduct is justified in the service of moral ends.

Displacement of responsibility and diffusion of responsibility impede that personal responsibility for detrimental effects is acknowledged. When individuals do not see themselves as actual agents of their conduct, displacement of responsibility occurs. Diffusion of responsibility allows the attribution of harm to the behavior of others.

Minimizing, ignoring, or misconstruing the consequences focus on the detrimental effects of immoral conduct. As a result, people ignore or misinterpret the detrimental consequences of their actions.

Dehumanization and attribution operate at the point of the victim. If the victim is not considered as a human being and therewith as equal, different moral standards become acceptable, and malign behavior can be justified (dehumanization). Wrongdoing can also be ignored if the victim is made responsible for the shown conduct (attribution of blame). In this case, the victim has provoked the action and is responsible for bringing suffering upon herself/himself. Figure 2 summarizes the moral disengagement mechanisms.

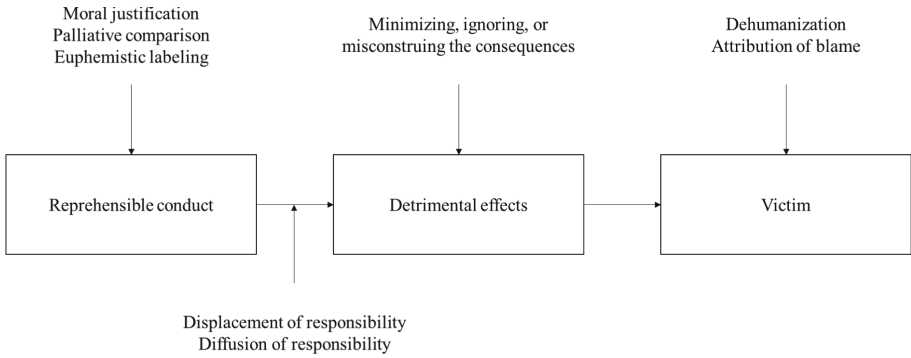


Fig. 2. Moral disengagement mechanisms [33].

### 3 Propositions and Conceptual Model Development

Previous research has provided evidence supporting the relationship between moral disengagement and unethical conduct. Baron et al. [4] found in the context of entrepreneurial behavior that moral disengagement is positively related to unethical behavior. Kish-Gephart et al. [35] showed that personal gain could trigger morally disengaged reasoning. Moore et al. [5] demonstrated that moral disengagement is linked to different unethical conduct in organizations such as fraud or self-serving decisions. Scheiner et al. [36] showed that moral disengagement occurs in the context of unethical behavior in idea competitions.

Next to the relationship between moral disengagement and unethical behavior, these studies are a reminder that context has to be considered when unethical behavior is examined.

As social interaction on social media is a subgroup of computer-mediated communication, the behavior of users on social media is affected by anonymity, asynchrony, and dissociative imagination.

When people can act anonymously, they can easily suspend, ignore, minimize, or misconstrue the detrimental effects of their conduct. Users find, in addition, an opportunity to separate their offline selves from their online actions [37].

Asynchrony can lead to reduced availability of social cues [37]. Being isolated from or experiencing only a small percentage of social cues, people

“feel safe from surveillance and criticism. This feeling of privacy makes them feel less inhibited with others. It also makes it easy for them to disagree with, confront, or take exception to others’ opinions” ([38], p. 48-49).

Dissociative imagination stands for the belief that the online world follows a different set of rules than the offline world [37]. The dissociation is grounded in the possibility to simply escape from it by turning the computer off or closing a browser or an app [37].

Past research on social media has been able to show that social media possesses, in addition, specific characteristics that make it easier for people to act in an unethical way. Baccarella et al. [16] draw attention to the “shallowing hypothesis” where some social

media activities such as sharing and conversing can lead to reduced reflective thinking and increased quick and superficial reasoning, which can cause moral triviality. So-called firestorms are another context-related characteristic where a large group of people often expresses unfounded and highly emotional opinions against a person, company, or group in an offending and immoral way [39]. As the responsibility of the individual can be diffused, normally unacceptable behavior can be conducted without any self-sanctioning.

Based on these findings, it can be assumed that moral disengagement is positively related to the tendency to show unethical behavior on social media.

Proposition 1: Moral disengagement is positively related to the tendency to show unethical behavior on social media.

Moral foundations are an innate moral preparedness which is then refined by experience. Haidt and Joseph [25] use the analogy of “taste buds” on the tongue to illustrate the functioning of moral foundations. While the taste buds on the tongue collect perceptual stimuli with respect to food and drinks, the moral foundations respond to abstract, conceptual stimuli (e.g., cruelty, dishonesty, disobedience, kindness, fairness, treason, cheating, cowardice, etc.). They are, in that sense, moral receptors.

When a moral receptor senses a stimulus, an affectively valenced experience is made. This experience is guiding the subsequent decision about how to react to this stimulus. In case this control mechanism leads to a positive experience (like), the object/agent in question is approached. If, however, a negative experience (dislike) is made, the object/agent in question is avoided [25].

At the same time, when a moral receptor senses a stimulus, a reaction can be triggered against the source of the stimulus. The reaction can be positive if the stimulus corresponds with the linked virtues of a single moral receptor or, rather, all affected moral receptors. In these cases, it is highly unlikely that people deactivate their self-regulation system to disengage from their own moral standards morally. Hence, it can be assumed that moral receptors are negatively related to moral disengagement.

Proposition 2a: Moral receptors are negatively related to moral disengagement.

If the stimulus is, however, in contradiction to virtues or corresponds with vices, a negative reaction can occur against the source of the stimulus. The moral receptor “purity/sanctity” contributes, for instance, in some societies and religious groups to a perception that the body is to be treated as a temple. Hence, the display of the body on social media in general, the presented treatment of the body (e.g., plastic surgery, tattoos), or the shown relationship with sexuality in social media posts may be perceived as a violation of the related virtues and as vices. As a consequence, people morally disengage from their own moral standards. Feeling provoked, they react to these actions in a highly emotional outburst and attribute the blame for this unethical conduct to the victim. The treatment of whistleblowers (e.g., Chelsea Manning, Edward Snowden) on social media is another example where a triggered moral receptor – in this case, the moral receptor “in-group and loyalty” – has led to similar emotionally charged reactions where people disengaged from their moral standards.

There are, in addition, circumstances and occurrences where the moral sense is not only violated but where something is perceived and evaluated as evil. Evil

“aren’t simply people or things that go against the foundation concerns, like vices. Evil is something more, something that threatens to hurt, oppress, betray, subvert, contaminate, or otherwise profane something that is held sacred” ([40], p. 17).

Evil is opposing the sacred. Fighting evil can, therefore, cause the highest unethical behavior; for instance, in form idealistic violence [40]. Table 3 shows the five moral foundations in relation to sacredness and evil.

Each moral foundation can, subsequently, be a source for unethical behavior on social media if something is judged as wrong or even evil. In light of these arguments, it can be assumed that moral receptors are also positively related to moral disengagement.

Proposition 2b: Moral receptors are positively related to moral disengagement.

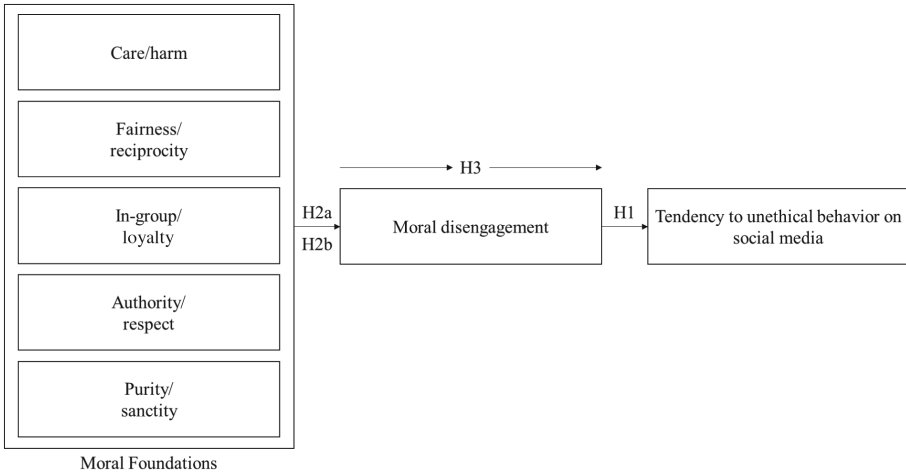
Together, Proposition 1 and 2a/b suggest an influence of moral receptors on the relationship between moral disengagement and the tendency to show unethical behavior on social media.

Proposition 3: Moral disengagement mediates the relationship between moral receptors and the tendency to unethical behavior on social media.

Figure 3 summarizes the propositions and the conceptual model.

**Table 3.** Moral foundations in relation to sacredness and evil [40]

| Foundation | Sacred values                         | Sacred objects  | Evil   | Examples of idealistic violence                             |
|------------|---------------------------------------|---|--|---|
| Harm       | Nurturance, care, peace, fairness     | Sacred objects innocent victims, nonviolent leaders     | Cruel and violent people                     | Killing of abortion doctors, underground bombings           |
| Fairness   | Justice, karma, reciprocity           | The oppressed, the unavenged                            | Racists, oppressors, capitalists             | Vengeance, killings, reciprocal attacks, feuds              |
| In-group   | Loyalty, self-sacrifice for the group | Homeland, nation, flag, ethnic group                    | Traitors, outgroup members and their culture | Ethnic grudges, genocides, violent punishment for betrayals |
| Authority  | Respect, tradition, honor             | Authorities, social hierarchy, traditions, institutions | Anarchists, revolutionaries, subversives     | Right-wing death squads, military atrocities, Abu Ghraib    |
| Purity     | Chastity, piety, self-control         | Body, soul, sanctity of life, holy sites                | Atheists, hedonists, materialists examples   | Religious crusades, genocides, killing abortion doctors     |



**Fig. 3.** Propositions and conceptual model.

## 4 Conclusion

The main goal of this study was to develop and provide a conceptual model that links moral receptors with moral disengagement in the context of unethical behavior on social media. For that reason, a set of propositions have been developed linking for the first time the moral foundations “harm/care,” “fairness/reciprocity,” “in-group/loyalty,” “authority/respect,” and “purity/sanctity” with moral disengagement as the deactivation of the self-regulatory system in order to explain immoral conduct on social media.

The conceptual model is rooted in existing research on the moral foundation theory, the social cognitive theory with a specific focus on moral disengagement as well as unethical behavior on social media.

The provided model can serve as a solid starting point for further empirical testing and theoretical discussion. At the same time, it contributes to the development of a theory of immoral behavior on social media.

## References

1. Bandura, A., Barbaranelli, C., Caprara, G.V., Pastorelli, C.: Mechanisms of moral disengagement in the exercise of moral agency. *J. Pers. Soc. Psychol.* **71**(2), 364–374 (1996)
2. Bandura, A.: Moral disengagement in the perpetration of inhumanities. *Pers. Soc. Psychol. Rev.* **3**(3), 193–209 (1999)
3. Bandura, A.: Impeding ecological sustainability through selective moral disengagement. *Int. J. Innov. Sustain. Dev.* **2**(1), 8–35 (2007)
4. Baron, R., Zhao, H., Miao, Q.: Personal motives, moral disengagement, and unethical decisions by entrepreneurs: cognitive mechanisms on the “slippery slope”. *J. Bus. Ethics* **128**(1), 107–118 (2014). <https://doi.org/10.1007/s10551-014-2078-y>

5. Moore, C., Detert, J.R., Treviño, L.K., Baker, V.L., Mayer, D.M.: Why employees do bad things: moral disengagement and unethical organizational behavior. *Pers. Psychol.* **65**(1), 1–48 (2012)
6. Scheiner, C.W., Krämer, K., Baccarella, C.V.: Cruel intentions? – the role of moral awareness, moral disengagement, and regulatory focus in the unethical use of social media by entrepreneurs. In: Meiselwitz, G. (ed.) *SCSM 2016. LNCS*, vol. 9742, pp. 437–448. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-39910-2\\_41](https://doi.org/10.1007/978-3-319-39910-2_41)
7. Statista: Social media - Statistics & Facts (2019). <https://www.statista.com/topics/1164/social-networks/>. Accessed 10 Jan 2020
8. Statista: Most popular social networks worldwide as of October 2019, ranked by number of active users (2020a). <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. Accessed 10 Jan 2020
9. Statista: Global social network penetration rate as of January 2019, by region (2020b). <https://www.statista.com/statistics/269615/social-network-penetration-by-region/>. Accessed 10 Jan 2020
10. Statista: Daily time spent on social networking by internet users worldwide from 2012 to 2018 (in minutes) (2020c). <https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/>. Accessed 10 Jan 2020
11. Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of social media. *Bus. Horiz.* **53**(1), 59–68 (2010)
12. OECD: Participative web and user-created content: Web 2.0, wikis, and social networking. Organisation for Economic Cooperation and Development, Paris (2007). <https://www.oecd-ilibrary.org/docserver/9789264037472-en.pdf?expires=1578651819&id=id&accname=guest&checksum=EC7BE0A763B2A4C69B63BA7508223477>. Accessed 10 Jan 2020
13. Papacharissi, Z.: Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media Soc.* **6**(2), 259–283 (2004)
14. Groshek, J., Cutino, C.: Meaner on mobile: incivility and impoliteness in communicating online. In: *ACM International Conference Proceeding Series*, pp. 1–7 (2016)
15. Kietzmann, J.H., Hermkens, K., McCarthy, I.P., Silvestre, B.S.: Social media? Get serious! understanding the functional building blocks of social media. *Bus. Horiz.* **54**(3), 241–251 (2011)
16. Baccarella, C.V., Wagner, T., Kietzmann, J.H., McCarthy, I.P.: Social media? It's serious! understanding the dark side of social media. *Eur. Manage. J.* **36**(4), 431–438 (2018)
17. Kohlberg, L.: Stage and sequence: the cognitive developmental approach to socialization. In: Goslin, D.A. (ed.) *Handbook of Socialization Theory and Research*, pp. 347–480. Rand McNally, Chicago (1969)
18. Trevino, L.: Moral reasoning and business ethics: implications for research, education and management. *J. Bus. Ethics* **11**(5/6), 445–459 (1992). <https://doi.org/10.1007/BF00870556>
19. Haidt, J.: Morality. *Pers. Psychol. Sci.* **3**(1), 65–72 (2008)
20. Rest, J.R.: The validity of tests of moral judgment. In: Meyer, J., Burnham, B., Cholvat, J. (eds.) *Morat Development: Current Theory and Research*. Lawrence Erlbaum Associates, Potomac (1975)
21. Sridhar, A.B.S., Camburn, A.: Stages of moral development of corporations stages of moral development of corporations. *J. Bus. Ethics* **12**(9), 727–739 (1993). <https://doi.org/10.1007/BF00881386>
22. Turiel, E.: Developmental processes in the child's moral thinking. In: Mussen, P., Langer, J., Covington, M. (eds.) *Trends and Issues in Developmental Psychology*. Holt, Rinehart & Winston, New York (1969)
23. Treviño, L.K., den Nieuwenboer, N.A., Kish-Gephart, J.J.: (Un)ethical behavior in organizations. *Annu. Rev. Psychol.* **65**, 635–660 (2014)

24. Haidt, J., Joseph, C.: Intuitive ethics: how innately prepared intuitions generate culturally variable virtues maps Strangeness. *Daedalus* (Special Issue Human Nature) **133**(4), 55–66 (2004)
25. Haidt, J., Joseph, C.: The moral mind: how five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In: Carruthers, P., Laurence, S., Stich, S. (eds.) *Evolution and Cognition. The Innate Mind. Foundations and the Future*, vol. 3, pp. 367–391. Oxford University Press, Oxford (2008)
26. Haidt, J., Graham, J.: When morality opposes justice: conservatives have moral intuitions that liberals may not recognize. *Soc. Just. Res.* **20**, 98–116 (2007). <https://doi.org/10.1007/s11211-007-0034-z>
27. Graham, J., Haidt, J., Nosek, B.A.: Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* **96**(5), 1029–1046 (2009)
28. Marcus, G.: *The Birth of the Mind*. Basic Books, New York (2004)
29. Haidt, J., Bjorklund, F.: Social intuitionists answer six questions about morality. In: Sinnott-Armstrong, W. (ed.) *Moral Psychology. The Cognitive Science of Morality*, vol. 2, pp. 181–217. MIT Press, Cambridge (2008)
30. Damasio, A.: *Looking for Spinoza*. Harcourt, Orlando (2003)
31. Skitka, L.J.: Do the means always justify the ends, or do the ends sometimes justify the means? A value protection model of justice reasoning. *Pers. Soc. Psychol. Bull.* **28**, 588–597 (2002)
32. Flack, J.C., de Waal, F.B.M.: “Any animal whatever”: Darwinian building blocks of morality in monkeys and apes. In: Katz, L.D. (ed.) *Evolutionary Origins of Morality*, pp. 1–29. Thorverton, Imprint Academic (2000)
33. Bandura, A.: *Social Foundations of Thought & Action – A Social Cognitive Theory*. Prentice Hall, New Jersey (1986)
34. Bandura, A.: Mechanisms of moral disengagement. In: Reich, W. (ed.) *Origins of Terrorism: Psychologies, Ideologies, Theologies, States of Mind*, pp. 161–191. Cambridge University Press, Cambridge (1990)
35. Kish-Gephart, J., Detert, J., Klebe, L., Baker, V., Martin, S.: Situational moral disengagement: can the effects of self-interest be mitigated? *J. Bus. Ethics* **125**(2), 267–285 (2014)
36. Scheiner, C., Baccarella, C., Bessant, J., Voigt, K.-I.: Participation motives, moral disengagement, and unethical behaviour in idea competitions. *Int. J. Innov. Manage.* **22**(4), 1850043-1–1850043-24 (2018)
37. Santana, A.D.: Virtuous or vitriolic: the effect of anonymity on civility in online newspaper reader comment boards. *Journal. Pract.* **8**(1), 18–33 (2014)
38. Sproull, L., Kiesler, S.: *Connections: New Ways of Working in the Networked Organization*. MIT Press, Cambridge (1991)
39. Pfeffer, J., Zorbach, T., Carley, K.M.: Understanding online firestorms: negative word-of-mouth dynamics in social media networks. *J. Market. Commun.* **20**(1–2), 117–128 (2014)
40. Graham, J., Haidt, J.: Sacred values and evil adversaries: A moral foundations approach. In: Mikulincer, M., Shaver, P.R. (eds.) *Herzliya Series on Personality and Social Psychology. The Social Psychology of Morality: Exploring the Causes of Good and Evil*, pp. 11–31. American Psychological Association (2012)



# Catfishing: A Look into Online Dating and Impersonation

Mariah Simmons and Joon Suk Lee<sup>(✉)</sup>

Virginia State University, Petersburg, VA 23806, USA  
msim6933@students.vsu.edu, joonsuk1@acm.org

**Abstract.** This study investigates *catfishing* and online impersonation. *Catfishing* is a relatively new social phenomenon that happens online. The term, *catfishing* is still foreign to many online users. It is still unclear to many people what constitutes *catfishing* and how it is the same or different from *online impersonation* or *phishing*. In this paper, we discuss catfishing and how it relates to other online threats like *online impersonation* and *phishing*. To see how catfishing affects online users, we interviewed sixteen college students who use social media and online dating platforms at a Historically Black College and University. Among the sixteen participants, nine said they were catfish victims, and four said they were online impersonation victims. Three participants said they had catfished other people online. In this paper, we share the stories of catfish and catfish victims. Our findings show that catfishing has affected our participants' social media use and prevented some of them from trying online dating services.

**Keywords:** Catfishing · Online impersonation · Phishing · Online dating · Social media

## 1 Introduction

### 1.1 Catfish

Merriam-Webster defines *catfish* as “a person who sets up a false personal profile on a social networking site for fraudulent or deceptive purposes” [1]. The term *catfish* was initially coined by a documentary film [14] and subsequently popularized by a reality television series titled *Catfish: The TV Show* [9]. The show taps into an emergent yet already prevalent form of online impersonation and presents stories about people who have dated *catfish* online without ever meeting the person in real life. While catfishing and online impersonation can happen without involving online romance, the show mainly focuses on cases around online dating.

Online impersonation materializes in the forms of *phishing* and *catfishing* in that both involve a wrongdoer pretending to be someone or something else. While *phishing* attacks typically lure people in through fake emails, websites and phone calls to steal sensitive personal information such as social security numbers or passwords for further financial exploitation of the victims, *catfishing* occurs when the culprit assumes someone



else's identity, typically by creating false online profiles. *Catfishing* is a bizarre social phenomenon [31] which has numerous social and legal ramifications. It is clear that catfishing is one of the downsides of digitalization and globalization that the internet brings to us [22]. Yet, the term *catfish* is still foreign to many of us. Such a social phenomenon warrants rigorous investigation.

However, as of this writing, the search term, *catfish*, on ACM digital library only returns 51 results. Only two (e.g., [18, 21]) out of the 51 articles seem to be tangentially related to the catfish phenomenon. Even though the eight seasons of the *Catfish* TV show portrayed a good enough number of catfish case stories, TV shows—reality show or not—are always curated and crafted. In other words, the catfish phenomenon is quite under-explored within the HCI community.

## 1.2 Research Questions

To understand how catfish prey on victims online, and to learn how young adults—the only age group in the U.S. that entirely consists of the internet users [4] and is known to be more vigilant on online privacy issues than their elders [29]—perceive and experience catfishing, we conducted an interview study of 16 internet users. These semi-structured interviews probe into internet users' day-to-day social media usages, catfish encounters, and potential catfishing experiences. In addition, we also asked our participants about *phishing* and *online impersonation* in general to see if and how they discern the seemingly interchangeable, yet subtly distinct three terms.

Three of the interviewees identified themselves as catfish and shared stories about how and why they pretended to be someone else. Nine reported that they were catfish victims, while four said they were impersonated victims. Two interviewees got their social accounts hacked. Through analyzing the interview data, we aim to explore (1) the reasons for catfish individuals to impersonate others, (2) the impacts created by catfishing incidences on catfish victims and their perception of online dating and social media, and (3) potential precautions social media users can take to safeguard themselves from catfishing.

## 2 Related Work

Previous catfishing and online dating research studies have looked into analyzing the MTV show *Catfish* (e.g., [9, 14, 30]), online dating platform designs and usage (e.g., [6, 8, 32, 33]), or potential solutions to the legal issues these threats have caused (e.g., [5, 26]). Catfishing has become a big problem on the internet; however, many research projects have failed to acknowledge catfishing and how it can affect its users.

### 2.1 Catfish: The TV Show

The creation of MTV's documentary *Catfish* [14] and the reality television series *Catfish: The TV Show* [9] has shed light on one of the newest forms of threats on the internet. Over the past eight seasons of *Catfish: The TV Show*, the show has revealed a dark side of online dating. All the episodes are created using a documentary-style presentation

and are filmed using personal, everyday handheld cameras, as well as TV production cameras to capture multiple scenes at once. The show is not a scripted TV show, and scenes are captured as they occur in real-time [30]. The episodes include cases about catfish impersonating celebrities. The victims were convinced that they were really in a relationship with these celebrities. The TV show presented the world of catfishing to the public by documenting victims as they seek their catfish to figure out why they were targeted. Catfishing has become a widely known issue of social media, mainly due to its representation on the MTV show.

## 2.2 Research on Catfish

The exposure of catfishing on the MTV show has led scholars to investigate catfishing. Kottemann [15], for instance, explores the notion of “*deliberate deception*” and catfishing in her doctoral dissertation and argues that online identity creation is a rhetorical action, and that understanding catfishing can help to learn modern rhetoric. Lovelock uses the TV show as a case study in investigating the articulation of the self on social media and explores the meaning of online identity and the authenticity of it [16].

Although the show played a key role in exposing the issue to the public, a widely known catfishing incident of a famous American football player, Manti Te’o, also caught national attention (e.g., [15, 23, 26]). Scholars have also started looking into legislative issues with catfishing, creating legal laws against catfishing and online impersonation (e.g., [5, 26]). According to the Better Business Bureau, 1 in 7 online profiles are believed to be fake, and the number of catfishing victims has increased by more than 50% in 3 years [2]. Catfishing acts are rapidly spreading across online platforms and continuing to cause multiple problems to many users. This necessitates and warrants further investigatory research.

## 2.3 Research on Online Impersonation and Phishing

Online impersonation in the form of a phishing attack became one of the first real threats that people noticed on the internet. According to the Internet Crime Report issued by the U.S. Federal Bureau of Investigation, 20,000 business email phishing attacks resulted in about \$1.2 billion losses [12]. It is hard to believe that so much money was lost in phishing scams, but the data shows that phishing has become one of the biggest threats to companies and online users. For instance, about 83 million Facebook accounts are fake, and many of these fake accounts are still actively used profiles [3]. Previous studies have investigated the issues and threats online impersonation brings to online users and how self-representation and deception are involved with these issues [10].

Previous studies have also investigated lying on online communities in four countries and warns that individuals and organizations need to be aware of deception on online communities [19, 24]. Phishing attacks have evolved over the years and became quite sophisticated [11]. The attackers use social engineering and deceptive techniques to extract sensitive information from the online users and subsequently cause financial harm to the victims (e.g., [11, 13]).

### 3 Methodology

To investigate these online issues, we conducted interviews with catfish, catfishing victims, impersonated victims, and general social media users to understand participants' experiences with online impersonation and catfishing.

#### 3.1 Interview

Interviews were semi-structured and conducted in person. In case participants do not wish to participate in an interview, but still want to share their experiences, an online questionnaire option was provided. Only one participant decided to complete the online questionnaire. Interviews lasted anywhere from 15–45 min. The interviews were audio-recorded and transcribed. We also collected demographic information for each participant prior to the interview through a pre-study questionnaire.

#### 3.2 Participants

We recruited participants from a pool of students from Virginia State University, a historically black public land-grant university in Petersburg, Virginia. In order to recruit participants who, have catfishing experiences, we also used a snowball sampling method [27]. A total of sixteen participants were recruited (7 females, 9 males). The participants' age ranged from 18 to 27 (Mean: 21.3, SD: 2.56). Fifteen participants were students at Virginia State University, and one participant was a technical college graduate. Ten participants (62.5%) identified themselves as African American, two participants (12.5%) as Hispanic/Latino, one (6.25%) as Asian, and three (18.75%) as other. Sixteen participants included three catfish, nine catfish victims, four impersonated victims, two hacked victims, and seven general everyday social media users. Participants received no monetary compensation for participating in the study.

#### 3.3 Interview Data Analysis

We conducted a thematic analysis on the interview data in order to identify patterns of meanings in the participants' responses. The interview data were fully transcribed and read multiple times by the first author. During the initial analysis, both first and second authors looked at the interview data together and developed an initial coding scheme. The first author went through the interview data in multiple iterations to find recurrent occurrences in participants' data about their experiences and knowledge of the topics being studied.

#### 3.4 Interview Data Collection

Our data collection consists of participants recount of their experiences with social media, online dating, online impersonation, phishing, and catfishing. Although we offered a questionnaire, only one participant chose the questionnaire over the in-person interview. The interview questions were grouped into the following themes: online dating, online

impersonation, catfishing, catfishing prevention, and the relationship between online impersonation, catfishing, and phishing. Three sets of interview questions were prepared for three different types of interviewees: general social media user, catfish, or catfishing victim. Some participants fell into more than one type. In such a case, the interviewer used questions from multiple interview question sets. At the beginning of the interview process, the interviewer stated that the interviewees should only share experiences that they feel comfortable with sharing.

## 4 Findings

Conducting multi-phased thematic analysis [25] on the interview data revealed several recurring themes. In this section, we report on some of the interesting findings from our qualitative analysis.

### 4.1 Participants Perception of Online Impersonation, Catfishing, and Phishing

At the beginning of the interviews, we asked participants whether they know of the three terms—online impersonation, catfishing, and phishing. We also asked participants how these three terms are related, similar, or distinct. All but one participant said they knew the terms, but only five participants said the three terms are distinct. Eleven participants said the three terms are somewhat related. Eight participants believed that online impersonation and catfishing are the same, or at least they thought the two terms go hand in hand. This shows that the term “*catfish*” might still be foreign to many people.

When we asked how online impersonation is different from catfishing, five participants said they believed the three terms are different, stating that catfishing involves ill-intention whereas online impersonation is a neutral term. For instance, both P3 and P4 stated that catfishing is to “*get someone*,” and online impersonation is an action without “*a goal in mind*.”

*I assume catfishing they are trying to **get the person**. Online impersonation I'll assume they are trying to impersonate the individual but not really have a goal in mind — P3*

*... catfishing is for like a specific form of online impersonation like you are trying to **get with a person**. Online impersonation you are just pretending to be someone — P4*

When participants did not have clear understandings of the three terms, we gave a brief description of each term and asked them if they could conjecture the logical relations of the three terms by ranking the terms into a hierarchical form. Out of six participants who did put the terms in a hierarchical relation, four put phishing at the top of the hierarchy. P15 stated, “*phishers I feel like they have a whole program. I feel like they get hired to do that because when you get those emails and you get those scams. They look legit sometimes they really look real.*” Two participants stated they believe that online impersonation would be on the top of the hierarchy. P9 stated, “*So with online*

*impersonation, it's like the head of it so you can get to catfishing as in appearing to be somebody else or you can get to phishing by as online impersonation of a donation group.*" It was interesting to see how the participants could and could not categorize and relate the terms with each other.

Despite the prevalence of phishing, catfishing, and online impersonation incidences in our society, our participants did not have a full consensus on the three terms. Many of our participants did not recognize the term, *phishing* until we explained what the term means. Only then the participants responded that they knew it as *online scamming*.

## 4.2 Catfish

Eleven participants were familiar with the Catfish TV show and said that they learn about catfishing from the TV show. Five participants said that they learned about the term from other sources such as the internet or from other people.

When we asked if they were ever involved in catfishing, nine participants responded that they were catfished at least once before. The catfish victims revealed that they had been catfished on various social media platforms: Facebook, Instagram, Tinder, WhatsApp, Snapchat, and an online gaming platform. Facebook, Tinder, and Instagram were the top three platforms on which our participants experienced being catfished.

Out of sixteen participants, three participants said that they had experiences of catfishing online users. Tinder, Hily, Bumble, and Instagram were the platforms that the three catfish participants said that they used to catfish. Tinder was the common platform that all three catfish participants used.

## 4.3 Catfishing Victims

In most cases, catfish victims claimed that they were tricked by fake online profiles on dating apps. For instance, P13 met a female on Tinder and liked her profile pictures. He stated that her profile pictures were *"a 9 out of 10."* So P13 finally decided to invite her over to meet. However, when he met his date, she was not the same person as her profile pictures. He said, *"she looked more like a 4 or 5,"* and that was him being *"generous."* We do not know why P13's date posted fake profile pictures on Tinder. Sometimes, people might post fake profile pictures or extensively photoshopped pictures on dating apps such as Tinder due to security concerns—some might think it is not safe to post real photos. People might post fake or exaggerated profiles pictures for the same reasons as we put on our "best dress" or make-up when we go out on a date. The fact that P13's date showed up to meet with him offline hints to us that her intention of putting her profile pictures might not be so malicious.

On the other hand, we also saw a case in which a catfish pretended to be someone the victim already knew. P7 had been involved in what she believed to be a catfishing relationship since 2012. For eight years, she had been involved with the catfish online and believed that the person was someone she knew from a school she attended before. When the catfish reached out to her to reconnect on Instagram and Facebook, she never doubted his identity because she already met the person in real life before. However, whenever she asked to video chat with him on social media platforms, the catfish avoided showing himself. P7 stated, *"Whenever I wanted to facetime, it was always a text message, it*

*was always I'm busy, or I'm at work.*" Again, we do not know whether the alleged catfish is actually someone who is pretending to be someone else, or the person has legitimate reasons to avoid video chat—the person might be allergic to video chat, the person's phone might have a broken camera, or the person might simply enjoy texting over video chat. There can be a million different reasons. For P7, however, it was enough reason to doubt the catfish's identity and put a stop on their online interactions. This case shows that catfishing is not a simply definable set of behaviors but is still an evolving phenomenon. As technologies get more and more complicated, we might see more variance of catfishing. For instance, with technologies like DeepFake [11], we might not be able to tell a person's true identity even if we interact with the person on an online video chat.

#### 4.4 Three Catfish

Among the sixteen participants, three said they had catfished other people before. We share their accounts of catfishing experiences. To our surprise, P7 was not only a catfishing victim, but she admitted that she catfished others before. For P7, catfishing was a way to scout whether her boyfriend was faithful or not.

*"I thought my boyfriend was cheating on me, so I made a fake account, I made several fake accounts to the point Instagram stopped me from making accounts, they were blocking them"* — P7

*"It was like a thing I felt like I was a uh.. mental thing I had to whine myself off of that, off of checking his Instagram every single day"* — P7

In P7's defense, she had been a catfish, but she did catfish only because she wanted to find out if her boyfriend was being unfaithful. She did not intend to trick anyone else into believing that she was someone else. P7 stated that the Instagram profiles she made were strictly used as a way to track what her boyfriend at the time was doing.

P11 was a male participant, but he said he created a female profile with pictures he gathered on Google. P11 stated that he became a catfish because he was curious to see what type of responses he would receive from creating a profile as an attractive female. He also stated that it was something that he did when he got bored.

*"So I just used some pictures I found on google and I just swiped left on everybody and some of them would pursue me"* — P11

*"Nobody was trying to get to know me... Most of them were about sex which was weird. All of them were funny"* — P11

The participant said he is no longer a catfish because all the conversations became similar or just became dull and boring, which led him to delete the profile. He also stated that since deleting the profile, he believes he will not catfish anyone again.

Our last catfish's motives were completely different from the previous two participants'. P15 stated that she had catfished about 30 or more people on almost all of the

dating platforms and websites to gain monetary benefits. P15 said, “*I made like \$100 in an hour for sending fake images and stuff like that... I really only did it because I was desperate and I needed money.*” The participant also said that she targeted online dating platforms because many of the users on the platforms were “*desperate*” and only were looking for attention or someone to talk to them.

Since P15 stated that she had catfished 30 people, we wanted to see how she was able to trick so many. When we asked, the participant said, “*I would search up places near them, you know Google knows everything.*” She explained that in order for her to be believable to her victims, she made them believe that she was from the area that she claimed to be from. She stated that she had victims in California, Virginia, North Carolina, and even the District of Columbia. For this participant, her goal was never to be in a romantic relationship; rather, she used these online dating platforms as a way to support herself when she was in need of money. She even stated that she had made about \$100 in a day by sending fake images to men throughout these platforms that had sexual intentions. She stated that she doesn’t feel good doing it; however, men should not be using these platforms because they are sexually desperate [20].

Our three catfish participants’ stories confirmed that many of our participants’ assumptions about online dating platforms were indeed true. If not all, at least a good portion of the users on these platforms are believed to be using the platform only for sexual purposes and not for the platforms’ original intended use. For instance, we believe that online dating platforms are not intended as “sexual search tools,” yet many users might just be using these platforms as “sexual search tools.” Many online dating users create profiles on these online dating platforms looking for potential partners. However, they might end up meeting with users who have bad intentions, potential catfish. These catfish have begun pushing away potential, innocent users and possibly hindering the success of these online dating platforms. P3 responded that she does not try online dating because “*I don’t want to meet someone who is pretending to be someone else when they are actually this kind of person.*” Just like P3, many other participants shared the same fears and reasons why they don’t want to use these online platforms.

## 5 Discussion, Limitations, and Conclusion

This paper contributes to the studies of social media and online dating by highlighting how catfishing and online impersonation have become a pervasive threat to many users on various online communities and platforms. Through an interview study, we explored how our participants perceived phishing, online impersonation and catfishing. We shared both catfish and catfish victims’ accounts on catfishing. Catfishing has indeed become a big threat to social media and online dating platform users. We related our findings to other studies that have investigated social media, online dating platforms, and online identity deception.

While we tried to interview as many catfish as we could, our data is limited to the responses we got from three catfish participants. Our findings are constrained to college students at a Historically Black College and University. While social media and online dating platform users consist a wide range of ages and ethnicities, our study only focused on a small group of the users.

Catfishing is still under-explored topic within HCI, and our work is the first attempt to study catfish victims as well as catfish individuals.

## References

1. Catfish: In Merriam-Webster.com Dictionary. <https://www.merriam-webster.com/dictionary/catfish>
2. Cooke, K.: When Love Bites: 2019 Catfishing Numbers by State (2019). <https://www.highspeedinternet.com/resources/states-with-most-catfishing-scams>
3. Clanton, K.: We are not who we pretend to be: ODR alternatives to online impersonation statutes. *Cardozo J. Conflict Resolut.* **16**, 323–356 (2014)
4. Clement, J.: U.S. Internet Usage Penetration 2019 by Age Group (2019). <https://www.statista.com/statistics/266587/percentage-of-internet-users-by-age-groups-in-the-us/>
5. Derzakarian, A.: The dark side of social media romance: civil recourse for catfish victims. *Loyola Los Angeles Law Rev.* **50**(2014), 741–764 (2014)
6. Ellison, N., Heino, R., Gibbs, J.: Managing impressions online: self-presentation processes in the online dating environment. *J. Comput. Mediated Commun.* **11**(2), 415–441 (2006). <https://doi.org/10.1111/j.1083-6101.2006.00020.x>
7. Engler, A.: Fighting deepfakes when detection fails (2019). <https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/>
8. Flug, K.C.: Swipe, right? young people and online dating in the digital age (2016). [https://sophia.stkate.edu/msw\\_papers/578](https://sophia.stkate.edu/msw_papers/578)
9. Hale, M.: There's Always a Catch (2012). <https://www.nytimes.com/2012/12/arts/television/catfish-the-tv-show-with-nev-schulman-exposes-deceit.html>
10. Jeffrey, T., Hancock, C.T., Ellison, N.: The truth about lying in online dating profiles. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007), pp. 449–452. Association for Computing Machinery, New York (2007). <https://doi.org/10.1145/1240624.1240697>
11. Hong, J.: The state of phishing attacks. *Commun. ACM* **55**(1), 74 (2012). <https://doi.org/10.1145/2063176.2063197>
12. Jentzen, A.: The Latest Phishing Statistics and News May 2019. <https://www.proofpoint.com/us/security-awareness/post/latest-phishing-may-2019>
13. Kirda, E., Kruegel, C.: Protecting users against phishing attacks with antiphish. In: 29th Annual International Computer Software and Applications Conference (COMPSAC05) (2005). <http://dx.doi.org/10.1109/compsac.2005.126>
14. Koch, C.M.: To catch a catfish: a statutory solution for victims of online impersonation. *Univ. Colorado Law Rev.* **88**(2017), 233–280 (2017)
15. Kottemann, K.: The rhetoric of deliberate deception: what catfishing can teach us. Dissertation, ProQuest, Ann Arbor, MI (2015)
16. Lovelock, M.: Catching a catfish. *Telev. New Media* **18**(3), 203–217 (2006). <https://doi.org/10.1177/1527476416662709>
17. Lumbres, D.: A qualitative study on friendship: comparing offline and exclusively online experiences. Dissertation. ProQuest, Ann Arbor, MI (2018)
18. Magdy, W., Elkhatib, Y., Tyson, G., Joglekar, S., Sastry, N.: Fake it till you make it. In: proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017- ASONAM 2017 (2017). <http://dx.doi.org/10.1145/3110025.3110075>
19. Marett, K., George, J.F., Lewis, C.C., Gupta, M., Giordano, G.: Beware the dark side: cultural preferences for lying online. *Comput. Hum. Behav.* **75**(2017), 834–844 (2017). <https://doi.org/10.1016/j.chb.2017.06.021>



20. Marwah, E.V.: Understanding how young people experience risk with online-to-offline sexual encounters: a second qualitative phase for the CH@T project. University of South Florida Scholar Commons: Graduate theses and dissertations (2015). <http://scholarcommons.usf.edu/etd/5988>.
21. Masden, C., Edwards, W.K.: Understanding the Role of Community in Online Dating. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems-CHI 2015 (2015). <https://doi.org/10.1145/2702123.2702417>
22. Meerts, A.: How did catfishing come into our society? (2018). <https://www.diggitmagazine.com/articles/catfishing>
23. Nolan, M.P.: Learning to circumvent the limitations of the written-self: the rhetorical benefits of poetic fragmentation and internet ‘catfishing’. *Persona Studies* **1**(1) (2015). <http://dx.doi.org/10.21153/ps2015vol1no1art431>
24. Obada-Obieh, B., Somayaji, A.: Can I believe you? Establishing trust in computer mediated introductions. In: Proceedings of the 2017 New Security Paradigms Workshop (NSPW 2017), pp. 94–106. Association for Computing Machinery, New York (2017). <https://doi.org/10.1145/3171533.3171544>
25. Saldaña, J.: *The Coding Manual for Qualitative Researchers*. SAGE, Los Angeles (2016)
26. Smith, L.R., Smith, K.D., Blazka, M.: Follow me, what’s the harm? Considerations of catfishing and utilizing fake online personas on social media. *J. Legal Aspects Sport* **27**(1), 32–45 (2017). <https://doi.org/10.1123/jlas.2016-0020>
27. Snowball sampling. <https://research-methodology.net/sampling-in-primary-data-collection/snowball-sampling/>
28. The State of Privacy in America (2016). <https://www.pewresearch.org/fact-tank/2016/09/21/the-state-of-privacy-in-america/>
29. Sumter, S.R., Vandenbosch, L., Ligtenberg, L.: Love me tinder: untangling emerging adults’ motivations for using the dating application tinder. *Telematics Inform.* **34**(1), 67–78 (2017). <https://doi.org/10.1016/j.tele.2016.04.009>
30. Van Neygen, E.: *The Representation of Online Daters in Reality TV: An analysis of ‘Catfish: The TV Show’*. Dissertation, University of Ghent, Ghent, Belgium (2015)
31. Waring, O.: What is catfishing and how can you spot it? (2019). <https://metro.co.uk/2018/03/18/catfishing-can-spot-7396549/>
32. Vipobtanaseth, P., Mujtaba, B.: A study of users’ perception of online dating websites. *Int. J. Curr. Sci. Technol.* **5**(7), 457–466 (2017)
33. Zytco, D., Grandhi, S.A., Jones, Q.: Online dating coaches’ user evaluation strategies. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA 2016), pp. 1337–1343. Association for Computing Machinery, New York (2016). <https://doi.org/10.1145/2851581.2892482>



# Riding the Wave of Misclassification: How We End up with Extreme YouTube Content

Christian Stöcker<sup>1</sup>  and Mike Preuss<sup>2</sup>  

<sup>1</sup> HAW Hamburg, Hamburg, Germany  
Christian.Stoecker@haw-hamburg.de

<sup>2</sup> LIACS, Universiteit Leiden, Leiden, The Netherlands  
m.preuss@liacs.leidenuniv.nl

**Abstract.** Content recommendations on big internet platforms such as YouTube are supported by crowd-based recommender systems. Users are presented with what other users in comparable situations spent time on. There are several examples, however, where sequences of suggested content quickly degenerated towards only slightly related, sometimes problematic content that we define as contextually inappropriate. We try to identify the basis of this effect from both the user and the technical side, and set up a simple simulation system in order to better understand the interactions. Simulation results provide evidence that autoplay is an especially problematic feature, but that completely preventing inappropriate suggestions is technically very hard if not infeasible because it is the nature of a recommendation system to take into account feedback from the user and adapt to it. We also propose some possible measures to mitigate the problem.

**Keywords:** Recommender systems · Simulation · YouTube · Contextually inappropriate content

## 1 Introduction

The companies that run the large online platforms serving media content of various kinds to users have long recognized that some content, especially, but not exclusively so-called user generated content (UGC) may be deemed inappropriate. Most large platforms offer mechanisms on their platforms that applies, for example, to Google, which not only runs the most widely used search engine in the western hemisphere but also owns the world's largest and most widely used video platform YouTube [19], and Facebook, the world's most widely used social network and also the owner of photo sharing platform Instagram. The platform companies have also been alerted to the fact that automatic distribution of content can lead to juxtapositions that customers find highly problematic, mainly in the area of advertising: A number of companies have, for example, discontinued advertising campaigns run on platforms after it was discovered that their ads ran

next to or in front of, extremist or otherwise undesirable content [9, 11, 12, 21, 22]. This paper, however, is concerned with a different kind of contextually inappropriate content: The kind that is caused by the serial, automatically generated video viewing recommendations on YouTube. As developers working on the platform have publicly explained, the main goal of the automatic curation system that suggests a video to watch right after a user has finished his or her current video on the platform is to maximize watch time [15], i.e. the total number of minutes and seconds as well as the number of further videos a given user will keep watching.

To further this goal, YouTube introduced a feature called “autoplay” in 2015. This means that the next “suggested video” in the queue produced by the algorithmic curation system of the platform for each user in each session starts automatically if the user doesn’t actively intervene by clicking pause or closing the browser tab or app. The watch time optimization goal can produce collateral damage: It sometimes leads to the promotion of extremist, misleading or otherwise problematic content to a wider audience [19]. YouTube has even been called “The great radicalizer” because of this mechanism [20]. YouTube’s parent company Google has acknowledged that the systems in place can lead to undesirable content being promoted: “When our services are used to propagate deceptive or misleading information, our mission is undermined.” [6] The company has vowed to apply strategies to counter “disinformation and misinformation” across “Google Search, Google News, YouTube, and our advertising products”.

The focus of this paper, however, is a concept more abstract and also broader than disinformation and misinformation. We are addressing a category of content suggested by such systems that we call “contextually inappropriate”. By this we mean content that is problematic not necessarily in and of itself - although disinformation and misinformation videos can generally be considered inappropriate in any context.

The goal of this paper is to develop a computational understanding for the mechanisms underlying such inappropriate content recommendations and thus potentially offer approaches to remedy some of the problems caused by this kind of content (see next section). We will first try to develop a clearer definition of different kinds of contextually inappropriate content, citing recent examples from the literature and also journalistic publications on this issue. We will then proceed to develop a conceptual model of the underlying mechanisms (Sect. 3) that we can simulate by first analysing the recommender system as described in [4] and other sources. After discussing the role of the user in Sect. 4, we describe our simulation model and provide first results in Sects. 5 and 6.

Our working hypothesis (as expressed in the title) is that even if a recommendation system makes few erroneous wrong suggestions that may be inappropriate in a specific context, the user interaction easily propels these to be played out more and more. It appears to be hard to prevent that kind of “unwanted malicious interaction” between user and recommendation system.

## 2 Contextually Inappropriate Content

We define contextually inappropriate content as content that violates the assumptions, intentions and goals of the viewer and/or the uploader of a specific video in the context of the current viewing session. To put it differently: Contextually inappropriate content is content recommended by an automatic curation system that reaches an audience that it is not intended for, or an audience that might be shocked, disturbed, misled or otherwise harmfully impacted by said content. Contextually inappropriate content is problematic for two main reasons: First, the content itself might, in a different context or targeted to a different audience, be entirely harmless and unremarkable. This means that human content raters that assess the appropriateness of content for the platforms have no choice but to leave the content online. Flagging systems and the like thus may not work in these cases. Second, contextually inappropriate content creates a situation where the - potential - harm a certain piece of content does may be confined to a few isolated viewing sessions. It is thus hard to detect and even harder to counter. Both of these reasons are exacerbated by the fact that YouTube is used by many children [13] who might be particularly vulnerable to the kinds of inappropriate content recommended to them. Traditional systems for preventing minors from watching harmful content might be circumvented by this process, as the second example listed below will show. To illustrate the concept of contextually inappropriate content, here are three examples:

- In 2016, former YouTube developer Guillaume Chaslot systematically explored automatically generated recommendations for the search terms “Trump” and “Clinton” in the run-up to the presidential election in the USA. He found that both search terms produced sequences of recommendations that skewed heavily towards “Trump-leaning” clips. He noted that “a large proportion of these recommendations were divisive and fake news”. Chaslot also reported that “a ‘Clinton’ search on the eve of the election led to mostly anti-Clinton videos” [3].
- In 2017, an independent writer [2] and subsequently several news outlets [10] [16] reported that the YouTube recommendation system led underage users towards videos that consisted of “parodies” of well-known cartoon shows for children like “Peppa Pig” or “Paw Patrol”. These “parody” videos contained, to quote the British “Guardian”, “well-known cartoon characters in violent or lewd situations and other clips with disturbing imagery that are occasionally - in a nice postmodern touch - set to nursery rhymes”. Children following the flow of recommendations ended up watching these videos, some reacting disturbed or fearful.
- In 2019, three researchers originally interested in political content on YouTube in Brazil, came across a network of “channels that were sexually suggestive” [7]. When examining those channels more closely, they found a number of sexually suggestive videos featuring “underage women” or adults posing in children’s clothing. Examining the suggestions for such videos in turn led them to “channels featuring videos of small children”, some in swimwear,

some doing gymnastics, “the common theme was that the children were only lightly dressed”. According to the “New York Times” [5], the newspaper the researchers had co-operated with, some of these videos accumulated hundreds of thousands of views within days, from viewers who had obviously been led there “through a progression of recommendations”. When the researchers published their own results as a working paper, they stressed that they specifically decided against a peer reviewed publication. They reasoned that “the children in the videos, nor the families that had uploaded some of the videos could not have possibly waited one year for YouTube to change their algorithm”.

All of these are examples of what we have defined as contextually inappropriate content. Each single video might be entirely legal, harmless and unremarkable, be it one promoting a presidential candidate, a drastic cartoon parody or a clip of children in swimming gear. But the context of the respective viewing session or reached audience makes them inappropriate, either for the audience or for the uploaders. For example, the parents of the children concerned in example three. This points to two of the core problems associated with contextually inappropriate content promoted by automated recommender systems: What emerges as problematic is hard to predict, but, once uncovered, it often seems to warrant immediate intervention. What is hard to gauge is the role that the behavior of individual users plays in interaction with the recommender systems. Certain very active subgroups of users rallying around certain types of content, be they pro-Trump videos, drastic cartoon parodies or clips of lightly-dressed children, seem to contribute extreme outsized signals that influence the recommendations generated by the system if not explicitly toned down. These recommendations in turn make the content available to more users. Some of these users might keep watching for entirely different reasons, be it curiosity, shock, alarm or any number of other motivations.

The aim of this paper is to in principle uncover the mechanisms that lead to the recommendation of contextually inappropriate content in automated recommender systems and discuss possible measures to mediate this problem.

In the next section (Sect. 3), we first investigate if the YouTube recommender system works in a way that it “drives” users to more extreme content. Based on the available material (the most recent description is the work of Covington, Adams, and Sargin [4]) we conclude that this is not the case.

Is the problem described above thus the users’ fault? The contributions of the users are much harder to judge, starting from the fact that we do not have a good description of the mechanisms inside *the user* that enables us to deduce if the user unintentionally ignites the process of degeneration. Users are of course not a homogeneous group, but rather a very diverse and very large population. Even a single user can interact with the system in a very different fashion depending on the environment or situation.

Our main hypothesis is that the injection of contextually inappropriate content described above mainly stems from classification errors the recommender system makes which are in turn amplified by more explorative users.

### 3 The Role of the Recommender System

Whereas we try to take into account what we know of the inner workings of the YouTube recommender system, we have to be aware that this is an evolving system that will not freeze at the development stage of 2016 [4]. In any case, we are far from knowing the details of this system well enough to recreate it. This is because the paper leaves out many details that would be necessary to do so. On top of that it could only be replicated with the full dataset of any point in time which we do not possess. Even if we did, we would probably be unable to process it due to its size. What we can do, however, is to roughly describe how the system works and then try to generalize to a much simplified form that can at least be simulated in a more qualitative style (what happens when this factor changes).

The YouTube recommender system that is described in [4] consists of two main parts:

- a collaborative filtering system that generates a set of suggestions on the order of several hundreds, and
- a personalized ranking system that selects the best fitting content from these that is then offered to the user.

Both parts rely on deep learning, and in the following, we will focus on the first system only because it requires much less user data and is partly oriented at serving groups of users well. It takes into account the last video views and searches of a single user, but also the age, gender, and geographic location. From [4] we can obtain a good overview of the general mechanisms of the recommender system, but a number of important concrete details remain in the dark. This, however, is not surprising as the exact function of the system is a) valuable intellectual property of YouTube, and b) would require a lot of concrete data that would not fit into a conference paper.

There are a number of attempts to understand the system beyond the sketch in [4], e.g.<sup>1</sup>, most likely with the aim to be able to explain how it behaves in a lot of reported cases where its behavior was unexpected. We will not dive very deep but attempt to obtain a rough picture on the mechanics of the collaborative filtering (suggestion generation) system.

Roughly, the recommendation system learns what to recommend to whom by merging several layers of embeddings. An embedding is a way to map rather sparse, very high dimensional data into a much lower, constant dimension space. This is necessary because technically, both dealing with variable dimensions and sparse data are very difficult. Also, the underlying architecture of *artificial neural networks* (ANN) requires a constant number of input and output dimensions.<sup>2</sup>

We can thus say that the embedding works as a means of compression. The input data is not fully known, and Google states that a fixed but large vocabulary

<sup>1</sup> <https://youtuberecommends.2018.ctcp506.georgetown.domains/>.

<sup>2</sup> For a general idea on how embeddings work see: <https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526>.

is used to describe the videos. We presume that this works like a tagging process where for each video a variable, possibly high number of tags is generated such that each video is described with a *bag-of-words*. A second source of information depends on the users and takes the users' last searches, views, and also group related information as the geographic location, gender, age, etc. into account (see [4]).

For both of these data sets (and possibly more), single embeddings are learned, and these are combined by averaging later on. During learning, this effectively means that videos that are often viewed subsequently shall be placed “closer” to each other during learning time. At serving time, a much faster process is needed and here a compromise between accuracy and performance has to be achieved. Whereas in the learning phase, a high dimensional but fixed space (according to [4] we presume that it most likely has 256 dimensions) is filled with video entries, at lookup time an approximate nearest neighbor scheme is used to find the videos that are closest to any given video. This is then the set of hundreds of candidate videos. In the second step, the recommender system—by taking even more information (e.g. user language, time since last watch, previous interactions) about individual users into account—brings down this number to something around some dozens of videos which are then presented to the user.

One obvious way to trick the system into giving more weight to the actions of some users than others is actually disabled by design: Regardless of the activity, only a fixed amount of user interactions are considered for the training process (50 last videos, 50 last searches). This means that particularly active users might still have comparatively more influence on the recommendations, but this influence is capped.

What we do not know, however, is how often the system is retrained or updated. Every conceivable influence of the user behavior on the recommendation system can only be realized when the user data is fed into the training, and in theory it is easy to imagine some sort of feedback loop: Unwanted recommendations are erroneously played out to the users and if these accept the recommendations, this amplifies the video to video relation of two videos that appear to be “similar” but might not be in reality. Generally, more frequent iterations speed up the process of single videos getting more popular or specific sequences of videos being stored in the system. Thus we find conflicting objectives here (very often, it is a highly desired effect that some videos get more popular quickly): updating the system very often leads to wanted (viral “hit” videos) *and* unwanted videos or combinations spreading quicker (timeliness vs error propagation). A slower update process (less trainings) would of course slow down how fast wanted and unwanted content combinations spread.

## 4 The Role of the User

Kahn [8] presents an overview of YouTube users' motivations, based on McQuail's [14] Uses and Gratifications framework for media choice. Kahn distinguishes

between five different motives: Seeking Information, Giving Information, Self-Status Seeking, Social Interaction and Relaxing Entertainment. Most of these factors, however, do not really bear on the question we are dealing with here: People who watch videos presented to them by YouTube's recommender system are most likely looking for, to use Kahn's terminology, "relaxing entertainment" or possibly "seeking information". To quote: "Seeking information and relaxing entertainment motives are factors that were highly significant in explaining a user's behavior of viewing videos." The possible user actions considered by Kahn and others in the field, like commenting, "liking" or "disliking" videos, uploading videos or sharing videos can be largely discounted for the purpose of this paper. One interesting result from Kahn's study, however, points to the kind of user situation we are most likely dealing with here: According to Kahn, users who seek entertainment on YouTube are likely to "like", "dislike" (thumbs down button) and share videos, but not more likely comment on or upload videos. In other words: Actions that can be performed, with one click, while watching videos at the same time are likely for people who use YouTube for entertainment purposes, while actions that would require interrupting the watching session are not. There is a number of data points that point to the fact that recommendations generated by YouTube's recommender system account for a high percentage of the views. A YouTube representative claimed in early 2018 that 70% of watch time is due to algorithmic recommendations [18]. Also, according to a Pew study, 8 in 10 Americans watch videos recommended by YouTube at least occasionally [17].

What is the user actually doing (in an process flow sense) when consuming videos on YouTube? This is going to be important in order to be able to simulate user behavior as well. We assume that a YouTube session usually starts with some kind of search, such that the first video that is seen is not depending on the recommender system but is a user choice (from a more general viewpoint the user starts with a random video). When the first video finishes, the user is provided with a list of ranked recommendations. If autoplay is activated and the user does not interact, the next video is simply the highest ranked video of these recommendations. If autoplay is disabled or the user chooses to interact, he/she may accept one of the videos from the presented recommendation list. These may be contextually appropriate, but some may actually be contextually inappropriate. If autoplay is disabled or disregarded, the user has to actively choose one of these in order to keep watching. We presume that if this is not the case, either the YouTube session stops or another "random" video is chosen (from the system viewpoint it may appear random; from the user viewpoint the video is probably chosen on the basis of criteria we do not know, possibly based on another manual search).

## 5 Simulation

Our simulation system attempts to replicate the most important mechanisms as given in the description of the original YouTube recommendation system [4]. However, we are well aware that:



- the system has probably already changed at least in many details since the publication of that paper, and
- due to the complexity of the deep learning system employed and the lack of exact data, a very close reproduction is not possible.

The basis of the simulation system is a coordinate system we model after the presumed latent space of the original system. By latent space we mean the hyperspace (presumably 256 dimensions) used by the recommender system for embedding all YouTube videos before later on using the positions of the videos in that space for looking up (via approximate nearest neighbor search) similar videos.

We thus use a 256 dimensional unit hypercube<sup>3</sup> (coordinate values between 0 and 1) in which each video is placed onto a “real” position, according to its properties. By real position we mean that if the tagging process for each video and its mapping to the 256 dimensional space were error-free (also taking into account the positions of similar videos), this is where it should be positioned. Next to this real position, we determine an “apparent” position. This is determined by taking the real position and adding some noise. This noise reflects the errors that have been made in tagging and placing. Whereas it is difficult to know what the error distributions are like in reality, we simply assume that the tagging process has about 10% error. The error is modeled by changing every coordinate of the apparent position with a 10% change to a random position. Note that the exact size of this error is not really important, it will just accelerate or decelerate the spread of classification errors (and thus possible contextually inappropriate content).

This placing of the videos in a high dimensional space simulates learning a video embedding in such a space, i.e. the training of the neural network without taking user information into account. How can we now integrate recent searches/views by the users? In the original system, the user information seems to be employed for learning another embedding layer. The layers are then connected via averaging. That means (and this is a functional building block of the system) that user histories and the sequential video pairs therein can drive videos closer together. We model this by means of a simpler mechanism:

We collect all pairs of videos that are viewed after each other by any user and apply a weak force between them in the apparent space. The force is toned down so that the two sources of information (original position, all viewed pairs including this video) are approximately on the same scale. In order to compute concrete force values we have to decide what an average distance between two videos is. That is in our case the expected distance between two random points in a unit hypercube  $[0, 1]^{256}$ . Computing this results in a complicated multiple

---

<sup>3</sup> It may be difficult to imagine such a huge space since we experience the world around us as only three dimensional. In principle we could also break this high dimensional space into a large number of 3D spaces, in our case around 85 3D cubes would be necessary. A point (video in our case) would then have a 3D position in each of these cubes and all cubes together provide its position in 255 dimensions.

integral that does not seem to be analytically known for such high dimensions<sup>4</sup>. However, there is an approximation for lower and upper bounds on this value in dependence of  $n$  (number of dimensions) [1], see (1) that is fairly easy to compute.

$$\frac{1}{3}n^{1/2} \leq \Delta(n) \leq \left(\frac{1}{6}n\right)^{1/2} \sqrt{\frac{1}{3}\left[1 + 2\left(1 - \frac{3}{5n}\right)^{1/2}\right]} \tag{1}$$

It is known that especially for lower values of  $n$ , the true value of  $\Delta(n)$  is much closer to the upper bound than to the lower bound. We thus simply use the upper bound (right side) as a stand-in and name it  $\hat{\Delta}(n)$ . We divide this estimated value by the overall count of a specific video in all user histories  $v_i$ , multiplied by the number of users  $U$ . The idea behind this is that the maximum applied forces between two respective videos should at least not be larger than the approximated distance of two random videos. As we will see later, the resulting force is probably still too strong, so we may see this approach rather as reflecting a trend and not as a concrete value we can build on.

This way, we can now compute the force we apply between the positions of video  $i$  and video  $j$  as in (2). The 2 in the denominator stems from the fact that almost all videos in history lists (except the first and the last) do have a predecessor and a successor (so each video has two neighbors).

$$f_{i,j} = \frac{\hat{\Delta}(n)}{2 \cdot v_i \cdot U} \tag{2}$$

```

Data: numbers of videos, users, size of user histories  $h$ , epochs
Result: number of suggested/accepted inappropriate/appropr. videos
set up real video positions in 256 dim space;
set up apparent video positions as slight variation of the above;
fill user histories: apply user behavior of Alg.2 for  $2 \cdot h$  videos each;
reset all counters: inappropriate/appropriate suggested/accepted videos;
while not terminated (due to number of epochs) do
    | update apparent video positions: apply forces from user histories;
    | simulate user consuming  $2 \cdot h$  videos, employ recommendation system;
    | measure the number of accepted/suggested approp/inapprop. videos;
end
accumulate numbers;
    
```

**Algorithm 1:** Simulation steps

The combined amount of forces (regardless of the directions) on one video thus cannot be larger than the approximate distance of any two random videos, and even this applies only if all of the single constituents point into the same direction. The real forces are probably much smaller, as we can expect a lot of movements to partly or fully cancel each other out (dragging a video in lots

<sup>4</sup> <http://mathworld.wolfram.com/HypercubeLinePicking.html>.

of different directions may actually result in a very similar final position). It is clear that this model is not particularly accurate and we will not be able to make good quantitative predictions with it, but it is definitively closely related to the main mechanics of the original system and should thus be sufficient to ask what-if questions. Of course, we apply the forces only to the apparent space position of each video, not to its real position. Presuming that we can take the relation of two videos to be contextually matching or mismatching as stable even if some users see both videos in a row, we also keep the real position fixed (the real tagging of each video should not depend on user view choices but the other way around).

Returning to our original goal, we now need to define how we want to recognize that a video suggestion is inappropriate, given the context of the most recently watched video(s). This is a difficult question to answer, so we simply rely on a rough estimate: If the real positions of two videos are more distant from each other than the distance of two random videos (which resembles  $\Delta(n)$ ) we assume that they are not contextually related. If they are closer together, they may or may not be contextually related, but we assume that they are, at least to a certain extent. Note that we operate in a very high dimensional space (256). This means that certain notions of distance and neighborhood are much more fragile than in our usual 3D world.

The overall course of the simulation is described in pseudocode in Algorithm 1, the user behavior for choosing  $2 \cdot h$  videos (in order to fill the user history with fresh content) is provided in pseudocode in Algorithm 2. Note that we did not reflect the erroneous nature of the approximate nearest neighbor lookup that is used in the original recommendation algorithm as described in [4]. Here, we simply presume that nearest neighbor searches are accurate, assuming that the possible effect of errors here is rather small.

## 6 Experimental Analysis

In the simulation system described above, we have several parameters that can be set. However, space is limited and simulation time is considerable (around 70 min for 20 repetitions). We thus decided to choose a parameter set that is a compromise between runtimes, memory use on the one hand and hopefully realistic values on the other hand. We set the number of videos in our simulation to 10,000, and the number of users to 1,000. It is clear that these values are much lower than the real world case (YouTube), but otherwise the computation times would explode due to the necessary distance computations. Likewise, we also model the users as a homogeneous group, with the same probabilities to accept recommendations or switch on autoplay. This is of course a very strong simplification. The simulation system could accommodate different probabilities for every user, but apart from the difficulty to obtain such detailed realistic data we do not use this feature now, as our first experiments are meant to focus on the big picture only.

**Data:** probabilities for autoplay  $p_{it}$ , inapprpr./apprpr. accept  $p_{in}, p_{ia}$

**Result:** list of watched videos

```

determine first video randomly;
while not consumed enough videos do
    | get recommendation;
    | check if inappropriate;
    | determine random number between 0 and 1 as  $r$ ;
    | if autoplay on then
    | | accept recommendation;
    | else
    | | if video appropriate and  $r < p_{ia}$  then
    | | | accept recommendation;
    | | end
    | | if video inappropriate and  $r < p_{in}$  then
    | | | accept recommendation;
    | | end
    | end
    | if not accepted then
    | | randomly choose video;
    | end
end
accumulate inappropriateness and acceptance counts;

```

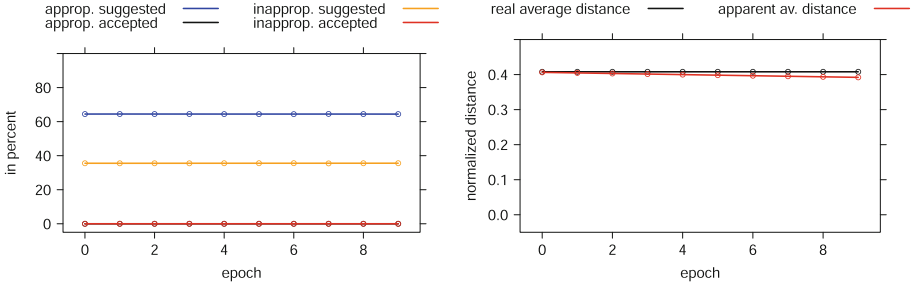
**Algorithm 2:** User video session

The experimental results thus only provide a rough, rather qualitative picture of the underlying effects. We also acknowledge that the applied forces that reflect the acceptance of recommendations may still be too strong and need to further be reduced to provide a realistic impression.

### 6.1 Function Test

At first, we perform a functional test, meaning that with autoplay turned off for all users ( $p_{it} = 0$ ) and acceptance rates for appropriate and inappropriate recommendations set to zero ( $p_{ia} = 0, p_{in} = 0$ ) the video corpus should stay relatively stable. In this case, the users completely ignore the recommendation system and only choose videos they search for, which is modeled as the users always choosing random videos. Figure 1 displays the rates for appropriate and inappropriate videos on the left side, and the real average video distances and apparent video distances on the right side over 10 epochs of training and recommendation. We can easily see that the rates for appropriate and inappropriate video suggestions stay the same over the whole simulation, and that the apparent distances slightly decrease. Standard deviations (we perform 20 runs and average all numbers) are near zero for all measured values.

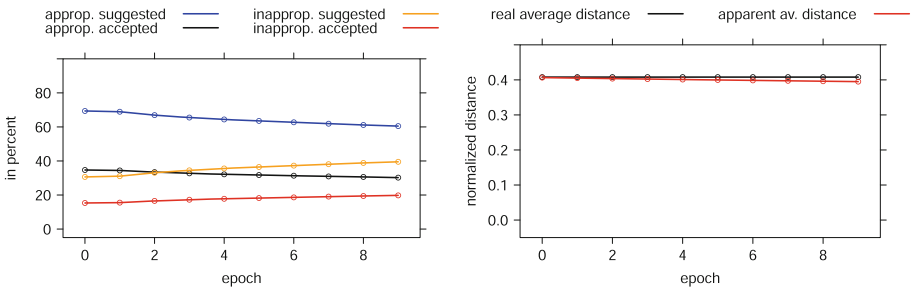
This result means that the simulation system performs as expected in this case, which is no surprise as all forces that are applied have random directions.



**Fig. 1.** Left: rates for appropriate suggested and accepted, and inappropriate suggested and accepted videos for autoplay = 0, appropriate acceptance = 0, inappropriate acceptance = 0. Right: real and apparent average distances between videos. Both over 10 training and recommendation epochs.

### 6.2 Autoplay Test

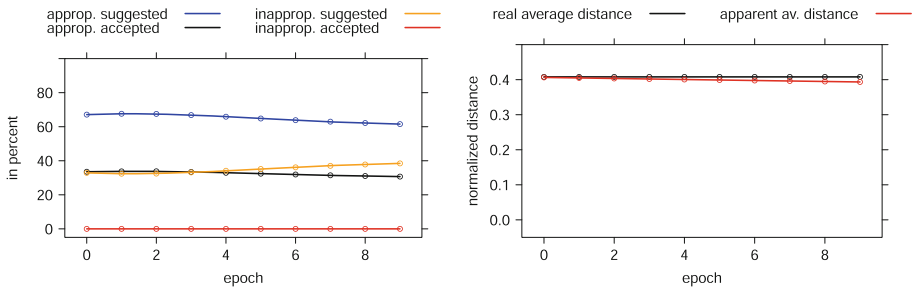
Leaving all other parameters the same as before, we now test what happens if we enable autoplay with a considerable chance of 0.5. This means that in 50% of the watches, the next recommended video is accepted without consideration if it may be appropriate or inappropriate. Figure 2 shows the averages of the obtained results (here, the standard deviations are a bit higher but still on a very low level around 1–2%). We see that the number of appropriate suggestions slowly decreases over the epochs, and the number of inappropriate videos that are suggested and also accepted rises steadily. Average distances of all videos show comparable behavior as in the case where nothing is accepted at all.



**Fig. 2.** Left: rates for appropriate suggested and accepted, and inappropriate suggested and accepted videos for autoplay = 0.5, appropriate acceptance = 0, inappropriate acceptance = 0. Right: real and apparent average distances between videos. Both over 10 training and recommendation epochs.

### 6.3 Appropriate/Inappropriate Accept Test

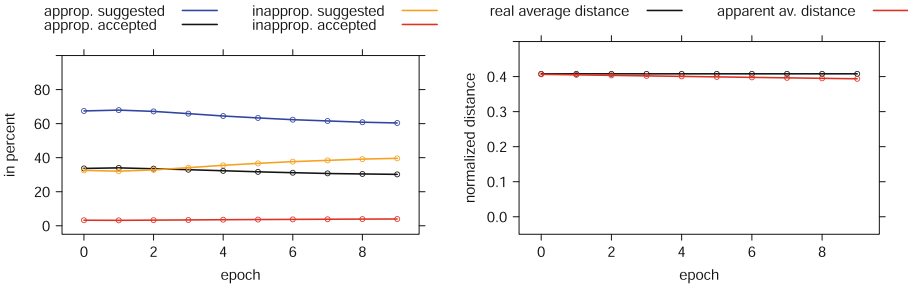
What happens now if autoplay is switched off and either only 50% of the appropriate videos but no inappropriate videos or vice versa are accepted? Figs. 3 and 4 show the results of our simulations, again averaged from 20 runs and with quite low standard deviations around 1% to 2%. Whereas the rate of accepted inappropriate videos stays at 0 in the first case and rises slowly in the second case, the overall impression is comparable. The only remarkable difference is that the number of inappropriate suggestions also rises more quickly than in the first case. Surprisingly, this means that the interaction with the recommendation system itself already leads to a higher rate of suggested inappropriate videos, although in the first case these are never accepted. It seems to be not of importance if the accepted videos are appropriate or not, it is rather important if they have been recommended. It seems that the applied forces make the suggestion of inappropriate videos more likely.



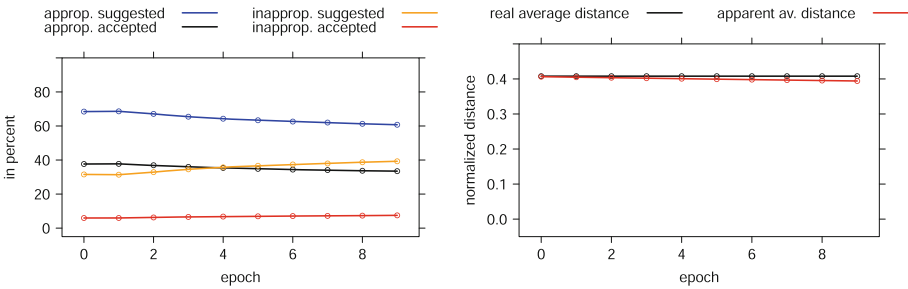
**Fig. 3.** Left: rates for appropriate suggested and accepted, and inappropriate suggested and accepted videos for autoplay = 0.0, appropriate acceptance = 0.5, inappropriate acceptance = 0. Right: real and apparent average distances between videos. Both over 10 training and recommendation epochs.

### 6.4 Realistic Setting with Low Autoplay?

For a hopefully realistic situation, we set the autoplay probability to 0.1, the acceptance of appropriate videos to 0.5, and the acceptance probability for inappropriate videos to 0.1. Figure 5 provides the response of the simulation system. As expected, inappropriate video recommendations are now accepted more often, but still much less frequently than in the 50% autoplay case. Compared to Fig. 4, the overall rate of accepted inappropriate videos ranges at around the double size (7 to 8%) after 10 epochs. This is at first surprising, but taking the autoplay figure into account, it seems that the autoplay feature, even if switched on with a much lower probability, dominates the handling of inappropriate videos (more are accepted and, after a while, more are also being suggested).



**Fig. 4.** Left: rates for appropriate suggested and accepted, and inappropriate suggested and accepted videos for autoplay=0.0, appropriate acceptance=0.0, inappropriate acceptance=0.5. Right: real and apparent average distances between videos. Both over 10 training and recommendation epochs.



**Fig. 5.** Left: rates for appropriate suggested and accepted, and inappropriate suggested and accepted videos for autoplay=0.1, appropriate acceptance=0.5, inappropriate acceptance=0.1. Right: real and apparent average distances between videos. Both over 10 training and recommendation epochs.

## 7 Possible Mitigations

Judging from the results of our simulation, the simplest way to mitigate the bulk of the problem of contextually inappropriate video suggestions would be to eliminate the autoplay option from YouTube. The feature seems to produce an inherent and cumulative error that makes contextually inappropriate video suggestions more likely the longer a viewing session lasts. Maximising watch time across videos is, however, the stated optimization goal of the YouTube development team. If the number given by a YouTube executive in 2018 (see above) is correct, algorithmically generated suggestions are responsible for 70% of watch time on YouTube. It thus seems unlikely that the company would be willing to accept this measure. And since the recommender system itself has no way of “knowing” whether certain users receive contextually inappropriate videos and watch them anyway, more finely tuned mitigation efforts on the supply side don’t seem to easily suggest themselves. One method of approaching the

problem would be to improve the quantity and quality of the feedback users can give. This should be complemented, and this is crucial, by effective measures on the side of the company to act on such feedback. Parents who find that their childrens' videos might have become a desirable kind of content for pedophiles should not have to rely on scientists or major news publications to be heard. A potential regulatory measure would be to limit the total time or number of successive videos that the autoplay function is allowed to deliver. There could, for example, be a mandatory cap on viewing sessions, so that a new chain of recommendation and autoplay would have to be initiated with another "random" video, i.e. one generated by a user initiated search or some other way. The cumulative effects of inappropriate videos entering the recommendation stream would thus also be capped. A first, quick approach to the problem is to alert users, including uploaders and passive YouTube users, to the existence of the problem. Parents, for example, might not even realize which mechanisms lie behind their own childrens' YouTube consumption. They might also be unaware of the danger of receiving contextually inappropriate content suggestions generated this way. Similarly, uploaders might not be aware that their videos might end up in front of an audience that they were not intended for. Alerting users to these facts might be a start.

## 8 Conclusion

We have been looking at the YouTube recommendation system and its effects from various angles. We departed from reports of surprised to annoyed users who have been confronted with strange recommendations as well as by uploaders shocked by view counts that pointed to an audience that they never intended to reach. We then looked at the technical side of how the algorithms in the recommendation system work. Finally, we tried to simulate the most important mechanisms in a "what-if" fashion. From the reports of users and researchers we know that the recommendation system sometimes does not work as expected but sometimes suggests videos assumed to be contextually inappropriate. In Sect. 2, we have attempted to characterize what that actually means, as it seems there is no useful definition available so far.

From the technical viewpoint, it seems that contextually inappropriate recommendations are a collateral damage rather than intended aim of the recommendation system. The aim of YouTube to maximize watch time indirectly enables pandering to all possible user motivations if they help increasing watch time, from joy to disgust. It also appears to be very difficult to get rid of such unwanted recommendations due to a) the huge amount of data that enforces compromises between performance and accuracy, and b) the necessity to incorporate user video watch information into the system that will impact the system's performance (as it is the nature of collaborative filtering systems) in unforeseen ways.

We do not want to exaggerate the relevance and representativity our simulation results, as there are a lot of shortcomings we have to accept (too small, too



simple, etc.). However, it is interesting to see that even if inappropriate videos are not accepted at all by the users and autoplay is switched off, the number of inappropriate suggestions rises over time. Most problematic in this respect seems to be a high autoplay rate, however, because this may lead to situations when users are not even aware that inappropriate videos are currently played, e.g., because they are away from the screen for a few minutes or YouTube is silently continuing to play videos in a background browser tab. Some YouTube users report that they have switched off the autoplay feature in order to obtain more control about what is played and when. The available data, however, point to a high impact of the recommendation system on what videos users actually consume on YouTube. This corresponds to available information on the importance of the recommender system for total watch time as reported by YouTube executives.

It seems clear that more research is needed in this direction. The number of scientific teams that can look into the journalistic, communication scientific, and computer science perspective at the same time seems to be too small to be able to sufficiently research the effects of hugely important socio-technical systems as social media.

## References

1. Anderssen, R.S., Brent, R.P., Daley, D.J., Moran, P.A.P.: Concerning  $\int_0^1 \dots \int_0^1 (x_1^2 + \dots + x_k^2)^{1/2} dx_1 \dots, dx_k$  and a Taylor series method. *SIAM J. Appl. Math.* **30**(1), 22–30 (1976)
2. Bridle, J.: Something is wrong on the internet. Medium (2017). <https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2>
3. Chaslot, G.: Youtube's A.I. was divisive in the us presidential election. Medium (2016). <https://medium.com/the-graph/youtubes-ai-is-neutral-towards-clicks-but-is-biased-towards-people-and-ideas-3a2f643dea9a>
4. Covington, P., Adams, J., Sargin, E.: Deep neural networks for YouTube recommendations. In: *Proceedings of the 10th ACM Conference on Recommender Systems*, New York, NY, USA (2016)
5. Fisher, M., Taub, A.: On YouTube's digital playground, an open gate for pedophiles. *The New York Times* (2019). <https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html>
6. Google. How google fights disinformation (2019). [https://www.blog.google/documents/37/How\\_Google\\_Fights\\_Disinformation.pdf](https://www.blog.google/documents/37/How_Google_Fights_Disinformation.pdf)
7. Kaiser, J., Rauchfleisch, A.: The implications of venturing down the rabbit hole. *Internet Policy Review*, June 2019. <https://policyreview.info/articles/news/implications-venturing-down-rabbit-hole/1406>
8. Khan, M.L.: Social media engagement: what motivates user participation and consumption on YouTube? *Comput. Hum. Behav.* **66**, 236–247 (2017)
9. Maheshwari, S.: Chase had ads on 400,000 sites. Then on just 5,000. Same results. *The New York Times*, March 2017
10. Maheshwari, S.: On YouTube kids, startling videos slip past filters. *The New York Times* (2017). <https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html>

11. Maheshwari, S.: Publishers retreat from the risks of Google-YouTube advertising. *The New York Times* (2017)
12. Maheshwari, S., Wakabayashi, D.: AT&T and Johnson & Johnson Pull Ads From YouTube. *The New York Times* (2017)
13. Martineau, P.: YouTube has kid troubles because kids are a core audience. *Wired* (2019). <https://www.wired.com/story/youtube-kid-troubles-kids-core-audience/>
14. McQuail, D.: *Mass Communication Theory: An Introduction*. Sage Publications, London (1983)
15. Meyerson, E.: YouTube now: why we focus on watch time (2012). <https://youtube-creators.googleblog.com/2012/08/youtube-now-why-we-focus-on-watch-time.html>
16. Naughton, J.: How Peppa Pig knock-offs bring home the bacon for Google. *The Guardian* (2017). <https://www.theguardian.com/commentisfree/2017/nov/12/content-google-youtube-kids-not-always-suitable-for-children-peppa-pig-brings-home-bacon>
17. Smith, A., Toor, S., van Kessel, P.: Many turn to YouTube for children's content, news, how-to lessons. *pewresearch.org* (2018). <https://www.pewresearch.org/internet/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons/>
18. Solsman, J.E.: Ever get caught in an unexpected hourlong YouTube binge? Thank YouTube AI for that. *cnet.com* (2018). <https://www.cnet.com/news/youtube-ces-2018-neal-mohan/>
19. Grimme, C., Preuss, M., Takes, F.W., Waldherr, A. (eds.): *MISDOOM 2019*. LNCS, vol. 12021. Springer, Cham (2020). <https://doi.org/10.1007/978-3-030-39627-5>
20. Tufekci, Z.: Opinion—YouTube, the great radicalizer. *The New York Times* (2018). <https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html>
21. Stöcker, C., Böhm, M.: YouTube: Werbung vor hetzvideos gegen flüchtlinge. *Der Spiegel* (2015). <https://www.spiegel.de/netzwelt/netzpolitik/youtube-werbung-vor-hetzvideos-gegen-fluechtlinge-a-1058254.html>
22. Wakabayashi, D., Maheshwari, S.: YouTube advertiser exodus highlights perils of online ads. *The New York Times* (2017)



# Characterizing Social Bots Spreading Financial Disinformation

Serena Tardelli<sup>1</sup> , Marco Avvenuti<sup>2</sup>, Maurizio Tesconi<sup>3</sup>,  
and Stefano Cresci<sup>3</sup> 

<sup>1</sup> IIT-CNR and Department of Information Engineering,  
University of Pisa, Pisa, Italy  
`serena.tardelli@iit.cnr.it`

<sup>2</sup> Department of Information Engineering, University of Pisa, Pisa, Italy  
`marco.avvenuti@unipi.it`

<sup>3</sup> IIT-CNR, Pisa, Italy  
`{maurizio.tesconi,stefano.cresci}@iit.cnr.it`

**Abstract.** Despite the existence of several studies on the characteristics and role of social bots in spreading disinformation related to politics, health, science and education, *financial social bots* remain a largely unexplored topic. We aim to shed light on this issue by investigating the activities of large social botnets in Twitter, involved in discussions about stocks traded in the main US financial markets. We show that the largest discussion spikes are in fact caused by mass-retweeting bots. Then, we focus on characterizing the activity of these financial bots, finding that they are involved in speculative campaigns aimed at promoting low-value stocks by exploiting the popularity of high-value ones. We conclude by highlighting the peculiar features of these accounts, comprising similar account creation dates, similar screen names, biographies, and profile pictures. These accounts appear as untrustworthy and quite simplistic bots, likely aiming to fool automatic trading algorithms rather than human investors. Our findings pave the way for the development of accurate detection and filtering techniques for financial spam. In order to foster research and experimentation on this novel topic, we make our dataset publicly available for research purposes.

**Keywords:** Social bots · Disinformation · Deception · Financial spam · Stock markets · Twitter

## 1 Introduction

Nowadays, social bots play a pivotal role in shaping the content of online social media [25]. Their involvement in the spread of disinformation ranges from the promotion of low-credibility content, astroturfing, and fake endorsements, to the propagation of hate speech propaganda, in attempts to manipulate public opinion and to increase societal polarization [26]. Indeed, recent studies have observed the presence of artificial tampering in a wide variety of online topic

debates, including political discussions, terrorist propaganda, and health controversies [8].

A growing field under scrutiny is the online financial ecosystem, in which social bots now pervade [14,24]. Indeed, such an ecosystem has proven to be of great interest as a valuable ground to entice investors. Although the leverage of social media content for predicting trends in the stock market has promising potential [6], the presence of social bots in such scenarios poses serious concerns over the reliability of financial information. Examples of repercussions of financial spam on unaware investors and automated trading systems include the real-world event known as the Flash Crash – the one-day collapse of the Dow Jones Industrial Average in 2010 induced by an error in the estimation of online information by automated trading systems [19]. Another most notable example is the hacking of the US International Press Officer’s official Twitter account in 2013, when a bot reported the injury of President Obama following a terrorist attack, causing a major stock market drop in a short time<sup>1</sup>. Finally, we witnessed to the abrupt rise of *Cynk Technology* in 2014 from an unknown unprosperous small company to a billions-worth company, due to a social bot orchestration that lured automatic trading algorithms into investing in the company’s shares based on a fake social discussion, which ultimately resulted in severe losses<sup>2</sup>. Therefore, investigating such manipulations and characterizing them is of the utmost importance in order to protect our markets from manipulation and to safeguard our investments.

**Contributions.** In an effort to shed light on the little-studied activity of social bots tampering with online financial discussions, we analyze a rich dataset of 9M tweets discussing stocks of the five main US financial markets. Our dataset is complemented with financial information collected from Google Finance, for each of the stocks mentioned in our tweets. By comparing social and financial information, we report on the activity of large botnets perpetrating speculative campaigns aimed at promoting low-value stocks by exploiting the popularity of high-value ones. We highlight the main characteristics of these financial bots, which appear as untrustworthy, simplistic accounts. Based on these findings, we conclude that their activity is likely aimed at fooling automatic trading algorithms rather than human investors.

Our main contributions are analytically summarized as follows:

- We outline the activities and role of social bots in the spread of financial disinformation on Twitter.
- We uncover the existence of several large botnets, actively involved in artificially promoting specific stocks.
- We characterize social bots tampering with financial discussions from various perspectives, including their content, temporal, and social facets.

<sup>1</sup> <https://www.bbc.com/news/world-us-canada-21508660>.

<sup>2</sup> <https://www.cnbc.com/2014/07/25/mysterious-stock-cynk-plummets-after-reopening.html>.

Our findings provide an important contribution towards understanding the role and impact of social bots in the financial domain, and pave the way for the development of accurate detection and filtering techniques for financial spam.

## 2 Related Work

As anticipated, in the present study we are interested in analyzing the activity, the behavior, and the characteristics of financial social bots. For this reason, in this section we do not survey previous works related to the *detection* of social bots – which is a different topic covered by many through studies [8] – but we rather focus on those works related to the *characterization* of malicious accounts.

Given the many issues caused by malicious accounts to our online social ecosystems, a large body of work analyzed the behavior of bots and trolls in disinformation campaigns aimed at influencing a variety of debates. To understand how trolls tampered with the 2016 US Presidential elections, previous work characterized the content they disseminated, and their influence on the information ecosystem [30]. Among other findings, authors discovered that trolls were created a few weeks before important world events, and that they are more likely to retweet political content from normal Twitter users, rather than other news sources. In [31], authors evaluated the behavior and strategy changes over time of Russian and Iranian state-sponsored trolls. By exposing the way Iranian trolls changed behavior over time and started retweeting each other, they highlighted how strategies employed by trolls adapt and evolve to new campaigns. Authors in [27] detected and characterized Twitter bots and user interactions by analyzing their retweet and mention strategies, and observed a high correlation between the number of friends and followers of accounts and their bot-likeness. In [1], authors characterized Arabic social bots spreading religious hatred on Twitter, and discovered they have a longer life, a higher number of followers, and an activity more geared towards creating original content than retweets, compared to English bots [5]. The previous works remarked the importance of understanding the inherent characteristics of bots and trolls. In fact, despite showing signs of bot-likeness, bots do not often get caught in time, thus potentially affecting the polarization and outcome of essential debates. Moreover, previous works also highlighted how bots and trolls evolve and adapt to new contexts. Despite such consistency in previous results, the characteristics of social bots disseminating financial information are yet to be explored. The few previous works that tackled automation and disinformation in online financial discussions, went as far as providing evidence of the presence of financial spam in stock microblogs and raised concerns over the reliability of such information [13,14]. However, the detection and impact estimation of such bots in social media financial discussions still represent largely unexplored fields of study. Conversely, the leverage of social bots in other sectors has been extensively examined, with previous works focusing on the interference of bots in health issues [2], terrorist propaganda [3], and political election campaigns in the US [5], France [16], Italy [10], and Germany [7], to name but a few.

In this work, we aim at filling in the missing piece of the puzzle – that is, the characterization of social bots in online financial conversations, with a focus on how they are organized and how they operate.

**Table 1.** Statistics about the financial and social composition of our dataset.

| Markets      | Financial data |                  |                  | Twitter data |           |                 |
|--------------|----------------|------------------|------------------|--------------|-----------|-----------------|
|              | Companies      | Median cap. (\$) | Total cap. (\$B) | Users        | Tweets    | Retweets (%)    |
| NASDAQ       | 3,013          | 365,780,000      | 10,521           | 252,587      | 4,017,158 | 1,017,138 (25%) |
| NYSE         | 2,997          | 1,810,000,000    | 28,692           | 265,618      | 4,410,201 | 923,123 (21%)   |
| NYSEARCA     | 726            | 245,375,000      | 2,227            | 56,101       | 298,445   | 157,101 (53%)   |
| NYSEMKT      | 340            | 78,705,000       | 256              | 22,614       | 196,545   | 63,944 (33%)    |
| OTCMKTS      | 22,956         | 31,480,000       | 45,457           | 64,628       | 584,169   | 446,293 (76%)   |
| <b>Total</b> | 30,032         | –                | 87,152           | 467,241      | 7,855,518 | 1,802,705 (23%) |

### 3 Dataset

By leveraging Twitter’s Streaming API [15], we collected all tweets mentioning at least one of the 6,689 stocks listed on the official NASDAQ Web site<sup>3</sup>. Companies quoted in the stock market are easily identified on Twitter by means of *cashtags* – strings composed of a dollar sign followed by the ticker symbol of the company (e.g., \$AAPL is the cashtag of Apple, Inc.). Just like the hashtags, cashtags serve as beacons to find, filter, and collect relevant content [18].

Our data collection covered a period of five months, from May to September 2017, and resulted in the retrieval of more than 9M tweets. We also extended the dataset by gathering additional financial information (e.g., capitalization and industrial classification) about the companies mentioned in our tweets, by leveraging the Google Finance Web site<sup>4</sup>. Table 1 shows summary statistics about our dataset, which is publicly available online for research purposes<sup>5</sup>.

### 4 Uncovering Financial Disinformation

In this section we describe the various analyses that allowed us to uncover widespread speculative campaigns perpetrated by several botnets. For additional details on the subsequent analyses, we point interested readers to [14].

<sup>3</sup> <http://www.nasdaq.com/screening/company-list.aspx>.

<sup>4</sup> <https://www.google.com/finance>.

<sup>5</sup> <https://doi.org/10.5281/zenodo.2686862>.

### 4.1 Dataset Overview

Each tweet in our dataset mentions at least one of the 6,689 stocks of the NASDAQ list. Companies from this list typically feature a large market capitalization and are traded in the 4 main US financial markets – namely, NASDAQ, NYSE, NYSEARCA, and NYSEMKT. However, among tweets mentioning our 6,689 stocks, we also found many mentions of other, less known, stocks. In particular, as shown in Table 1, overall tweets of our dataset also mention 22,956 stocks traded in the OTCMKTS market. Contrarily to the main stock exchanges, OTCMKTS has less stringent constraints and mainly hosts stocks with a small capitalization. Unsurprisingly, if we analyze our whole dataset, no company from OTCMKTS appears among those that are discussed the most. In fact, the most tweeted companies in our dataset are in line with those found in previous works [18], and include well-known and popular stocks such as \$AAPL, \$TSLA, and \$FB. Nonetheless, a few concerns rise if we consider the rate of retweets for OTCMKTS stocks, which happens to be as high as 76% and in sharp contrast with the much lower rates measured for all other markets. Since automated mass-retweets have been frequently exploited by bots and trolls to artificially boost content popularity [23], this result might hint at the possibility of a manipulation related to OTCMKTS stocks.



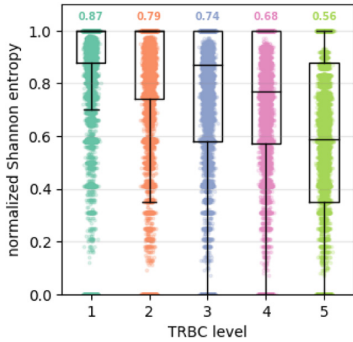
Fig. 1. Examples of tweets in which a few high-capitalization companies (green-colored) co-occur with many low-capitalization ones (red-colored). (Color figure online)

### 4.2 Investigating Financial Discussion Spikes

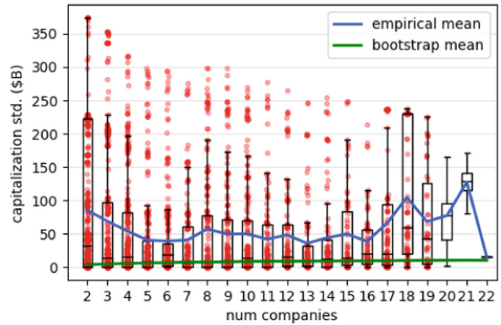
In order to deepen our analysis of financial conversations, we now focus our attention on discussion spikes about the 6,689 stocks of our starting list. For identifying discussion spikes, we compute for each stock the hourly time series of the volume of tweets mentioning that stock, to which we apply a simple anomaly detection technique. In detail, we label as anomalies those peaks of discussion that exceed the mean hourly volume of tweets by more than 10 standard deviations, finding in total 1,926 financial discussion spikes.

Within the discussion spikes, we found more retweets than in the rest of the dataset – namely, 60% retweets for spikes *vs* 23% for the whole dataset,

on average. This finding alone does not necessarily imply a coordinated inauthentic activity, since also organic surges of interest in social media typically result in many retweets. What is unusual however, is that tweets posted during the identified discussion spikes contain, on average, many more cashtags (i.e., mentioned stocks) than the ones in the rest of the dataset. Moreover, such co-occurring stocks seem largely unrelated, and the authors of those tweets do not provide any information to explain the co-occurrences, as shown in the examples of Fig. 1.



**Fig. 2.** Entropy of the industrial classes of co-occurring stocks in discussion spikes. As shown, the high measured entropy implies that co-occurring companies are largely unrelated.



**Fig. 3.** Standard deviation of the capitalization of co-occurring companies in discussion spikes, and comparison with a bootstrap. The large measured standard deviation implies that high-cap companies co-occur with low-cap ones.

### 4.3 Co-occurring Stocks

To investigate the reasons behind this large number of co-occurring stocks, we follow two different hypotheses: (i) stocks might co-occur because of a similar industrial sector (i.e., companies involved in the same business are more likely to be mentioned together) or (ii) they might co-occur because of a similar market value (i.e., high capitalization companies are more likely to be compared to others with similar capitalization).

To assess whether our co-occurring stocks have a similar industrial sector, we leverage Thomson Reuters Business Classification (TRBC)<sup>6</sup>. In particular, we compute the normalized Shannon entropy between the TRBC classes of co-occurring stocks for each tweet that contributes to a discussion spike. This analysis is repeated for all 5 TRBC levels. Each entropy value measured for a given

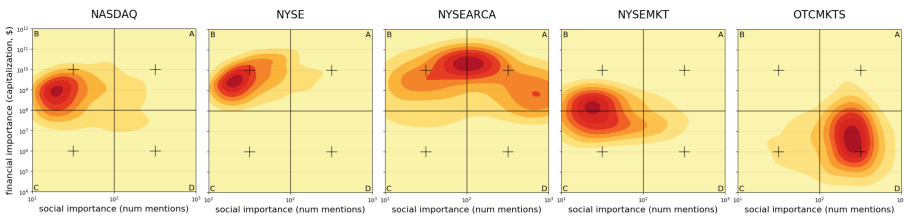
<sup>6</sup> TRBC is a 5-level hierarchical sector and industry classification, widely used in the financial domain for computing sector-specific indices: [https://en.wikipedia.org/wiki/Thomson\\_Reuters\\_Business\\_Classification](https://en.wikipedia.org/wiki/Thomson_Reuters_Business_Classification).



TRBC level for discussion spikes is then compared with the corresponding one computed out of the whole dataset. Results of this analysis are shown in Fig. 2 and depict a situation characterized by large entropy values (i.e.,  $\simeq 1$ , which is the maximum possible value of normalized entropy). In turn, this implies that co-occurring companies in discussion spikes are almost completely unrelated with regards to their industrial classification. Moreover, entropy values measured for discussion spikes are always higher than those measured for the whole dataset.

Regarding financial value, we assess the extent to which co-occurring companies have a similar value by measuring the standard deviation of their market capitalizations. To understand whether the measured standard deviation is due to the intrinsic characteristics of our dataset (i.e., the underlying statistical distribution of capitalization) or to other external factors, we compared mean values of our empiric measurements with a bootstrap. Results are shown in Fig. 3 and highlight a large empiric standard deviation between the capitalization of co-occurring companies, such that a random bootstrap baseline – accounting for the intrinsic characteristics of our dataset – can not explain it. These results mean that not only high-capitalized companies indeed mostly co-occur with small-capitalized ones, as shown in Fig. 1, but also that this phenomenon is rather the consequence of some external action.

In summary, we demonstrated that tweets responsible for generating financial discussion spikes mention a large number of unrelated stocks, some of which are high-cap stocks while the others are low-cap ones.

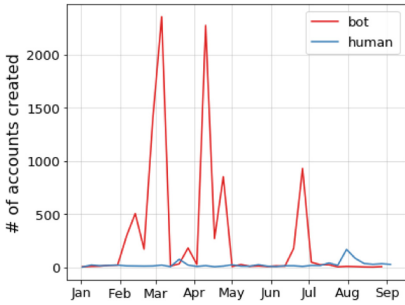


**Fig. 4.** Kernel density estimation investigating the relation between social and financial importance, for stocks of the 5 considered markets. OTCMKTS stocks have a suspiciously high social importance despite their low financial importance.

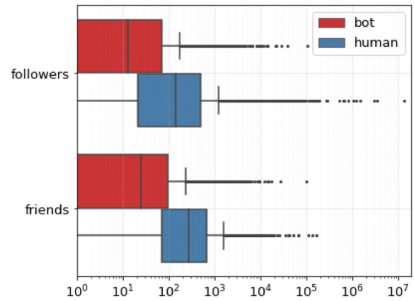
#### 4.4 Financial vs Social Importance

Several existing systems for forecasting stock prices leverage the positive correlation between discussion volumes on social media around a given stock, and its market value [22]. In other words, it is generally believed that stocks with a high capitalization (i.e., high *financial* importance) are discussed more in social media (i.e., high *social* importance) than those with a low capitalization. In this section we verify whether this expected positive relation exists also for the stocks in our dataset.

In fact over our whole dataset, we measure a moderate positive Spearman’s rank correlation coefficient of  $\rho = 0.4871$  between social and financial importance, thus confirming previous findings. However, when focusing on discussion spikes only, we measure a suspicious behavior related to OTCMKTS stocks, which feature a negative  $\rho = -0.2658$ , meaning that low-value OTCMKTS stocks are more likely to appear in discussion spikes than high-value ones. To thoroughly understand the relation between social and financial importance, in Fig. 4 we report the results of a bi-dimensional kernel density estimation of social and financial importance for stocks of the five considered markets. Confirming previous concerns, OTCMKTS stocks feature a suspiciously high social importance, despite their low financial importance, in contrast with stocks of all other markets.



**Fig. 5.** Number of accounts created per week in 2017. Bot accounts display coordinated creation activities, while humans are more evenly spread across the year.



**Fig. 6.** Distribution of the number of followers and friends. Bot accounts show a lower number of followers and friends with respect to human accounts.

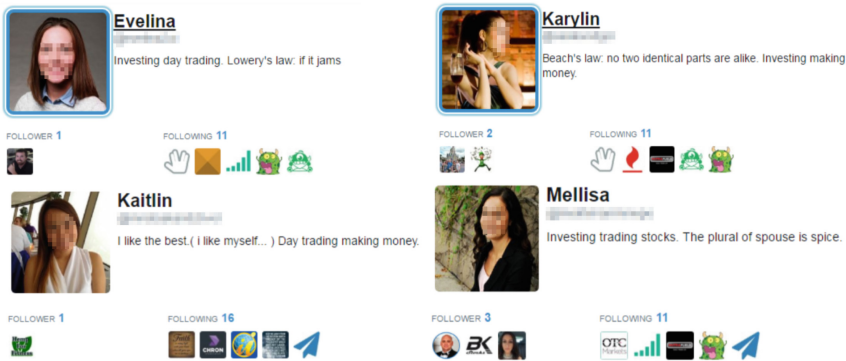
## 5 Bot Detection and Characterization

In the previous section, we described several suspicious phenomena related to stock microblogs. In detail, discussion spikes about high-value stocks are filled with mentions of low-value (mainly OTCMKTS) ones. Such mentions are not explained by real-world stock relatedness. Moreover, the discussion spikes are largely caused by mass retweets.

### 5.1 Bot Detection

In order to understand whether the previously described disorders in financial microblogs are caused by organic (i.e., human-driven) or rather by synthetic activity, here we discuss results of the application of a bot detection technique to all users that contributed to at least one of the top-100 largest discussion spikes. In this way, we analyzed roughly 50% of all our dataset, both in terms of tweets and users, in search for social bots.

To perform bot detection, we employ the state-of-the-art technique described in [10], which is based on the analysis of the sequences of actions performed by the investigated accounts. Strikingly, the technique classified as much as 71% of all analyzed users as bots. Moreover, 48% of the users classified as bots were also later suspended by Twitter, corroborating our results. Given these important findings, we conclude that social bots were responsible for perpetrating the financial disinformation campaigns that promoted OTCMKTS low-value stocks by exploiting the popularity of high-value ones. In the remainder of this section we report on the general characteristics of the 18,509 users classified as financial social bots and we compare them to other bots and trolls previously studied in literature as well as to the 7,448 accounts classified as humans.



**Fig. 7.** Examples of a subset of users classified as bots. The accounts show similarities in their names, screen names, numbers of followers and followings, and in their description. Such similarities support the hypothesis that these accounts are part of large, organized botnets.

## 5.2 Profile Characteristics of Financial Bots

The *creation date* is an unforgeable characteristic of a social media account that has been frequently used to spot groups of coordinated malicious accounts (e.g., bots and trolls) [29]. Its usefulness lies in the impossibility to counterfeit or to masquerade it, combined with the fact that “masters” typically create their bot and troll armies in short time spans<sup>7</sup>. As a consequence, large numbers of accounts featuring almost identical creation dates might represent botnets or troll armies. Given this picture, the first characteristic of financial social bots that we analyze is the distribution of their account creation dates. The creation dates of the accounts in our dataset are distributed between 2007 and 2017. However, the majority of bots (53%) were created in 2017, as opposed to humans (12%). Figure 5 shows the distribution of creation dates of bots and humans in

<sup>7</sup> <https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html>.

2017, at a weekly granularity. Interestingly, bots display coordinated creation activities, while the creation of human accounts is more evenly distributed across the year. In detail, 45% of bots were created between February and April, with a particularly significant spike of 1,346 bots created on March 2. These findings further confirm the manufactured nature of the accounts classified as bots, and their pervasive presence in stock microblogs.

**Table 2.** Top-5 words and 3-grams used in account descriptions by bots and humans. Descriptions for humans are more heterogeneous and repetitions are less frequent.

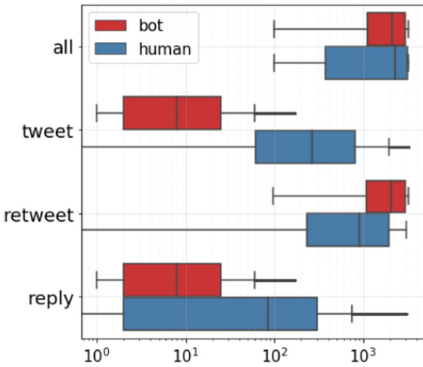
| Social bots |       |                          |       | Humans |       |                       |       |
|-------------|-------|--------------------------|-------|--------|-------|-----------------------|-------|
| Word        | Freq. | 3-gram                   | Freq. | Word   | Freq. | 3-gram                | Freq. |
| Trading     | 8,138 | Day trading making       | 848   | Love   | 314   | Follow follow back    | 7     |
| Day         | 4,195 | Trading making money     | 848   | Life   | 209   | Never say never       | 7     |
| Money       | 4,173 | Investing day trading    | 838   | Follow | 168   | Always strive prosper | 6     |
| Stocks      | 4,056 | Trading stocks investing | 821   | Music  | 164   | Live life fullest     | 5     |
| Trading     | 4,047 | Investing trading stocks | 814   | Like   | 112   | Stock market investor | 4     |

Colluding groups of bots and trolls have also been associated to peculiar patterns in their *screen names* [21]. This is because they represent fictitious identities whose names and usernames are typically generated algorithmically. Looking for artificial patterns in the screen name, we first analyze the distribution of the screen name length. Interestingly, 50% of bots have a screen name length between 14 and 15 characters, while only 26% of humans share such characteristic. By examining the structure of suspiciously long bot screen names, we observe two main patterns. The first denotes the presence of screen names composed of a given name, followed by a family name. Such users also use the given name, which in almost all the cases is a female English name, as their display name. The second pattern exposes bots with a screen name composed of exactly 15 random alpha-numeric characters, accompanied by a given name as a display name. Such phenomenon has been observed before for numerous bot accounts involved in two different political-related events [4], and it’s a strong confirmation of the malicious nature of our accounts labelled as bots. Figure 7 provides some examples of such bots. Moreover, by cross-checking information related to the creation dates, we observe that 11% of such bots are created on the same day.

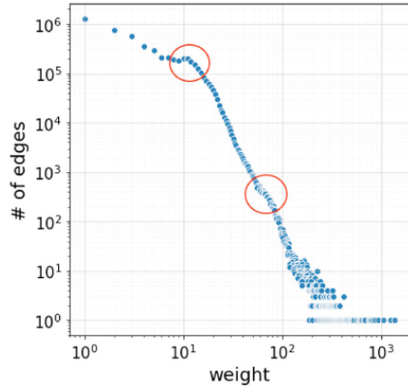
Next, we inspect account *descriptions* (also known as *biographies*). We find a total of 575 users (3%) sharing the exact same description with at least 3 other users. In other words, there are 174 small groups of at least 3 users having in common the same description. Such repeated descriptions follow a specific pattern – in particular, they are composed of a famous quote or law, and of a set of financial keywords that are totally unrelated with the rest of the description. Interestingly, the use of famous quotes by bots to attract genuine users has already been documented before, for bots acting in the political domain [11].

We find 373 occurrences of such pattern, and none amongst the users with this pattern is classified as human. Some bot accounts exhibiting this characteristic are shown in Fig. 7. Table 2 summarizes the words and 3-grams mostly used in account descriptions by bots and humans. As shown, striking differences emerge. In summary, all previous findings support the hypothesis that users classified as bots did not act individually but that are rather part of large, organized and coordinated botnets.

Finally, we measure differences between bots and humans with respect to their *social relationships*. In particular, Fig. 6 shows differences in the distributions of followers and followings. Bots are characterized by a significantly lower number of both followers and followings, indicating accounts with few social relationships. It has been demonstrated that accounts with many social relationships in online social platforms are perceived as more trustworthy and credible [9]. Thus, to this regard, our financial bots appear as rather untrustworthy and simplistic accounts. Having few social connections also implies a difficulty in amplifying and propagating messages. In other words, only few users can read – and possibly re-share – what these bots post.



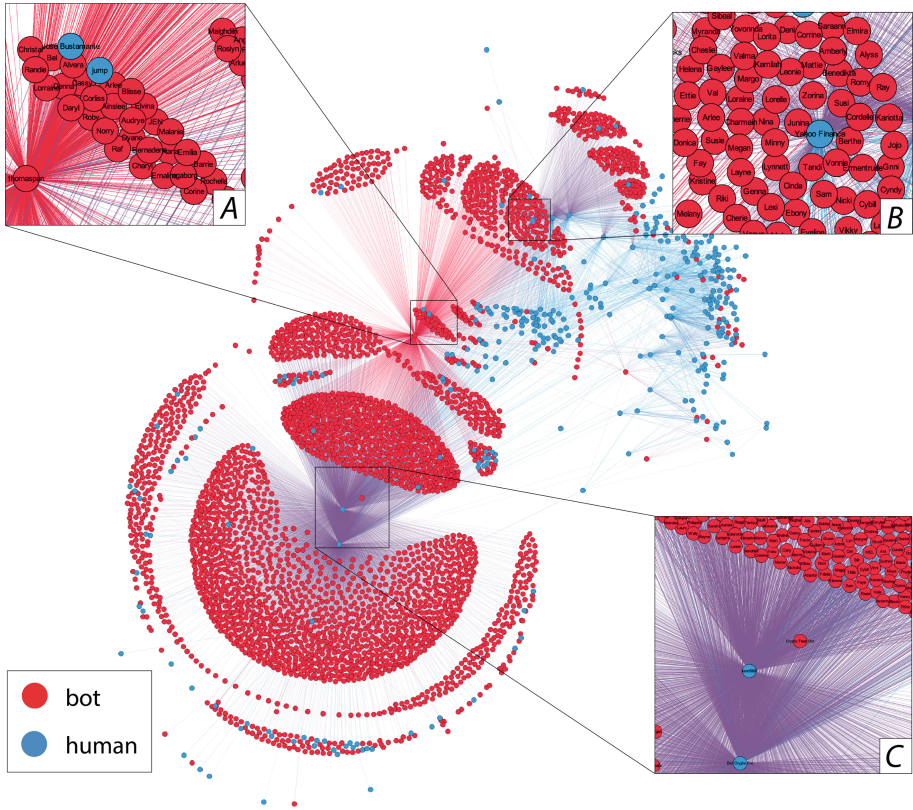
**Fig. 8.** Distribution of the number of tweets per user. Although bots and humans feature similar volumes of shared tweets, financial bots tend to retweet rather than to create original content.



**Fig. 9.** Edge weights distribution in the user similarity network. The distribution approximates a power law with 2 notable exceptions, marked with red circles. (Color figure online)

### 5.3 Tweeting Characteristics of Financial Bots

Studying general profile characteristics, as we have done in the previous subsection, allows to assess the credibility and trustworthiness of financial bots (or lack thereof). Instead, in the remainder of this section we focus on their tweeting



**Fig. 10.** A portion of the user similarity network. Nodes are colored according to their classification as bots or humans. Several different botnets are clearly visible as dense clusters. Bots are typically connected to a single human-labeled user, with which they share the majority of their mentioned companies.

activity. Our aim for the following analyses is to understand the likely target of financial bots as well as their inner organization.

We first analyze the *distribution of the number of tweets* posted by bots and humans, for each possible type of tweet (that is, original tweets, retweets and replies). As displayed in Fig. 8, financial bots and humans share a comparable total number of tweets. In other words, financial bots do not seem to post excessively (i.e., to spam), as other simplistic types of bot do [8], but instead they have an overall content production that is similar to that of humans. However, bots exhibit a strong preference for retweeting rather than for creating original content or for replying. Therefore, retweets are the primary mechanism used by financial bots to propagate content. It is worth noting however that the focus of

these financial bots is likely posed on the retweet itself, rather than on retweets as an efficient mean to rapidly reach broader audiences [17, 23]. This is because financial bots are characterized by few social relationships, as discussed in the previous section. As such, few users would be exposed to their retweets. This strategy, applied to the financial context, may nonetheless deceive trading algorithms listening to social conversations in search for hot stocks to invest in. As a consequence, synchronized mass-retweets of stock microblogs may contribute to artificially overstate the interest associated with specific stocks.

We conclude our analyses by studying the use of *cashtags* by bots and humans. Here, we are particularly interested in identifying groups of users that systematically tweet about the same stocks, because this might reveal the inner structure of financial disinformation botnets. Interesting questions are related as to whether we are witnessing to a single huge botnet or whether there are multiple botnets individually promoting different sets of stocks.

To answer these questions, we first build the bipartite network of users (comprising both bots and humans) and companies. In detail: Twitter users are one set of nodes, companies represent the other set of nodes, and a link connects a user to a company if that user mentioned that company in one of its tweets. This bipartite network is directed and weighted based on the number of times a user mentions given companies. In order to study similarities between groups of users, we then project our bipartite network onto the set of users. This process results in two users being linked to one another if they both mentioned at least one common company. The projected network, henceforth called *user similarity network*, is undirected and weighted. The weight of a link connecting two users measures the number of companies mentioned by both users.

For the sake of clarity, in the following we report results of the analysis of a subset of the user similarity network. In particular, Fig. 9 shows the distribution of edge weights in the considered portion of the network. As shown, the edge weights distribution approximates a power law, with 2 notable exceptions marked in figure with red circles. Peculiar patterns that deviate from the general law for specific portions of a network distribution have been previously associated with malicious activities [20]. For this reason, we focus subsequent analyses on the network nodes and edges that are responsible for the deviations highlighted in Fig. 9. In particular, Fig. 10 shows the resulting user similarity network, visualized via a force-directed layout, where nodes are colored according to their classification as bots or humans. Interestingly, the vast majority of nodes in this network were previously labeled as bots, during our bot detection step. This explains the deviations observed in the edge weights distribution plot. In addition, the vast majority of bots is organized in a few large distinct clusters. Each cluster of bots is typically connected to a single human-labeled user, with which bots share the majority of their mentioned companies. In other words, the visualization of Fig. 10 clearly allows to identify several distinct botnets, as well as the accounts that they are promoting. The few human-labeled users of the network show more diverse patterns of network connections. They are not organized in dense clusters and, in general, feature more heterogeneous connectivity

patterns with respect to the bots, confirming previous literature results [12]. A few interesting portions of the network are magnified in the **A**, **B** and **C** insets of Fig. 10, and allow to identify the users to which the botnets are connected (including the *@YahooFinance* account visible in inset **B**), as well as the similar names (e.g., all English and female, as shown in insets **A** and **B**) of the accounts that constitute the botnets.

## 6 Discussion

Results of our investigations highlighted the widespread existence of financial disinformation in Twitter. In particular, we documented a speculative campaign where many financially unimportant (low-cap) stocks are massively mentioned in tweets together with a few financially important (high-cap) ones. In previous work, this fraud was dubbed as *cashtag piggybacking*, since the low-value stocks are piggybacked “on top of the shoulders” of the high-value ones [14]. Considering the already demonstrated relation between social and financial importance [22], a possible outcome expected by perpetrators of this advertising practice is the increase in financial importance of the low-value stocks, by exploiting the popularity of high-value ones. To this regard, promising directions of future research involve assessing whether these kinds of malicious activity are correlated to, or can influence, stock prices fluctuations, the stock market’s performance, or even the macroeconomic stability [14].

Analyses of suspicious users involved in financial discussion spikes, revealed that the speculative campaigns are perpetrated by large groups of coordinated social bots, organized in several distinct botnets. We showed that the financial bots involved in these manipulative activities present very simple accounts, with few details and social connections. Among the available details, many signs of fictitious information emerge, such as the suspicious profile descriptions where some financial keywords are mixed with other unrelated content. The simplistic characteristics of these bots, their relatively recent and bursty creation dates, and their limited number of social connections give the overall impression of untrustworthy accounts. The financial social bots discovered in our study have different characteristics with respect to the much more sophisticated social bots recently emerged in worldwide political discussions [8,11]. Financial social bots thus appear as a rather easy target for automatic detection and removal, as also confirmed by the large number of such bots that has already been banned by Twitter.

Based on these findings, we conclude that these bots should not pose a serious threat to human investors (e.g., noise traders) looking for fresh information on Twitter. However, the aim of financial bots could be that of fooling automatic trading algorithms. In fact, to the best of our knowledge, the majority of existing systems that feed on social information for predicting stock prices, do not perform filtering with regards to possibly fictitious content. As such, these systems could potentially be vulnerable to coordinated malicious practices such as that of *cashtag piggybacking*. The fact that no study nor existing system actually



hunted financial bots before our present works, could also possibly explain the simplistic characteristics of these bots. In fact, it is largely demonstrated that recent social bots became so evolved and sophisticated as an evasion mechanism for the plethora of existing bot detection techniques [8]. In other words, financial bots could be this simple, just because nobody ever hunted them. If this proves to be the case however, we should expect financial bots to become much more sophisticated in the near future. A scenario that would pose a heavier burden on our side with regards to their detection and removal.

The user-centric classification approach that we adopted in this study demands the availability and the analysis of large amounts of data, and requires intensive and time-consuming computations. This is because, in order to assess the veracity of a discussion spike, all users involved in that discussion are to be analyzed. This could easily imply the analysis of tens of thousands of accounts for evaluating a single spike of discussion. On the contrary, another – more favorable – scenario could involve the classification of the discussion spikes themselves. In other words, future financial spam detection systems could analyze high-level characteristics of discussion spikes (e.g., their burstiness, the number of distinct accounts that participate, market information of the discussed stocks, etc.), with the goal of promptly detecting promoted, fictitious, or made up discussions. This approach, previously applied to other scenarios [28], is however still unexplored in the online financial domain. As such, it represents another promising avenue of future research and experimentation.

## 7 Conclusions

Our work investigated the presence and the characteristics of financial disinformation in Twitter. We documented a speculative practice aimed at promoting low-value stocks, mainly from the OTCMKTS financial market, by exploiting the popularity of high-value (e.g., NASDAQ) ones. An in-depth analysis of the accounts involved in this practice revealed that 71% of them are bots. Moreover, 48% of the accounts classified as bots have been subsequently banned by Twitter. Finally, bots involved in financial disinformation turned out to be rather simplistic and untrustworthy, in contrast with recent political bots that are much more sophisticated.

Our findings about the characteristics of fake financial discussion spikes as well as those related to the characteristics of financial bots, could be leveraged in the future as features for designing novel financial spam filtering systems. Hence, this work lays the foundations for the development of specific – yet still unavailable – methods to detect online financial disinformation, before it harms the pockets of unaware investors.

**Acknowledgements.** This work is partially supported by the European Community’s H2020 Program under the scheme INFRAIA-1-2014-2015: **Research Infrastructures**, grant agreement #654024 *SoBigData: Social Mining and Big Data Ecosystem* and the scheme INFRAIA-01-2018-2019: **Research and Innovation action**, grant agreement #871042 *SoBigData++: European Integrated Infrastructure*

for *Social Mining and Big Data Analytics*, and by the Italian Ministry of Education and Research in the framework of the *CrossLab Project* (Cloud Computing, Big Data & Cybersecurity), Departments of Excellence.

## References

1. Albadi, N., Kurdi, M., Mishra, S.: Hateful people or hateful bots? Detection and characterization of bots spreading religious hatred in Arabic social media. In: Proceedings of the ACM on Human-Computer Interaction (HCI), vol. 3, no. CSCW, pp. 1–25 (2019)
2. Allem, J.P., Ferrara, E.: Could social bots pose a threat to public health? *Am. J. Public Health (AJPH)* **108**(8), 1005 (2018)
3. Berger, J.M., Morgan, J.: The ISIS Twitter census: defining and describing the population of ISIS supporters on Twitter. *The Brookings Project on US Relations with the Islamic World*, vol. 3, no. 20, pp. 1–4 (2015)
4. Beskow, D.M., Carley, K.M.: Its all in a name: detecting and labeling bots by their name. *Comput. Math. Organ. Theory (CMOT)* **25**(1), 24–35 (2019). <https://doi.org/10.1007/s10588-018-09290-1>
5. Bessi, A., Ferrara, E.: Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* **21** (2016). <https://doi.org/10.5210/fm.v21i11.7090>
6. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. *J. Comput. Sci. (JCS)* **2**(1), 1–8 (2011)
7. Brachten, F., Stieglitz, S., Hofeditz, L., Kloppenborg, K., Reimann, A.: Strategies and influence of social bots in a 2017 German state election - a case study on Twitter. In: *The 28th Australasian Conference on Information Systems (ACIS 2017)* (2017)
8. Cresci, S.: Detecting malicious social bots: story of a never-ending clash. In: Grimme, C., Preuss, M., Takes, F.W., Waldherr, A. (eds.) *MISDOOM 2019*. LNCS, vol. 12021, pp. 77–88. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-39627-5\\_7](https://doi.org/10.1007/978-3-030-39627-5_7)
9. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: Fame for sale: efficient detection of fake Twitter followers. *Decis. Support Syst. (DSS)* **80**, 56–71 (2015)
10. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. *IEEE Trans. Dependable Secure Comput. (TDSC)* **15**(4), 561–576 (2017)
11. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: The paradigm-shift of social spambots: evidence, theories, and tools for the arms race. In: *The 26th International Conference on World Wide Web Companion (WWW 2017 Companion)*, pp. 963–972 (2017). IW3C2
12. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: Emergent properties, models and laws of behavioral similarities within groups of Twitter users. *Comput. Commun.* **150**, 47–61 (2020)
13. Cresci, S., Lillo, F., Regoli, D., Tardelli, S., Tesconi, M.: \$FAKE: evidence of spam and bot activity in stock microblogs on Twitter. In: *The 12th International AAAI Conference on Web and Social Media (ICWSM 2018)*. AAAI (2018)
14. Cresci, S., Lillo, F., Regoli, D., Tardelli, S., Tesconi, M.: Cashtag piggybacking: uncovering spam and bot activity in stock microblogs on Twitter. *ACM Trans. Web (WEB)* **13**(2), 11:1–11:27 (2019)

15. Cresci, S., Minutoli, S., Nizzoli, L., Tardelli, S., Tesconi, M.: Enriching digital libraries with crowdsensed data. In: Manghi, P., Candela, L., Silvello, G. (eds.) IRCDL 2019. CCIS, vol. 988, pp. 144–158. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-11226-4\\_12](https://doi.org/10.1007/978-3-030-11226-4_12)
16. Ferrara, E.: Disinformation and social bot operations in the run up to the 2017 French Presidential election. *First Monday* **22**(8) (2017). <https://doi.org/10.5210/fm.v22i8.8005>
17. Giatsoglou, M., Chatzakou, D., Shah, N., Faloutsos, C., Vakali, A.: Retweeting activity on Twitter: signs of deception. In: Cao, T., Lim, E.-P., Zhou, Z.-H., Ho, T.-B., Cheung, D., Motoda, H. (eds.) PAKDD 2015. LNCS (LNAI), vol. 9077, pp. 122–134. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-18038-0\\_10](https://doi.org/10.1007/978-3-319-18038-0_10)
18. Hentschel, M., Alonso, O.: Follow the money: a study of cashtags on Twitter. *First Monday* **19**(8) (2014). <https://doi.org/10.5210/fm.v19i8.5385>
19. Hwang, T., Pearce, I., Nanis, M.: Socialbots: voices from the fronts. *Interactions* **19**(2), 38–45 (2012)
20. Jiang, M., Cui, P., Beutel, A., Faloutsos, C., Yang, S.: CatchSync: catching synchronized behavior in large directed graphs. In: The 20th SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2014), pp. 941–950. ACM (2014)
21. Lee, S., Kim, J.: Early filtering of ephemeral malicious accounts on Twitter. *Comput. Commun.* **54**, 48–57 (2014)
22. Mao, Y., Wei, W., Wang, B., Liu, B.: Correlating S&P 500 stocks with Twitter data. In: The 1st International Workshop on Hot Topics on Interdisciplinary Social Networks Research (SIGKDD 2012 Workshops), pp. 69–72. ACM (2012)
23. Mazza, M., Cresci, S., Avvenuti, M., Quattrociochi, W., Tesconi, M.: RTbust: exploiting temporal patterns for botnet detection on Twitter. In: The 11th International ACM Web Science Conference (WebSci 2019), pp. 183–192. ACM (2019)
24. Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., Tesconi, M., Ferrara, E.: Charting the landscape of online cryptocurrency manipulation. *arXiv preprint arXiv:2001.10289* (2020)
25. Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.C., Flammini, A., Menczer, F.: The spread of low-credibility content by social bots. *Nat. Commun.* **9**(1), 4787 (2018)
26. Stella, M., Ferrara, E., De Domenico, M.: Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Natl. Acad. Sci.* **115**(49), 12435–12440 (2018)
27. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: detection, estimation, and characterization. In: The 11th International AAAI Conference on Web and Social Media (ICWSM 2017). AAAI (2017)
28. Varol, O., Ferrara, E., Menczer, F., Flammini, A.: Early detection of promoted campaigns on social media. *EPJ Data Sci.* **6**(1), 1–19 (2017). <https://doi.org/10.1140/epjds/s13688-017-0111-y>
29. Viswanath, B., et al.: Strength in numbers: robust tamper detection in crowd computations. In: The 2015 ACM Conference on Online Social Networks (COSN 2015), pp. 113–124. ACM (2015)
30. Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., Blackburn, J.: Disinformation warfare: understanding state-sponsored trolls on Twitter and their influence on the Web. In: The 2019 World Wide Web Conference Companion (WWW 2019 Companion), pp. 218–226 (2019). IW3C2
31. Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., Blackburn, J.: Who let the trolls out? Towards understanding state-sponsored trolls. In: The 11th International ACM Web Science Conference (WebSci 2019), pp. 353–362. ACM (2019)



# Cyber Risks in Social Media

Linda R. Wilbanks(✉)

Towson University, Towson, MD 21252, USA

lwilbanks@towson.edu

**Abstract.** Social media is an internet tool often used by businesses to communicate with customers, clients, vendors and other businesses. It is used for financial transactions with customers and to pay employees and manage the company. But social media and the internet are fraught with risks that businesses should be aware of, risks that could impact their company, their employees and their clients. Companies need to be aware of what the social media risks are and mitigations that can be taken to reduce those risks. While social media threats cannot be removed, actions can be taken to reduce the probability that they will be successful and to reduce the impact if they succeed. The impact of a threat gaining internet access to a business could cripple a company financially, others could damage relationships, but all would impact the company's ability to grow and thrive.

**Keywords:** Cyber risk · Social media risk · Social media cybersecurity

## 1 Introduction

Information systems are subject to serious threats by cyber criminals' exploitation of known and unknown vulnerabilities resulting in the compromise of the confidentiality, integrity, or availability of the company's information. Threats to information and information systems can include purposeful attacks, environmental disruptions, and human errors resulting in damage to the company. Many of these attacks succeed by exploiting both known and unknown social media vulnerabilities. Most businesses have a program to manage cyber security risks, but companies may not consider social media usage a risk, it is considered part of doing business.

Any company utilizing the internet is subject to a cyber-attack; social media usage increases the probably of successful cyber-attack. The economic costs of cyber-attacks exceed those associated with natural disasters. The costs of data breaches are expected to amount to US \$2.1 trillion globally by 2019. Most estimates of cyber-attack costs do not include the harms associated with damage and destruction of data, lost productivity, theft of intellectual property, disruptions to the business and reputational harm. Including those costs, it is estimated that cybercrimes costs the world US \$3 trillion in 2015 and that is expected to increase to US \$6 trillion annually by 2021 [10].

On April 10, 2018, in a hearing held in response to revelations of data harvesting by Cambridge Analytica, Mark Zuckerberg, the Facebook chief executive, faced questions from senators on a variety of issues, from privacy to the company's mishandling

of data. This was Mr. Zuckerberg's first appearance before Congress, prompted by the revelation that Cambridge Analytica, a political consulting firm linked to the Trump campaign, harvested the data of an estimated 87 million Facebook users to psychologically profile voters during the 2016 election. Zuckerberg was pressed to account for how third-party partners could take data without users' knowledge. This hearing made the cyber risks associated with social media visible and public.

Businesses need to understand their cyber risks associated with the use of the internet. The company needs to identify the threats and the vulnerabilities of their networks and to their data. The additional risks of social media usage need to be identified and then the impact assessed if the threat is successful. Can the company withstand the impacts of a cyber-attack through social media? Once the threats, vulnerabilities and risks are understood, mitigations need to be identified to reduce the risk to an acceptable level. Risk can never be totally mitigated, cyber threats are dynamic, continually changing. Hence the company's management and the cyber security team need to identify feasible mitigations and then implement them. But the first step is understanding what is cyber risk, what are the cyber risks associated with social media, and what are possible mitigations.

## **2 Cyber Security, Risk and Risk Management**

### **2.1 Cyber Security**

Cyber Security plays an important role in social media, managing the information technology risks, ensuring the confidentiality (information is not released without authorization or knowledge) and the integrity (information is not altered) of the data. The goal is to maintain the confidentiality and integrity at an acceptable level, but on social media, this has been proven difficult, if not impossible.

### **2.2 Risk**

Risk is the likelihood that a loss will occur. Losses occur when a threat exposes a vulnerability. To identify risks, first identify the threats and vulnerabilities and then estimate the likelihood of a threat exploiting a vulnerability. After understanding the risks, then take appropriate actions to mitigate the risks [16].

### **2.3 Cyber Security Risk**

Cyber security risk is the risk to an organizational operation's mission, reputation, assets, and individuals due to the potential for unauthorized access, use, disclosure, modification or destruction of information and/or information systems. Cyber risks are those that arise through the loss of confidentiality, integrity and/or availability of the data and/or networks. The risks also relate to the authentication to access/view the information. Cyber risks, like any other type of risk, cannot be eliminated, they must be managed [14].

**Threats.** A threat is any activity that represents a possible danger, any circumstance or event with the potential to adversely impact the confidentiality, integrity or availability of personal data or business assets. Threats cannot be eliminated, they are always present, but it is possible to reduce the potential for a threat to occur and reduce the impact of a threat [16].

**Vulnerabilities.** A vulnerability is a weakness in an information system, system security procedures, internal controls, or implementation that could be exploited by a threat [16].

**Loss.** It may be difficult to associate a loss with social media but a variety of losses can occur. These losses associated with social media relate to the businesses who use social media. There are risks to corporate data and reputations.

**Risk Impact.** The impact of a risk is the magnitude of harm that can be expected to result from the consequences of a threat exploiting a vulnerability resulting in unauthorized disclosure, modification, destruction, or loss of data. The loss can result in a compromise to business functions or assets that adversely affects the business or a person [14].

## 2.4 Risk Management

Risk management is the recognition that you cannot protect against everything – it is about prioritization and the acceptance of risk. Identification of the risks and their management are the keys to appropriately protecting networks and data. In order to identify what the social media risks are, the concepts of social media that make it vulnerable to cyber threats must be identified. The risks can then be determined and possible mitigation strategies can be identified [18].

## 3 Risk Concepts of Social Media

Social media is in many respects an unstoppable cultural force, ubiquitous and powerful. It is interactive computer-mediated technologies that facilitate the creation and sharing of information and business via virtual communities, networks and offices. Social media plays an important role in business in every industry, such as the ability to connect with their customers, making major corporations more accessible to the consumer. But this ease of accessibility creates risks, and it is better to manage those risks than to ignore them and then address the consequences.

Organizations that fail to harness the potential value of social networking run the risk of lagging behind and losing business to competitors. Social media enables companies to listen to and learn from satisfied and dissatisfied customers regarding their ideas, experiences and knowledge and they offer business an opportunity to reach out and proactively respond to views and reactions. These significant benefits also create significant risks that businesses need to be aware of, taking precautions and implementing mitigations if deemed appropriate [3].

Compounding the mitigation problem is that social media risks are difficult to quantify. Most corporate initiatives are not approved without a strong business case, comprehensive cost/benefit analyses are not always feasible. According to a survey that looked

at the corporate social media risks, almost three out of four executed said their companies believe the risks can be mitigated or avoided, but another 13% stated they felt their company does not currently have any appreciable risks [1].

## 4 Social Media Threats and Vulnerabilities

While social media is a powerful tool for interacting with others, many people and social media organizations have jumped into using it without considering the risks. The threats users face can affect their safety and job; businesses can also face many diverse social media threats such as hackers and employees [2].

### 4.1 Threats

Risk is the likelihood that a loss will occur. Losses occur when a threat exposes a vulnerability. To identify risks, identify the threats and vulnerabilities. In order to understand the risks to social media, the first step is to identify the threats.

A threat is any activity that has the potential to adversely impact the confidentiality, integrity or availability of business assets. There are different types of threats, such as a natural disaster (e.g. hurricane, earthquake) which can disrupt a business's operations. While a disruption to the ability to access social media has the potential to impact business operations, the threats we will discuss are the human, software and network threats [13, 17].

**Users threats.** The most viable threats related to social media is the users. Social media is driven by people who create the content and control usage of social media. They can unintentionally make statements that can harm businesses through misinformation. The threat is also the intruder intent on using the information for espionage, damaging reputations (personal, professionally and to the businesses) or otherwise causing a negative impact to a business. The threat can be criminals, intent on fraud or theft, or a focused and persistent attacks on a specific target. Or the threat could be hackers whose intrusions on a social media site may be curiosity or a malicious intent, to do damage, steal or commit fraud. It can be disgruntled employees' intent to get even or to damage the reputation of people or businesses. The threat may be a result of greed, espionage or anger, or desire to do damage. By definition, people are unpredictable, making calculating the risk associated with this threat a very difficult task [16, 17].

**Malware Threats.** Malware is another serious threat in social media as it can be purposefully planted in a social media site. Malware is any software intentionally designed to cause damage to a computer, server, client, or computer network, it is a software threat to social media sites. Malware can be placed on a computer, unknowingly to the business. This then allows for the someone to access the computer, utilize social media and poses a threat to the business [12].

Malware can be in many forms:

- A virus is commonly spread through file sharing, web downloads, email attachments.
- A worm which can crawl through networks without human interaction.

- A Trojan designed to specifically extract sensitive data from network, and may take control of infected system, or open up back door for attacker to access later.
- Spyware which can infect web browsers making them nearly inoperable, or secretly record behavior and usage patterns.

Social media is a compendium of many highly accessible media, many with billions of users – corporate blogs, video sharing sites, social networks, microblogging tools among many others. These media leverage the power of the internet and mobile technologies to facilitate the creation, exchange, use and modification of user generated content. It is this user created data that leads to the vulnerabilities for businesses using social media [3].

In order for a risk to occur, a threat must exploit a vulnerability. In the use of social media, one vulnerability is the same as the threat, the people using social media, posting, commenting and sharing information. All of these transactions can allow a threat access to the system and data which can have negative consequences such as a business loss.

**System Configuration Vulnerability.** There are other vulnerabilities that can be exploited such as the way a system is configured and the system security components, such as firewalls and access controls. While these network configurations are critical for a business to manage as part of their cyber security program, they are also important to manage the vulnerabilities of employees using personal media devices, such as tablets, laptops and smartphones if the company allows these devices access their networks [5].

**Access Vulnerabilities.** Secure access to social media sites is a requirement for businesses to ensure their data and networks are not compromised. Vulnerabilities relate to the ability for an attacker to easily take control of systems and risking the data could be compromised. Any password or security question that is easily guessed is not a secure. In social media, if a password is stolen, personal and employer accounts could be accessed, the confidentiality and integrity of the data on social media can be compromised resulting in a negative impact to the business.

*Password Structure.* A vulnerability with social media is the improper use of passwords. Most social media sites require a user name and password to access the site. An extremely secure password, one that could not be easily guessed or identified through brute force, would be at least 16 characters in length, comprised of numbers, letters (upper and lower case) and special characters, and be changed at least every 3 months. This type of password is cumbersome for users to remember and could discourage a user from accessing that social media site. For easier access, many social media sites only require an 8-digit password, letters and numbers; some do require a special character. This type of password is a vulnerability due to short length and minimal complexity [12].

*Guessable Passwords.* A use of any of these passwords increase the security risks for social media accounts because they can be guessed by anyone trying to access the user's social media. For businesses that own social media sites, allowing weak passwords is a vulnerability. Businesses should implement a strong password policy, not allowing the use of these passwords.



A study done in 2017 of the most common passwords in 2016 showed these as some of the most common passwords:

- 123456789 – as many numbers as needed, repeat if necessary
- Qwerty – use the keyboard, this is the top row of letters
- 111111111 -pick any number or letter, use as many as required
- ????????? - same ideas as the number, use any special character
- Password, google, welcome – any common word [7]

Other common techniques for creating passwords are using friends, family or pet names and adding a birthday or number to meet the length and content requirement. Using the same password on all accounts is another vulnerability, which many social media users do, it's much easier to remember only one password. But this approach to password management creates the risk to all of the user's systems, once the common password is captured, the intruder has access to all of those systems. If employees access personal social media accounts utilizing company assets, they should be trained in strong password management [11].

*Password Management.* Another common vulnerability on social media is the use of security questions to change passwords or profiles. Many of the questions contain information that can be found within social media, such as a pet's name, a high school, or even the user's mother's maiden name. Many of these security question answers can also be found through pictures in offices and rooms or on calendars. Many social media sites use the same questions, such as high school, pets and family names. Once an attacker knows the answers to security questions, they may have access to multiple sites and data, creating a risk for the business that owns the social media site. The risk to businesses is also that if an attacker knows an employee's security question responses on personal accounts, it is highly likely they use the same security questions at work, thus potentially giving the attacker access to the business's data and/or networks [8].

## 5 Social Media Risks

For a business, a risk of social media is that they can receive negative exposure which can result in lost trust and lost revenues. There are additional serious business risks: Strategic, business, regulatory, legal, and market. If not effectively mitigated, these risks can lead to serious negative consequences including fraud, intellectual property loss, financial loss, privacy violations and failure to comply with laws and regulations. Social media is a powerful tool that gives organizations the ability to expand their brand value, but it can also tarnish a brand quickly.

### 5.1 Potential Impacts

For businesses to maintain a secure network, they must understand the risks and constantly monitor all types of social media; the risks are considerable. According to security

experts, a majority of current attacks on industries use the social platforms as a delivery mechanism. These attacks can have a negative impact on business. Some financial institutions have had to shut down social media forums due to unanticipated negative feedback; the stock markets have been buffeted by fraudulent social network postings, other businesses have suffered brand damage due to the power of social media, the ability to send negative impressions almost immediately around the world [1, 9].

## 5.2 Reliability

The extensive use of social media by all enterprise, companies and across all nations, implies that social media has become perceived as a trustworthy source of information, that the data and connections are reliable. The continuous interpersonal connectivity on social media may lead to businesses regarding recommendations as indicators of the reliability of information sources. This acceptance on the reliability of the data creates vulnerabilities and exposes social media users to the acceptance of inaccurate information as fact. The threat is that businesses may act on information gleaned from social media with the incorrect assumption of reliability, creating the risk for all actions taken based on the information [15].

## 5.3 Ownership

Social media content is generated through interactions done by the users on the site. A question that arises is the ownership of the content on social media platforms. It is generated by the users and hosted by a company. The threat is that once the data, accurate or inaccurate, is posted to a social media site, it is not clear who has the responsibility to remove it. Recognizing that once information is posted to a social media site it may be permanent, even if inaccurate, is a risk that all users of social media users tend to accept, or just ignore. A business can take down a social media site, but some posts made but disgruntled employees or customers, many not be that easy to remove [2].

## 5.4 Risks

When using social media, the risks to a business are very diverse in the vulnerability, threat and impact. The spectrum of risks is very large but some of the major risks to business's using social media are discussed here. Attacks against social media networks are able to leverage a user's contacts, location, and business activities. This information can then be used to develop targeted advertising campaigns toward specific users, or even increase crime in the virtual or real world. The information can also make businesses vulnerable to a personal attack, such as negativity campaign. These are the risks that business accept when using social media users, as they continue to post information about their company and employees; social media users are vulnerable [9, 11].

**Reputational Risks.** A social media mishap can permanently damage the business's reputation. This risk has increased due to the rise of online and social media usage and can happen very quickly. Reputation damage may result from inappropriate employee

behavior, setting unrealistic product or customer expectations, rogue tweets of inappropriate messages, intended for internal or personal use or inability to measure up to the openness, straight talk and transparency expected by customers and prospects seeking to engage. Customers or other parties can use social media to say negative things about the company, impacting their reputation [3].

**Compliance or Regulatory Violations Risks.** Many organizations use social media as a business tool for increased visibility. There are risks of communicating data and information that violates applicable laws and regulations, including infringement of trademarks and copyrights, data security issues, and violations of privacy rights. There also are potential risks based on the organization's retention regulations or e-discovery requirements. Commentary on company performance that could impact the stock price or violate insider trading, "quiet period" and other rules under applicable securities laws is also a risk [3].

**Rival Enterprise Risks.** Enterprises may also leverage social platforms for "reconnaissance attacks" either directly or through third parties to collect valuable user and organization information about rivals. This data can provide businesses with a competitive edge in future business endeavors, and these attacks are expected to increase [1].

**Social Engineering Risks.** Employees in almost all organizations have a social media presence; each platform has the potential risk of providing a hacker information about employees, creating the risk of a social engineering attack on the company. The release of confidential information and activities about the company may also be used to the detriment to the company [9].

**Unauthorized System Access Risk-Phishing.** One major risk to all organizations is criminals is to get unknowing individuals to disclose personal information while posing as a fictitious representative of a legitimate professional or company – exercising a phishing attack. Employees fall victim to email, phone call and website phishing schemes, resulting in attackers having access to the business's network resulting creating an enterprise-level risk [9].

**Unauthorized System Access Risk - Viruses and Malware Risks.** Hackers' ability to penetrate the organization's network via social media websites, and accounts utilized by employees, creating the risk of viruses and malware entering the network. This creates the risk of inappropriate release, leakage or theft of information strategic to the company and exposure of company network and systems to viruses and malware due to human error, phishing scams, sophisticated attackers and identify thieves. This type of attack can also lead to a denial of service attack, where the networks is not accessible due to the worm or virus consuming all resources [3].

**Release of Personally Identifiable Information (PII) Risk.** Social engineering, phishing, viruses and malware all have a negative impact on the operations of a company, but they also create a risk that personal information about employees or clients could be released and cause harm to those people. Personally Identifiable Information (PII) is any

information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information [17].

**Social Media Information Permanence.** The content posted can and will be held against people. It is difficult to remove videos, photos, tags and posts that others put on social media, risking that once information is posted, it cannot really be removed. False information, and the release of private, personal information is also a risk people accept when using social media. But a business needs to be aware of these risk employees face as it can also impact the company depending on the information in the posts [2].

**Digital Footprint Risk.** When Internet users visit various Web sites, they can leave behind evidence of which sites they have visited. This collective, ongoing record of one's Web activity known a digital foot print. The information posted on social media and the digital footprint of users' interactions with a company's social media can be beneficial to the company. But if this information is utilized by an attacker or a competitor, there is a risk that it could have a negative impact on the company due to espionage or redirection of customers to a competitor's social media [3].

**Security Breach Risks.** Social engineering, phishing, viruses and malware all have a negative impact but they also are types of security breaches. A security breach is an event that compromises the confidentiality, integrity, or availability of an information asset. A security breach may result in a data breach with a confirmed disclosure of information to an unauthorized party [6].

**Identify Theft Risk.** The risk of identity theft is usually associated with social media users, but a company is also subjected to this threat. The concept is simple, the process can be complex, but the consequences for the company can be quite severe. The concept is that a new company takes on the identify of another, using a new website that appears to be the website of the original company. This usually is for the purpose of capturing customers making purchases, which allows the fake company to capture credit cards purchases as well as the customer's name or even driver's licenses, birth certificates and social security cards. This user information captured can then be sold as additional revenue to the fake company. The original company faces reputational risks since to the consumer their credit card was changed for merchandise that will never be received [3].

## 6 Cybercrime

The virtual world can be a wonderful place, allowing businesses to interact with networks of people, have a virtual salesforce, provide training with groups or one-on-one help for customers. But not everyone or everyplace is safe. The Internet can create overconfidence and leaving the business vulnerable to predators, charlatans, and scams. There are many threats on the Internet, and even though a business may have a rigorous security program, they still face risks on social media. Social media is about sharing information, promoting business, engaging others, and having trust. Users often work on good

faith in the links clicked, the programs installed, requests that are made, and questions asked. Unfortunately, this also allows cyber criminals to take advantage of companies and employees [2].

## 6.1 Cyber Criminals

Cybercrime is a criminal activity that involves computers and the Internet. There are many threats on the Internet, since social platforms rely on sharing information, engaging others, and having trust. Businesses utilize social media to engage, connections, clients and to increase revenue. Users trust the companies hosting the social media sites and the users of them. The expectation is that the links, the programs, requests that are made, and questions asked are valid and without threats. Unfortunately, this is not always the case. Cyber criminals target information that is of value to them, such as bank accounts, credit cards, or intellectual property that can be converted into money. They often are structured to operate as any well-run, legitimate business with experts specialized in each area and position. This creates the risks of who do you trust? [2].

*“Cyber criminals run rampant across every social network today. We often see headlines about social marketing fails and celebrity account hacks, but they’re just the tip of the iceberg. Far more nefarious activity takes place across these social channels, while most organizations remain oblivious and exposed. Companies’ poor social media security practices put their brands, customers, executives, and entire organizations at serious risk” [5].*

## 6.2 Cyber Criminal Threats

Consumers implicitly trust people’s activity on social media thus creating a risk with the extensive data available for cyber criminals. Attackers now have incredibly broad reach and can easily manipulate users and execute a variety of widespread cyber-attacks and scams, including everything from social engineering to exploit distribution to counterfeit sales to brand impersonations, account takeovers, customer fraud, and much more [5].

There is a threat to businesses that a cybercriminal may use online resources to obtain information about the business, as a medium to contact potential victims (employees), gain access to systems, and/or damage data and bring down systems. The types of offenses will vary, but they will always involve technology and connectivity. The targets of cybercrime may be computers or people. A cybercriminal may focus an attack on systems using viruses or other malicious code, or by directly hacking a system to gain unauthorized access. They may also focus their attention on employees to steal their identity, commit fraud, stalk, or bully them or to gain information on their employer. It may also be a precursor for other offenses, where the crime is initiated online, but later physical access or impact [2].

While social media sites create completely new cyber threats, they also substantially amplify the risk of existing ones. From reconnaissance to brand hijacking and threat coordination, cyber criminals use social media to boost the effectiveness of their attacks. Social media risk isn’t solely about brand and reputation damage but is a cyber-security threat that can lead to major data breaches, numerous compliance issues, and large

amounts of lost revenue due to fraud and counterfeit sales, along with a many of other risks [5].

### 6.3 Fraud Threat

Fraud is a common risk of social media. As more people utilize social media as a conduit for selling and buying, there is a greater opportunity for fraud. Fraud does not require the technical expertise of hacking, most internet fraud does not rely on in-depth technological expertise, internet fraud merely used the computer as venue.

Categories of social media fraud include failure to send the merchandise, sending some of lesser values, failure to deliver in a timely manner and failure to disclose information about the item. There can also be hidden chargers or bait-and-switch, offering a fun game or quiz. The site entices the user to enter personal information and cell number which is then captured and entered as a subscription to dubious services. Online auctions, such as eBay pose risks. They are a wonderful way to find merchandise at very good prices, however any auction site can be fraught with risks of not getting the merchandise ordered or the quality is not what was expected [4, 7].

Another area of social media fraud is perpetrated by tricking the user to click on a link which will take them to their social media page. The fraudster can then capture the personal account information, password resulting in total control of the account. This occurs because both the email and landing page were fake. That link went to a page that only looked like the intended social site. Phishing – tricking someone to release personal information, account names and numbers, and passwords, is often used as a component of this type of fraud. All businesses are at high risk for phishing which then can result various attacks such as phishing and malware. Clicking on shortened URLs is also a risk as this type of URL hides the location. The location that the user is redirected to could install malware on the computer [3].

## 7 Reducing Social Media Risk

Although the threats on social media will always exist, a company can reduce the probability of a success attack through a comprehensive corporate approach to managing the social media risks.

A professional risk culture is needed that is attuned to both the significant benefits and the distinctive risks of social media, and putting in the place the appropriate risk mitigations that maximize the benefits yet minimize the risks associated with social media.

Social media risks can fall into two categories: technology and the people using it. Technical problems can be dealt with by implementing proper security and having the right tools in place to handle any issues that arise. When it comes to people, businesses need to change behavior through policies, training, and communication. Users need to be aware of the risks and know what is needed to mitigate the risks so they can safely enjoy the benefits of social media. The user-generated problems of social media are diminished as a person learns what they should and should not do [2].

Not understanding how social media is used and its role in business and the impact on businesses creates a risk. A business must understand the impact and risks of social media, recognizing the great power it can have, both positive and negative.

## **8 Social Media Security Program**

There is no prescribed rulebook for eliminating social media risk to an enterprise but there are some areas that should be part of a social media security program. These components will also strengthen the overall cyber security program.

### **8.1 Framework**

Both security professionals and marketers alike should start treating social channels like the dangerous security threat they truly are, and align strategies to effectively fend against the range of cyber techniques currently in use. The first step in the right direction is to develop a framework and assess the social risk plan. A business should understand the external risk environment and continuously monitor social channels for cyber threats outside of the company's direct control. This plan will include the threats, vulnerabilities and risks for a comprehensive enterprise mitigation [5].

### **8.2 Protocols**

Given the visibility, risk, and real-time monitoring and response required to effectively manage social media channels, companies must establish extensive protocols for use by their organization in order to engage with external channels. Companies representing themselves externally should engage the appropriate and authorized spokespersons and executives designated by their communications department in order to speak to, initiate, provide and/or post information within the social media space. While there are several key risk factors relative to social media, there are many rewards as well [2].

### **8.3 Vigilance**

Social media impacts all departments of the organization, and each department, whether it is IT, finance, marketing or human resources, has a different perspective of how social media can or will be utilized by the department on behalf of the organization. Control of the social media crisis that is bound to happen to any organization as its social media presence increases over time. This requires significant amounts of collaboration and communication across key departments and personnel at all times. To prevent social media breaches, protect user information, and secure company data, increased vigilance by individual users and enterprise policies are the best ways to ensure data breaches are avoided [9, 11].

## 8.4 Social Media Risk Program Components

In addition to the program components above, these components will also reduce the business's social media risks

- Approved social media platforms—assessing social media solutions outside of the company to demonstrate how the social media presence will contribute to achievement of the overall goal and, the platform's degree of utilization and security.
- Training and awareness—It is key for the organization to embody the proper use of social media etiquette as part of the on boarding training of all employees. Risk is not just about “regulated users.” Executive staff, legal teams and other key stakeholders can also use social media or internal collaborative tools inappropriately. Compliance teams need to broaden their supervisory lens to include these groups.
- Monitoring information release - The risks from social media misuse apply to all organizations. All publicly traded corporations need to consider how they are monitoring for potential disclosure of non-public information through all communications channels.
- Messaging - Social media and mobile messaging need to be built into ongoing content inspection processes. Content containing risk or value can live anywhere, and processes need to catch up with today's communications tools.
- Content inspection - Pre-review of social content can pay huge dividends. Tools that allow content to be inspected and approved prior to delivery or posting can generate an enormous return on investment [4].

## 8.5 Basic Security Concepts

In addition to the social media risk mitigates discussed previously, basic security concepts need to be implemented, some of which are listed below. When implemented, the actions will reduce the company's cyber risk, which in turn will reduce the risks to social media.

- Defaults – Remove or change defaults after installations or upgrades, especially passwords.
- Attack service – Reduce the attack surface, only run the services and servers needed, remove all unneeded services and protocols (reduces risk).
- Patch Management – Keep systems and access controls current. Ensure all applications and systems have the current patches installed, especially on personal devices.
- Intrusion prevention and detection – Install Firewalls and Intrusion detection programs to filter traffic coming in, monitor traffic. Activate Intrusion detections systems (IDS) & monitoring detect and log threats, cannot prevent threat, may modify environment to block after detected.
- Virus attacks – Install Antivirus software before systems are activated to reduce the probability of a successful virus, worm or spyware attack.
- Incident response planning– A faster response when an incident is detected can lead to a lower impact; containment first, identify source vulnerabilities and remove; determine impact.



- Access management – Require strong passwords, protect passwords and access ability, use strong security questions for user ID and password resets. Limit access to only those systems required. Not allowing password sharing on corporate accounts. Immediately changing or removing passwords for corporate accounts.
- Public WIFI – If its free your information may also be freely visible to anyone else on the network. Ensure employees know how to securely use public WIFI to connect to corporate and private networks [6].

## 9 Conclusion



The widespread use of social media and interactive collaboration tools for business purposes will continue to grow. Since prohibition is not effective, the only way to mitigate the governance risks of social media use will be to implement new security strategies for digital information capture and archival. Organizations should exercise awareness and vigilance. While there are many benefits to using social media, there are detriments and risks. Organizations utilizing social media platforms should have a clearly defined policy and communication plan in place. Considerations should be made for good governance, a communication plan in the event of a breach, a social media policy and monitoring tools. Organizations must identify the risks and determine the most appropriate enterprise-level social media risk mitigations.

## References

1. Accutture: A Comprehensive Approach to Managing Social Media Risk and Compliance (2014)
2. Cross, M.: Social Media Security: Leveraging Social Networking While Mitigating Risk. Elsevier Science & Technology Books (2014)
3. Deloach, J.: 10 Ways Social Media Impacts Your Risk Profile, Corporate Compliance Insights (2018)
4. Cruz, R.: Scrutiny Highlight Risks of Social Media. <http://www.rmmagazine.com/2019/09/03/sec-scrutiny-highlights-risks-of-social-media/>
5. Hayes, N.: Why social media sites are the new cyber weapons of choice. DarkReading (2016). <http://www.darkreading.com/attacks-breaches/why-social-media-sites-are-the-new-cyber-weapons-of-choice/a/d-id/1326802>
6. Keeper Security (2017). <https://keepersecurity.com/blog/2017/01/13/most-common-passwords-of-2016-research-study>
7. Khan, M.: Social Media Rewards and Risk. ISACA J. **4** (2017)
8. Korris, N.: ISACA 2 For Whom the Web Trolls: Social Media Risk in your Organization, InterVeritas International. [http://www.isaca.org/Knowledge-Center/Blog/Lists/Posts/Post.aspx?ID=1037&utm\\_referrer=direct%2Fnot%20provided](http://www.isaca.org/Knowledge-Center/Blog/Lists/Posts/Post.aspx?ID=1037&utm_referrer=direct%2Fnot%20provided)
9. Kshetri, N.: The Economics of Cyber-Insurance, ComputingEdge, pp. 10–15 (2019)
10. McAfee, How Cybercriminals Target Social Media Accounts. <https://www.mcafee.com/us/security-awareness/articles/how-cybercriminals-target-social-media-accounts.aspx>
11. NIST Special Publication 800-39, Managing Information Security Risk Organization, Mission, and Information System View (2011)
12. Wilbanks, L.: What's your IT risk approach. IT Prof., 13–14 (2018)



# Misinformation in the Chinese Weibo

Lu Xiao<sup>1</sup> (✉)  and Sijing Chen<sup>2</sup> 

<sup>1</sup> Syracuse University, Syracuse, NY 13210, USA

lxiao04@syr.edu

<sup>2</sup> Wuhan University, Wuhan 430071, Hubei, China

sichen@syr.edu

**Abstract.** Social media users are increasingly influenced by misinformation and disinformation as the techniques offer affordances to rapidly spread information to large groups of people. Most of the existing studies about misinformation and disinformation are in the context of Western cultures, the influence of misinformation in Chinese context is underexplored. To fill this research gap, this study analyzed 26,138 Weibo posts that are marked as containing misinformation. We performed a frequency analysis of these posts' metadata and the top 50 frequent nouns, verbs, and adjectives in the dataset, and examined the sentiment in the content. Our results show that many posts that contain misinformation tactically target topics that Chinese people are already concerned about. The persuasion literature implies that these characteristics increase the persuasive power of the posts. With the forward-asking verbs are frequently used in the posts, one behavior that the receivers are persuaded to perform is to share these posts with the others, which can contribute to the virality of the misinformation. Another alarming finding is that a large proportion of our collected posts asked the receivers for help and the posts showed gratitude to acknowledge the forwarding and helping behavior. Based on the trust literature and the notion that trust as a social reality, we discuss the potentially severe negative impact these posts can impose on the society as they undermine Weibo users' trustfulness to others and to the social media platform.

**Keywords:** Misinformation · Weibo · Persuasion · Trust

## 1 Introduction

Verifying the validity of a social media post is challenging as the affordances of the technology allow people to rapidly spread information to large groups of people in an anonymous fashion. This challenge imposes a potential threat on the Internet - social media users are influenced by misinformation and disinformation instead of facts and evidence. Consider this scenario: an Internet user receives a Tweet (or Weibo if in China), without being able to verify the validity of the information in the article, what will the individual do? This user may not believe it or may ignore it. But the user may also be convinced with the content, retweet or forward it to others. Indeed, social media users are found to be persuaded by views that have no factual basis [1, 2], and with the

pervasiveness of the social media tools and the large number of users in participation, this threat can bring significant and negative impact to the society [3].

Researchers are paying more and more attention to this threat. Humprecht's study [4] explored the topics in fake news in four Western democracies (the US, the UK, Germany, and Austria) and found that the choice of topics in producing disinformation reflect the national information environment. Chadwick et al. [5] analysed four data sets they constructed during the 2017 UK election campaign discovered that the format of tabloid news plays an important role in the spread of misinformation and disinformation. To tackle the problem of misinformation and disinformation going viral, researchers also develop computational techniques that check and verify the facts or evidence in social media content in real time [6–8].

Most of the existing studies about false information and fake news are in the context of Western cultures, such as those mentioned in the previous paragraph. In a recent study by Kow et al. [9], Internet users' awareness of false information and fake news was explored in Eastern culture. Specifically, the researchers interviewed 21 Hong Kong residents. The authors found out that while most of their interviewees were aware of such information, most did not act on it. Nevertheless, the characteristics of the social media content that contains false information were not examined.

Contributing to fill this research gap, this study explores characteristics of misinformation content in the Chinese context by analyzing various aspects of Chinese Weibo posts that contain misinformation. This paper presents the current progress of this study. It first describes the collection of the Weibo posts, and then presents the features that we have explored. It discusses the implications of the findings and concludes with the future tasks.

## 2 Related Work

Attributes of social media communication make it challenging for a user to interpret someone's comment and to examine the truthfulness of the information. For example, a social media message—often intended to express one's views or to influence others—can be anonymous, from real people, or automatically generated, making it difficult to identify its source. Because of this challenge to interpret and evaluate a social media message, misinformation and disinformation can spread over social media easily, negatively influencing a large number of users and society [1, 2, 10].

To address the problem of misinformation and disinformation in social media, Western researchers have been exploring computational techniques that verify the validity of the information [6, 8], and automatically classify the truthfulness of the online information [7, 11]. Researchers also build open datasets for fact checking research, e.g., FEVER [12], LIAR [13], and Emergent [14].

Also in the Western context, the body of studies that attempt to characterize misinformation and disinformation is growing as well. For example, the choice of topics of the fake news is found to reflect the national information environment in some Western democracies such as US and UK [4]. In the context of UK, the use of tabloid news format is shown to play an important role in the spread of misinformation and disinformation [5].

Related to misinformation and disinformation study, the Chinese research community is interested in false rumor study. Defined by Dictionary.com, misinformation is ‘false information that is spread, regardless of whether there is intent to mislead’, disinformation is ‘deliberately misleading or biased information; manipulated narrative or facts; propaganda’, and rumor is ‘a story or statement in general circulation without confirmation or certainty as to facts’. False rumors are a type of misinformation, as indicated in Wikipedia’s page about misinformation. To our best knowledge, the first reported study of false rumor in the China context is by Yang et al. [15], which focused on the social media content. Since then, more and more scholars have turned their attention on the false rumor detection in social media content [16–19]. Most of these studies consider false rumor detection as a binary classification problem, incorporating various features afforded by the platform such as whether it contains a picture, video or URL links, the sentiment of the post, and the number of verbs. Recent studies are mainly concerned about identifying false rumors as early as possible, while keeping a reasonable detection accuracy [20, 21]. Based on literature review and expert interviews, [22] identified several cues for detecting misinformation from the post content (e.g., language errors, the URL links and references in the posts, the topic, etc.), the source of the post, and the patterns in its dissemination (e.g., the number of times it is forwarded). In the case study of “poisonous bean sprouts” where the misinformation is that some bean sprouts were injected with hormone and would be carcinogenic, [23] shows that there is a high rate of content repetition in fake posts and the combination of factual information and false rumors is common. While there is a good progress on the false rumor detection in the Chinese context, there is lack of studies that analyze the misinformation content, not to say the identification of characteristics of misinformation-contained posts. Although various features were applied in the machine learning approaches, how they offer insights on the detection of misinformation has not been investigated.

Contributing to fill this research gap, this study explores the characteristics of misinformation contained in the Chinese context by analyzing various aspects of Sina Weibo posts that contain misinformation. Sina Weibo is the largest social media platform in China. It includes features like those from Twitter or Facebook. It also enables users to post with a 140-character limit (increased to 2,000 as of January 2016 with the exception of reposts and comments) and insert graphical emoticons or attach their own images, music, and/or video files in every post.

### 3 Our Data Set: A Sample of Weibo Posts that Contain Misinformation

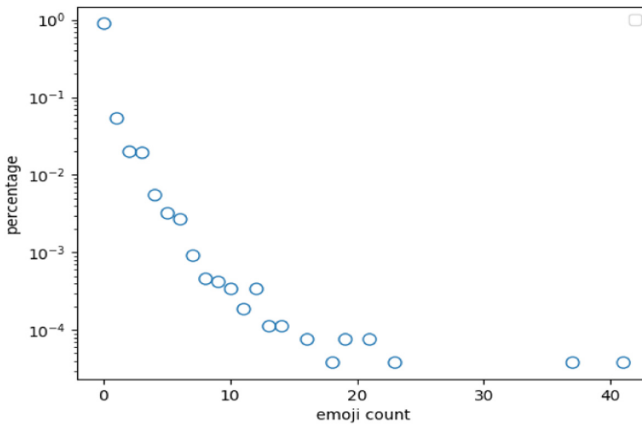
Weibo Community Management Center (<https://service.account.weibo.com/>) is a platform that officially announces and publishes posts containing false information and other violations in Sina Weibo. If Sina Weibo users suspect that a Weibo post contains misinformation, they can report to the company through this center. Once the Sina Weibo platform verifies that the reported posts indeed contain false information, it puts these posts online. Between 2nd and 5th of May, we scraped the reported posts from this source that are marked as containing false information by the Weibo Community Management Center. In total, we collected 26,138 Weibo posts posted between August 24th, 2011

and May 5th, 2019 and written by 24,127 unique users. Through parsing the original website of these posts, we scraped content and various metadata of these posts. Example metadata of a post includes timestamp, user information, number of forwards, number of comments, etc. The metadata also includes the links to images or videos if they are in the post. Collection scripts were implemented in Python 3.7 using the Selenium package. We removed the URLs and specific string patterns, like “@XXX”, “L微博视频”, “O网页链接” from the posts content in the analysis of word frequencies.

## 4 Data Analysis and Results

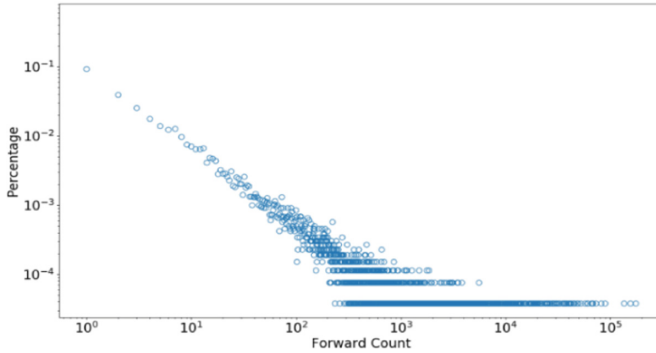
### 4.1 Frequencies of Weibo Features

Besides text, Weibo allows users to include images, videos, hyperlinks, hashtags and emojis in a Weibo post. In this dataset, about 45.5% of the posts use image and 38.1% of the posts have one image. About 10.2% of the posts contain external hyperlink, 5.5% include a video, and 9.6% use hashtag. Over 86% of the posts that use hashtag use just one hashtag. About 10.7% of the posts include emoji. In general, the frequency of emoji use obeys the heavy-tailed distribution (see Fig. 1).



**Fig. 1.** The frequency distribution of emoji use; the x axis represents the number of emoji use in a post, the y axis represents the percentage of posts after log transformation.

Weibo users also use various features to interact with the others, including mention, like, forward, and comment. About 13.6% of the posts in our dataset use mention. The frequency of mention use also obeys the heavy-tailed distribution. Additionally, about 49.2% of the posts are forwarded, 64.0% have comments, and 38.1% are liked. The frequency of these features follow the fat-tailed distribution. Figure 2 shows the distribution for the forwarding case.



**Fig. 2.** The frequency distribution of forward; the x axis represents the number of forwards after log transformation, the y axis represents the percentage of posts after log transformation.

## 4.2 Word Frequencies in the Posts

Our preliminary analysis included the examination of word frequencies in the posts. Interested in the content of these posts, we studied the top 50 nouns, verb, and adjectives as these word types reflect the content of the texts. The frequency range for these nouns is 834 to 6,439. From the analysis of these nouns, we observed that majority of the misinformation posts in our dataset is related to food/drink, people, and school/education. Interestingly, these topics are of great interest to the Chinese society in general. For instance, it is known that Chinese people are very health conscious [24, 25]. According to the survey of BCG [26], Chinese consumers have become the most health conscious in the world. Also, Chinese people perceive food safety as the most influential factor on their life quality nowadays [27].

A closer examination of the nouns in people category suggests that posts about child (孩子child, 小女孩little girl, 小孩kid, 宝宝baby) has appeared a lot in the dataset. We randomly selected 500 posts and found that the posts that contained these nouns were about child abduction and the category of “Location” is mainly about this context as well – “帮朋友转急找孩子求帮忙寻人启示。13940292999 线索酬金10万。今天上午一个三岁小女孩锦绣花园小区人拐走。小孩说出爸爸手机号。监控上看一个四十多岁男人抱走。现大人都急疯。知情者请告。万分感谢信息。朋友留意。联系人宁继春 13940292999 请转转” (Translation: (I’m) helping friends forward this in urgency. Child went missing 13940292999 a reward of 10,000 Yuan is provided if a clue is provided. This morning a three-year old girl was abducted by someone in JinXiu Garden area. The child can tell her parent’s cell phone number. The surveillance shows that a main in forties took her away. Her family is so desperate now. (If you have a clue) please tell. Appreciate it very much. Watch out for this friend. Contact Ning Jichun 13940292999”. A search of the collected posts shows that there are 4,542 out of 26,183 posts about child abduction stories similar to this. Human trafficking in children is still a very serious concern for many families in China [28, 29].

We found that the selected posts about school/education are mainly about this scenario – someone lost his/her exam permit on a national college entrance exam (NCEE) day. For example, the post “朋友捡粗心高考生准考证请相互转

下耽误孩子考试白娅倩考点市一中考场013座号11.准考证号 2024101311.请相互转发联系电话13830468131” (Translation: Friend found the NCEE exam permit left by a careless student. Please forward to others (about this to avoid) delay of the student’s exam. (The student’s name is) Bai Yaqian. Exam location: No. 1 High School of the City, Room 013, Seat No. 11. Permit No. 2024101311. Please forward to others. Contact Number: 2024101311). A search of the collected posts shows that there are 971 out of 26,183 posts about this scenario with the same or different locations and contact information. NCEE is commonly considered to be the toughest exam in the world in China because students are under great pressure during the period and the competition is very high - less than 10% of exam takers will be able to enroll in top-tier universities and less than 0.2% of the students will gain admittance into China’s top five universities. It is therefore imaginable what a disaster would be for a person if the person has lost his/her exam permit on the day. Therefore, our observation of these posts that contain misinformation suggests that the creators of such content tend to generate stories or offer misinformation around issues that people already have concerns about or are afraid of – be it about food safety, child trafficking, or NCEE.

A large percentage of the posts also asked the receivers to act, specifically, to provide information. Our reading of the selected posts suggest that these categories are all related to this action - Inform, Name, Number, and Appreciation. The two examples above illustrate this observation. Table 1 below presents the major categories we identified and the corresponding nouns and their frequencies.

The top 50 frequent verbs appeared in the dataset from 486 to 6,003 times. The actions indicated from these verbs are diverse, but several big categories did form. As shown in Table 2 below, the most frequent actions indicated in these Weibo posts are asking the receivers to forward the post (category “Forward”) and asking them to provide information as a way to help (category “Help” and “Inform”). Again, the two aforementioned examples about child abduction and NCEE exam permit also illustrate this aspect. From these most frequent verbs, we also observed that child abduction and showing gratitude are often mentioned in the posts, consistent with what we have found in the analysis of the 50 most frequent nouns.

Compared to the nouns and verbs, there are not many adjectives in this dataset. The top 50 frequent adjectives appeared from 72 to 1,328 times in the dataset. These adjectives are also very diverse making it challenging to group them into several conceptual categories. Nevertheless, there is one category that includes many counts of adjectives (4,075 in total): a category that indicates the mentioning of urgent/unexpected/severe or bad situation. The adjectives and their frequency counts are the follows: 紧急(urgent, 1328), 严重(severe, 1147), 麻烦(troublesome, 373), 突然(sudden, 240), 危险(dangerous, 210), 恐怖(horrible, 206), 不幸(unfortunate, 177), 慌乱(flustered, 74).

In summary, the analysis of the most frequent nouns, verbs, and adjectives imply that the posts that contain misinformation tend to inform the receivers about issues that bring anxiety or worry to the moment. They ask the receivers to inform others by ask them to forward the posts in their social network, and ask for the receivers’ help by asking them to provide information. In addition, the posts tend to thank the receivers in advance for these expected actions of the receivers.

**Table 1.** The categories of top 50 nouns

| Category                     | Nouns  |
|------------------------------|--|
| Contact                      | 线索(clue, 4133), 联系人(person to contact, 2640)   |
| Country                      | 中国 (China, 948), 日本(Japan, 834), 美国 (America, 1515)  |
| School/Education Related     | 小学(primary school, 3482), 学校(school, 1203), 准考证 (exam permit, 6439), 考点(exam location, 1308)   |
| Food/Drink Related           | 菠萝 (pineapple, 1976), 杆菌 (bacillus, 1394), 肉毒(Botox, 2359), 添加剂 (additive, 1441), 饮料 (beverage, 1383), 娃哈哈(Wahaha, 1379), 牛奶 (milk, 1356), 钙奶(calcium milk, 1354), 粒奶 (grain milk, 1350), 源果(Minute Maid, 1320), 美汁(Minute Maid, 1312), 旺仔(Hot-Kid Milk, 1486) |
| Hospital                     | 医院 (hospital, 1261), 妇幼保健(Maternal and Child Health, 4396)   |
| Inform/Information/Informant | 请告 (please inform, 1374), 知情者(insider, 4449), 信息(information, 3725)  |
| Location                     | 小区(residential community, 3105), 花园(garden, 3066), 锦绣(Jinxu, 917)  |
| Appreciation                 | 爱心 (warmhearted, 4444), 谢谢(thanks, 4249)   |
| Name                         | 宁继春 (Jichun Ning, 1290), 张静杰(Jingjie Zhang, 882)   |
| Number                       | 号码(number, 1083), 手机号码(phone number, 2413)   |
| People                       | 人民(people, 3884), 男人(man, 1787), 朋友 (friend, 1341), 孩子 (child, 4648), 大人(adult, 4468), 爸爸(father, 3962), 小女孩(little girl, 3732), 兄弟姐妹(brothers and sisters, 2852), 小孩(kid, 1504), 宝宝(baby, 1198), 妈妈 (mother, 1180), 家长(parent, 4042), 大家 (everybody, 951)     |
| Reward                       | 酬金(remuneration, 1047)   |
| Time                         | 今天上午 (this morning, 4244)  |



**Table 2.** The categories of top 50 verbs

| Category     | Verbs   |
|--------------|---|
| inform       | 说出(speak out, 4454), 告诉(tell, 1674), 通知(inform, 1277), 留意(look out, 3855), 注意(pay attention to, 1518), 提示(notice, 1324) |
| forward      | 转发(forward, 4002), 转转(forward, 1276), 求转(please forward, 1197), 接力(relay, 1023), 转下(forward, 796), 扩散(distribute, 4834) |
| help         | 帮忙(help, 6003), 帮助(help, 543)   |
| witness      | 看到 (see, 5020)  |
| abduct       | 拐走 (abduct, 4537)   |
| appreciation | 感谢 (thank, 4255)  |

## 5 Discussion

Our dataset is small compared to the enormous amount of Weibo posts that contain misinformation. Nevertheless, our finding is alarming as it implies the widespread of this misinformation on the platform. For instance, while less than 15% of the posts used mention to directly interact with specific users; almost half of the posts were forwarded to others. The fact that over 60% of them have comments suggests that users were not only being informed by these false information but also engaged in higher cognitive processing.

As shown in the previous section, the topics of these posts are mainly those that Chinese people already feel concerned about. Prior studies in persuasion research imply that such topics increase the persuasive power of the posts by two contextual factors. First, it is considered to be an effective persuasive strategy that triggers expected behavior or change people's behavior to arouse their fear [30–33]. Second, when one needs to assess the information in a post, two effects, namely, the contrast effect and the assimilation effect, explain how one's prior position affects the assessment processes [34]. Specifically, when the proposed comment is about something that the person finds acceptable, then the person subconsciously minimizes the differences between the comment's position and his/her own position (assimilation effect), and vice versa (contrast effect). Therefore, as Chinese people are already worried about problems such as food safety and child trafficking in the society and feel susceptible to them, it is relatively easier to arouse their fears and trigger them to perform expected actions when the misinformation is closely related to these topics. Also shown in the results, the verbs about asking receivers to forward the post appeared a lot in these posts, and indeed almost half of the posts were forwarded. This indicates the persuasive power of these posts. However, to confirm the influence of these contextual factors, further investigation is needed such as a follow-up questionnaire study that probes Weibo users' affective experiences when reading such posts. We also note here that a sentiment analysis is likely not an appropriate method to use to check whether a fear emotion is aroused, because the narrative can be neutral but arouse fear to the readers. For example, this post is describing an incident without showing the negative sentiment but can arouse fear to the receivers who have children themselves

(i.e., not feeling safe about the living world) – “帮朋友转急找孩子求帮忙寻人启示。 13940292999 线索酬金10万。今天上午一个三岁小女孩锦绣花园小区人拐走。小孩说出爸爸手机号。监控上看一个四十多岁男人抱走。现夫人都急疯。知情者请告。万分感谢信息。朋友留意。联系人宁继春 13940292999 请转转” (Translation: (I'm) helping friends forward this in urgency. Child went missing 13940292999 a reward of 10, 000 Yuan is provided if a clue is provided. This morning a three-year old girl was abducted by someone in JinXiu Garden area. The child can tell her parent's cell phone number. The surveillance shows that a man in forties took her away. Her family is so desperate now. (If you have a clue) please tell. Appreciate it very much. Watch out for this friend. Contact Ning Jichun 13940292999”. In fact, according to the Linguistic Inquiry & Word Count (LIWC) tool and SnowNLP, the sentiment of this post is classified as positive, because of words like please, appreciation, help. Note that LIWC is capable of providing a broad range of social and psychological insights from the language including the sentiments and some emotions such as anger. It is commonly used in the sentiment analysis of texts [35], and has a Chinese dictionary [36]. SnowNLP is an open source Python-based tool for analyzing the sentiment of a Chinese sentence (<https://github.com/isnowfy/snownlp>).

Another frequent action that the posts asked the receivers to do is to offer the help. The posts also showed gratitude to encourage these actions. Since these are the posts that containing misinformation, such “help wanted” scenarios are expected to be fake. What kind of impact do these scenarios have on the society? From the famous fable “The Boy Who Cried Wolf”, one would expect that once people find out that these scenarios are fake they would be more skeptical and hesitate to offer their help the next time they receive similarly posts even when the requests are real. The trust literature implies that such doubts can have severe negative impact to the society. Trust is “a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another” [37]. Mayer, Davis, and Schoorman [38] explained that trust propensity is an individual's general willingness to trust others. McKnight, Carter, Thatcher, and Clay [39] showed that people have different trust propensity towards computer technologies. It has been acknowledged that one's trust propensity is affected by various factors such as one's personality characteristics, cultural background, and experiences [38]. Therefore, as Weibo users become more and more skeptical when receiving the posts that ask for help, their trust propensity to others in this social media environment and/or to this environment may be undermined. Lewis and Weigert [40] maintained that trust is a social reality and all social relationships ultimately depend on trust and it is “a functional prerequisite for the possibility of society” (p. 968). One can therefore imagine that such misinformation can have severe negative impact on the society as they make people trust each other less. Further investigation is desired to identify the relationships between online misinformation and the trust in the contemporary society.

These observations demonstrate that posts containing misinformation are featured by several linguistic signals (e.g., citizen-concerned topics, emotional appeals, extensive use of forward-requested words). It sheds light on implementing intervention and moderation strategies against the spread or virality of the misinformation by social computing system. Furthermore, integrating emotional aspects with cognitive beliefs can extend

the understanding of individual social media use [39]. we call for research effort that designs intelligent interventions to fraise users' awareness of how they are affected by misinformation based on our observations.

## 6 Conclusion

Social media users are found to be persuaded by misinformation and disinformation instead of facts and evidence [1, 2]. With the pervasiveness of social media and the wide participation of users, this can bring significant and negative impact to the society. This study aims at exploring characteristics of misinformation content in the Chinese context. We collected and analyzed various aspects of the 26,138 Weibo posts that contain misinformation. These posts span from late August 2011 to early May 2019. Our results show that many posts that contain misinformation tactically target topics that Chinese people are already concerned about. The persuasion literature implies that this may trigger fear emotion among the receivers which increases the persuasive power of the posts. Additionally, forward-asking verbs are frequently used in the posts. Therefore, if one is persuaded by the posts, one expected behavior is that the individual would share the posts with the others. This can contribute to the virality of the misinformation. Our analysis also discovered that a large proportion of the posts asked for help, and the posts showed gratitude to acknowledge the forwarding and helping behavior. Based on the trust literature and the notion that trust is a social reality, we consider these posts can impose potentially severe negative impact on the society as they undermine Weibo users' trustfulness to others and to the social media platform.

A series of tasks is desired to further investigate the characteristics of the Weibo posts that contain misinformation. For instance, to confirm the fear arousal experiences with posts that are about child trafficking, food safety, etc., one can conduct a follow-up questionnaire study that probes Weibo users' affective experiences when reading such posts. Also, besides the use of pathos, whether and how other persuasion strategies are used in the online content that contain misinformation? For example, do they tend to offer more numbers which is shown to be effective in increasing the persuasive power of the online content [41]?

We also make an attempt to connect the characteristics of misinformation with its influence on people's trust of the others and of the social media platform. We found that these posts frequently asked the receivers for help. From the trust literature, this can undermine Weibo users' trustfulness to others. From the perspective that trust is a social reality, this impact can be severely negative to the society. We call for more studies to explore deeper the relationship between online misinformation and the trust in the contemporary society.

## References

1. Cohen, M.: Fake news and manipulated data, the new GDPR, and the future of information. *Bus. Inform. Rev.* **34**(2), 81–85 (2017)
2. Guo, L., Rohde, J.A., Wu, H.D.: Who is responsible for Twitter's echo chamber problem? Evidence from 2016 US election networks. *Inf. Commun. Soc.* **23**(2), 234–251 (2020)

3. Carlson, M.: Fake news as an informational moral panic: the symbolic deviancy of social media during the 2016 US presidential election. *Inf. Commun. Soc.* **23**(3), 374–388 (2020)
4. Humprecht, E.: Where ‘fake news’ flourishes: a comparison across four Western democracies. *Inf. Commun. Soc.* **22**(13), 1973–1988 (2019)
5. Chadwick, A., Vaccari, C., O’Loughlin, B.: Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. *New Media Soc.* **20**(11), 4255–4274 (2018)
6. Ciampaglia, G.L., Shiralkar, P., Rocha, L.M., Bollen, J., Menczer, F., Flammini, A.: Computational fact checking from knowledge networks. *PLoS ONE* **10**(6), e0128193 (2015)
7. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsl.* **19**(1), 22–36 (2017)
8. Vlachos, A., Riedel, S.: Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pp. 18–22 (2014)
9. Kow, Y.M., Kou, Y., Zhu, X., Sy, W.H.: Just my intuition: Awareness of versus acting on political news misinformation. In: Taylor, N.G., Christian-Lamb, C., Martin, M.H., Nardi, B. (eds.) *iConference 2019. LNCS*, vol. 11420, pp. 469–480. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-15742-5\\_45](https://doi.org/10.1007/978-3-030-15742-5_45)
10. Marchi, R.: With Facebook, blogs, and fake news, teens reject journalistic “objectivity”. *J. Commun. Inq.* **36**(3), 246–262 (2012)
11. Zubiaga, A., Aker, A., Bontcheva, K., Liaka-ta, M., Procter, R.: Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv. (CSUR)* **51**(2), 32 (2018)
12. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for Fact Extraction and VERification, 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT) (2018). <https://aclweb.org/anthology/N18-1074>
13. Wang, W.Y.: Liar, Liar Pants on Fire: A new benchmark dataset for fake news detection. In: 55th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 422–426 (2017). <https://www.aclweb.org/anthology/P17-2067>
14. Ferreira, W., Vlachos, A.: Emergent: a novel data-set for stance classification. In: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1163–1168 (2016)
15. Yang, F., Liu, Y., Yu, X., Yang, M.: Automatic detection of rumor on Sina Weibo. In: *The ACM SIGKDD Workshop on Mining Data Semantics* (2012)
16. Chen, T., Li, X., Yin, H., Zhang, J.: Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 40–52, June 2018
17. Ma, J., Gao, W., Wei, Z., Lu, Y., Wong, K.F.: Detect rumors using time series of social context information on microblogging websites. In: 24th ACM International on Conference on Information and Knowledge Management, pp. 1751–1754 (2015)
18. Sun, S., Liu, H., He, J., Du, X.: Detecting event rumors on Sina Weibo automatically. In: *Asia-Pacific Web Conference*, pp. 120–131 (2013)
19. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on sina weibo by propagation structures. In: 2015 IEEE 31st International Conference on Data Engineering, pp. 651–662 (2015)
20. Liu, Y., Wu, Y.F.B.: Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018)
21. Zhou, K., Shu, C., Li, B., Lau, J.H.: Early rumour detection. In: *Proceedings of NAACL-HLT 2019*, pp. 1614–1623 (2019)

22. Yuan, Q., Gao, Q.: The analysis of online news information credibility assessment on Weibo based on analyzing content. In: International Conference on Engineering Psychology and Cognitive Ergonomics, pp. 125–135 (2016)
23. Dang, Z.R.: An analysis of China's rumor spreading process via Sina Weibo, Master's Thesis, Bangkok University (2016). <http://dspace.bu.ac.th/jspui/handle/123456789/2027>
24. Liu, R., Grunert, K.G.: Satisfaction with food-related life and beliefs about food health, safety, freshness and taste among the elderly in China: A segmentation analysis. *Food Qual. Prefer.*, p. 103775 (2019)
25. Wu, Q., Liang, X.: Food therapy and medical diet therapy of traditional Chinese medicine. *Clin. Nutr. Experimental* **18**, 1–5 (2018)
26. BCG (Boston Consulting Group): Capturing a share of China's consumer health market from insight to action (2014). <https://www.bcg.com/publications/2014/center-consumer-customer-insight-globalization-insight-action-capturing-share-chinas-consumer-health-market.aspx>. Accessed 5 May 2018
27. Wu, X., Yang, D.L., Chen, L.: The politics of quality-of-life issues: food safety and political trust in China. *J. Contemp. China* **26**(106), 601–615 (2017)
28. Shen, A., Antonopoulos, G.A., Papanicolaou, G.: China's stolen children: internal child trafficking in the People's Republic of China. *Trends Organized Crime* **16**(1), 31–48 (2013)
29. Xin, Y., Cai, T.: Child trafficking in China: evidence from sentencing documents. *Int. J. Popul. Stud.* **4**(2), e817–e817 (2018)
30. Rogers, R.W.: Cognitive and physiological processes in fear appeals and attitude change: A revised theory of protection motivation. In: Cacioppo, J.T., Petty, R.E. (eds.) *Social Psychophysiology: A Sourcebook*, pp. 153–176. Guilford Press, New York (1983)
31. Rogers, R.W., Deckner, C.W.: Effects of fear appeals and physiological arousal upon emotion, attitudes, and cigarette smoking. *J. Pers. Soc. Psychol.* **32**, 222–230 (1975)
32. Ruiters, R.A., Abraham, C., Kok, G.: Scary warnings and rational precautions: A review of the psychology of fear appeals. *Psychol. Health* **16**(6), 613–630 (2001)
33. Tannenbaum, M.B., et al.: Appealing to fear: a meta-analysis of fear appeal effectiveness and theories. *Psychol. Bull.* **141**(6), 1178–1204 (2015)
34. O'Keefe, D.J.: *Persuasion: Theory and Practice*. Sage, Newbury Park (1990)
35. Reagan, A.J., Danforth, C.M., Tivnan, B., Williams, J.R., Dodds, P.S.: Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. *EPJ Data Sci.* **6**(1), 28 (2017)
36. Gao, R., Hao, B., Li, H., Gao, Y., Zhu, T.: Developing simplified Chinese psychological linguistic analysis dictionary for microblog. In: International Conference on Brain and Health Informatics, pp. 359–368 (2013)
37. Rousseau, D.M., Sitkin, S.B., Burt, R.S., Camerer, C.: Not so different after all: A cross-discipline view of trust. *Acad. Manage. Rev.* **23**(3), 393–404 (1998)
38. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. *Acad. Manage. Rev.* **20**(3), 709–734 (1995)
39. Mcknight, D.H., Carter, M., Thatcher, J.B., Clay, P.F.: Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Manage. Inf. Syst. (TMIS)* **2**(2), 12 (2011)
40. Lewis, J.D., Weigert, A.: Trust as a social reality. *Soc. Forces* **63**(4), 967–985 (1985)
41. Xiao, L., Khazaei, T.: Changing others' beliefs online: online comments' persuasiveness. In: Proceedings of the 10th International Conference on Social Media and Society, pp. 92–101 (2019)



# Ethical, Legal and Security Implications of Digital Legacies on Social Media

Paige Zaleppa<sup>(✉)</sup> and Alfreda Dudley<sup>(✉)</sup>

Towson University, Towson, MD, USA

pzalepp1@students.towson.edu, adudley@towson.edu

**Abstract.** The internet has transformed the way that humans interact not only online, but also in real life. Through social networking sites and other community based online sites a person can amass multiple online identities through their active presence on different online platforms that can affect the user in the real world [2]. These identities can be home to different types of digital artifacts in the form of pictures, posts, or videos and are attached to a specific user. The combination of these identities can be used to create a digital narrative about the user which can then be used to create a timeline of the user's life and their relationships. The content of this digital narrative is the property of the user that created it and can be considered their digital legacy. It is up to the user to determine how their digital legacy will be handled after their real-world death.

**Keywords:** Digital legacy · Social media · Ethical · Legal and social implications · Death

## 1 Introduction

Historically, legacies have included physical items handed down from one individual to another either through understanding or legal documentation. The information age has brought about a shift in physical items such as pictures, letters, and documents being replaced by digital equivalents [1]. These digital files can be shared or stored by individuals as they please in the same way those physical items were. However, this fundamental shift to digital representations of physical items has not been paired with a fundamental shift in how these digital representations are passed down to the next generation.

In real world legacies, the intertwining of these various values applied to a single asset can cause a number of ethical, legal, and security concerns. The responsibility of providing access to these digital assets after a person dies depends on a number of factors. One factor is that the deceased individual took it upon themselves to properly inventory their digital assets and provide a way for their heirs to access that data. Another factor is that the deceased individual took it upon themselves to designate a digital executor that they trust to carry out their wishes for their digital assets. A third factor is that the services that the deceased individual was using during their lifetime have methods in place to provide a way to shut down an account, retrieve data, or transfer accounts depending on

their terms of service. Based on the terms of service the deceased individual should also take into consideration that some aspects of the digital legacy may not be transferable to their heirs.

### 1.1 What Is a Digital Legacy?

A Digital Legacy can be defined as the collection of information available digitally about a user after they have died [2, 6]. A digital legacy is made up information that the user has personally posted, co-authored, shared, or has had posted or shared about them. Digital Legacies consist of digital assets that contain data about a user from every facet of their digital and real-world lives. This information can be found in many different places such as Social Networking Sites, emails, websites, online forums, cloud providers, digital wallets, and other similar online sites. This collection of information is a reflection of the user's profile. And as users create more digital content, this information becomes an even richer reflection of the user in the real-world context [3]. Table 1 outlines some of the common categories that data falls into as well as the relevant data that would need to be considered in those categories during the estate planning process [6].

**Table 1.** Based on the contents of a digital legacy [6].

| Accounts              | Types of data              | Examples of data         |
|-----------------------|----------------------------|--------------------------|
| Online                | Credit cards, etc.         | PayPal, Bitcoin, Netflix |
| Social media profiles | Login, news, media         | Instagram, Facebook      |
| E-mail                | Addresses, online profiles | Google, etc.             |
| Software services     | Login, news, media         | WordPress, NetSuite      |
| Licenses              | Login, news, media         | Software, video games    |
| Hardware              | Documents, projects        | Tablets, USBs, Nooks     |

The increase in use of internet-based technologies for many different facets of life has increased the amount of information available about a user online. Social Networking Site data can be used to reveal information about a users' connections or communities online and also their social, political, moral, or ethical views of the world. Online shopping data can be used to reveal information about a users' spending habits, shopping preferences, or financial standing. Email and cloud storage accounts can contain photos, videos, documents, correspondence, and many other forms of information. Various online service accounts can contain information pertaining to a users' business, medical, financial, and legal information [1, 4]. These accounts as well as other information available can be combined together make up the digital legacy of a user. It is up to the user to determine what parts of their legacy they would like passed on to the next generation and who will be responsible for their information once they have died.

## 1.2 Planning for a Digital Death

Death is not something that people like to think or speak about, but it is a fact of life. It is important to plan for in order to ensure that ones' wishes are carried out after they are no longer able to carry them out themselves. Carroll and Romano use the term digital executor to describe the person or web service that will act objectively on ones' behalf after they have passed on [3]. This individual should be given an inventory of all digital assets, access to those digital assets, and the owner's wishes for that digital asset.

## 1.3 Creating a Digital Assets Inventory

It is important for individual with digital assets to create an inventory of their assets. This inventory should include any credentials needed to access the account that the digital asset is stored in [3]. This can be done in a number of ways including spreadsheets, written instructions, or using password managers that allow for digital heirs to be defined [12]. This will allow for the digital executor or designated individual to access the digital assets and carry out the deceased individuals wishes for that asset. In the publication Digital Afterlife [3], there are listed suggestions for content when constructing a digital asset inventory:

- Asset                Name and other contents
- Access             Location, Username, Password
- Wishes             Instructions, Recipient

After an inventory has been created the next step is for the owner of the inventory to decide where, who, when, and how the inventory is given to the digital executor [3]. The individual must also decide where their inventory will reside until their death. Although the main goal of creating an inventory is to provide access after one has passed, putting all of this information into one place can become a security issue if not stored in a secure location during the person's life. Deciding on when this inventory will be released to the digital executor must also be taken into account when deciding on a storage location in order to ensure that it will reach them at the right time and in the manner that the owner of assets wanted.

## 1.4 Choosing a Digital Executor

Any individual that has digital assets should take the steps to designate a digital executor that they trust to carry out the wishes they have for their digital assets once they have died. A digital executor is not currently a position recognized by law, but they can be made as co-executor which would give them the same powers as a normal estate executor. Depending on the size of the digital and physical legacy and the technology skills of the estates executor these two roles can be given to one person [2, 6]. A digital executor has a number of responsibilities related to passing on digital assets to the designated heirs. Whether this is a person or a digital service there are number of tasks that the digital executor must be able to do in order to ensure that the deceased individuals' digital legacy is protected and handled according to their wishes.



### **Responsibilities of the Digital Executor**

Digital Executors have the same responsibilities as ‘regular’ executors of non-digital estates. Digital Executors must understand the basics of the federal and state laws that pertain to third-party online services [15].

- *Understand the terms of service associated with various online accounts.*
- *Understand the basics of the various technologies that are being used to store the assets listed in the inventory.*
- *Freeze and close any accounts that are designated by the deceased.*
- *Be able to securely and safely archive any data within accounts set to be closed or frozen.*
- *Be able to securely and safely distribute digital assets to heirs according to the wishes described in the inventory.*
- *Be able to securely and safely contact those designated as heirs to the digital assets [3, 6].*

### **1.5 The Value of Digital Legacies**

The content that makes up Digital Legacies are called digital assets. Digital assets are data or information, which is posted and/or shared by the user themselves or information that other users have posted and/or shared about them. Classifying these digital assets into collections and those collections into categories is important in order for a user to properly plan for their death. This would include such information as a user’s name and password, the stored contents, the file location, instructions for the digital executor, and the recipient of the legacy [6, 12].

These digital assets have value that is affected by what the digital asset is and how it can be used by the descendants of the deceased. The type and amount of value placed on these digital legacies can have real world ethical, legal, and security consequences. For the purposes of this paper, the type of value that can be placed on a digital asset is monetary, sentimental, cultural, or a combination of the three. Depending on the asset and the individual that is placing the value on the asset the type and the amount of the value will vary. Users must consider the legal, ethical, and social implications of who they choose to pass their digital assets onto.

### **The Different Value Types of Digital Assets**

A legacy can be considered an aggregation of all the assets that an individual accumulates over their lifetime. Legacies in the real world have value either monetarily, sentimentally, culturally or a combination of these three. Digital legacies also share this characteristic but, because of the inherent differences between physical legacies and digital legacies, the value of the assets that make up digital legacies is not always easy to discern. The value of digital assets is being affected by the increase in the shared experiences that are becoming increasingly prevalent through the use of social media and other online services [2]. The type and amount of value placed on these digital legacies can have real world ethical, legal, and security consequences. Similar to real world legacies, digital legacies are made up of assets which have their own value either in terms of monetary, sentimental, or cultural value.

- Monetary Value – the property of having material worth (Monetary Value).
- Sentimental Value – the value of something to someone because of personal or emotional associations (Sentimental Value).
- Cultural Value – the value of something based on a social groups' evaluation of the objects worth.

Digital values are measured in different ways, applied using different methods, and applied in different increments depending on many factors when assessed from a digital prospective [10, 11, 13].

## 1.6 Responsibilities of Account Providers

The responsibilities of the account providers depend on the terms that they have laid out in their terms of service that must be agreed to in order for an account to be opened. These terms of service vary from service to service and impact the level of responsibility that the account provider has when dealing with a deceased users account. Some services like iCloud, make accounts non-transferable which essentially remove any involvement that they could have in ensuring the contents of their accounts are accessible after the account holders' death [8]. Other services like Facebook, allow for users to designate a legacy contact that allows for an account holder to designate the person who can manage a users' account if it is memorialized [14]. However, if a user does not designate a legacy contact the account is lost to the individual who was supposed to carry out the wishes of the deceased. Twitter allows for the designated estate executor to contact Twitter in order to have the account deactivated but cannot gain access to the account [7]. When one is planning for their digital legacy it is important for them to consider the terms of service that they have agreed.

## 2 Ethical, Legal, and Security Considerations

Digital assets value add to the overall monetary, sentimental, or cultural value of the digital legacy. In real world legacies, the intertwining of the various values applied to a single asset can cause a number of ethical, legal, and security concerns. As the number of assets increases these concerns are added together which increases the complexity of the task of dealing with a legacy both before the individual has died and after they have passed on. A digital legacy shares these characteristics, but because certain assets that make up a digital legacy are slightly different from their real-world counterparts or do not exist in the real world, special consideration must be taken from an ethical, legal, and security perspective.

In order to demonstrate the content dynamics of digital legacies in relation to content created about a user after they have died by other user's data was collected from Twitter. Tweets containing the words ethical, legal, condolences, R.I.P., memorials, bereavement, and death posted between January 1, 2014 and April 17, 2019 were collected using the Twitter API.

When the resulting tweets that contained the word ethical or legal were combined with the resulting tweets that contained the words R.I.P., condolences, memorial, death,

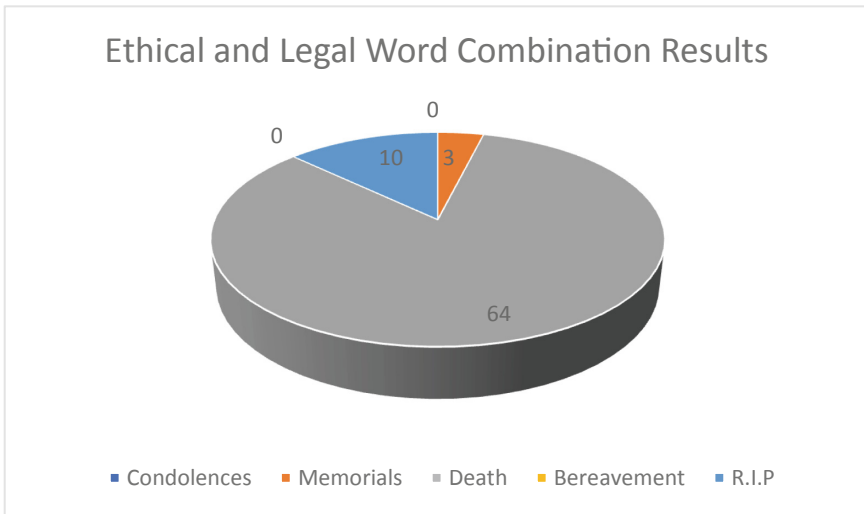
legal, or bereavement the number of tweets that contained both varied depending on the word. The word combinations of ethical and bereavement; legal and condolences; and legal and bereavement resulted in no tweets that contained both words. When the words legal and ethical were combined with the words condolences or bereavement there were also no tweets that contained all three words. The combination of words that resulted in the highest number of resulting tweets was the combination of ethical and legal with 10,141 total tweets. The next highest number of tweets was from the combination of ethical and death which returned 1,702 tweets that contained both words. Interestingly, the combination of ethical, legal, and death resulted in the highest number of tweets returned for any three-word combination. The next highest number of tweets returned for any three-word combination was ethical, legal, and R.I.P. which resulted in 10 tweets returned that contained all three words (Tables 2 and 3, Fig. 1).

**Table 2.** Single word combinations

| Word 1  | Word 2      | Total tweets |
|---------|-------------|--------------|
| Ethical | R.I.P.      | 452          |
| Ethical | Condolences | 21           |
| Ethical | Memorial    | 76           |
| Ethical | Death       | 1702         |
| Ethical | Legal       | 10,141       |
| Ethical | Bereavement | 0            |
| Legal   | R.I.P.      | 189          |
| Legal   | Condolences | 0            |
| Legal   | Memorial    | 6            |
| Legal   | Death       | 111          |
| Legal   | Bereavement | 0            |
| Legal   | Ethical     | 10,141       |

**Table 3.** Two word combinations

| Word combination  | Word        | Total tweets |
|-------------------|-------------|--------------|
| Ethical and legal | Condolences | 0            |
| Ethical and legal | Memorials   | 3            |
| Ethical and legal | Death       | 64           |
| Ethical and legal | Bereavement | 0            |
| Ethical and legal | R.I.P.      | 10           |



**Fig. 1.** Ethical and legal word combination results

## 2.1 Ethical Analyses

Ethically a number of issues arise when dealing with the value of a digital asset and digital legacy. Some of these dilemmas can be solved through the implementation of various methods typically used for estate planning. However, as the digital assets that make up digital legacies continue to diverge from their physical real-life counterparts the issue becomes what can be passed on to heirs. For example, iCloud accounts are nontransferable and if the password is not left behind for an heir to save those digital assets on a local device all of the digital assets will be lost because no one can be granted access to the contents of the account [8]. This raises the question of if companies should be allowed to make all of the contents of an individual's account nontransferable?

From a Utilitarian perspective, more individuals would benefit from companies being required to make accounts transferable. Heirs, lawyers, family members of the deceased, and friends of the deceased would benefit from making accounts transferable. This would allow for the contents to be backed up locally or on other services in order to preserve the digital assets. However, companies as well as the creators of certain content would not benefit from making the content transferable. iCloud cites that because the music, movies, and other digital files purchased through their platform are licenses to the content and not physical items the license agreement ends at the death of the owner [8]. By making the content of accounts transferable companies could be negatively impacted financially. Companies and creators would benefit from keeping the contents nontransferable because they would be able to continue to gain revenue through the sale of licenses to individuals. They would also benefit because they would not need to develop a process and/or software to support the transferring of accounts from one individual to another. Heirs, lawyers, family members of the deceased, and friends of the deceased would not benefit because the contents might have sentimental value to the individual or to the individual that they are representing in the case of a lawyer.

From a Deontological perspective, companies should not have to make the contents of the accounts transferable. This is because by requiring the companies to make the contents of the accounts transferable the company would be neglecting their duty to the creators of the material and would therefore be violating their right to the material that they created. It could be argued that the company would neglecting their duty to the deceased individual and their content. However, even if the content was created by them such as pictures, videos, or files they are creating the material and it should be protected in the same fashion as the creators that sell access to their material.

Looking at this issue from the perspective of Kant's Categorical Imperative Principle, companies should not be required to make the contents of users' accounts transferable. By making the contents of accounts transferable the companies would be treating the creators of the materials within those accounts with disrespect. This is because the content that they created would be shared with others that were not specifically authorized by the creator of the material to access them. The companies would be treating the material creators differently because they would be allowing individuals that do not have a license to the materials, they created to access it.

Analyzing this question from the Virtue Ethics theory also results in companies not being required to make the contents of users' accounts transferable. Companies would be acting in virtue because they would be protecting the rights that creators have over their materials. This is because the any proprietary licenses, data, or files would be protected from individuals that have not been given explicit authorization to access it. This would protect the privacy of the deceased as well because the content of their online accounts that were not explicitly passed on to an heir would remain secure. This would prevent unauthorized access and possible security issues.

Analyzing this from an ethical perspective, companies providing methods to access a deceased loved one's accounts can be a great benefit but can also do great harm. Although most people would want to have access to their loved one's accounts because they contain digital assets that are valuable to them for either sentimental, monetary, or cultural reasons it would violate the privacy of the deceased user. This could also violate the rights of the content creators that may or may not be the deceased user, because without explicit permission either through legal channels or giving their login credentials the heir does not have any right to access the digital assets contained within the account. Also from a business perspective, most services make money from the user actually using their site or application and when the user is no longer able to do that there is no reason to allocate resources to the management of that account [12].

## 2.2 Legal Perspectives

Legally the real world and digital assets of an individual belong to that individual until they die at which time, they have the opportunity to designate an heir to that asset. Both real world and digital assets have either monetary value, sentimental value, cultural value, or a combination of the three. The amount of these values can be assessed at different amounts by different people each ranking the values on which should be considered the most important when assessing the value of the asset. The application of the different values at different amounts presents a power struggle that present a number of legal hurdles if not planned for accordingly. The digital legacy of an individual is their

responsibility to legally plan for and if they do not the fate of their digital legacy lies in the hands of the services where their digital assets reside.

Through their terms and conditions service providers have been able to clearly state that they have no legal obligation to provide access to the data stored on their services when a user dies. This is because most services make money from the user actually using their site and when the user is no longer able to do that there is no reason to allocate resources to the management of that account [12]. Some services also make the account and its corresponding assets nontransferable which means that the account cannot be given to someone else. The ownership of the digital assets might come into question as well because these terms and conditions because the right to ownership might be shared or some or all of the ownership might be taken away [2, 3]. The value of these assets is important to consider when planning a digital legacy because the value could be impacted by the terms and services that binds the asset.

### 2.3 Security Considerations

Security is an ever-present concern when it comes to all categories of data stored digitally and online. While a user is alive their own, and the individuals around them, physical safety could be impacted by the data collected on their location. Their privacy can be impacted by the personal, medical, and financial data that is available through their various digital assets. After the user dies these are still data security concerns and proper steps must be taken by the user in order to ensure that their data is secure after they die. As mentioned before, it is good practice for a user to ensure that access is provided to the various services they use in the event of their death. However, the user must also consider who will have control of their digital assets. The choice of who has access falls to the user planning for their digital legacy to decide. Making another individual an authorized user to any service is an important decision because the digital assets available on those sites might impact other users that might still be alive [5].

Depending on the party providing the access to the deceased users accounts the security issues remain. If a service were to provide access to the heir or digital executor this could put them at risk of sharing personal information that the deceased user never wished to be made public. If the deceased user provides access through a digital inventory or other method to share their credentials to various services, they are putting their own information at risk of being used for purposes that they did not authorize. If a digital executor provides access to the deceased individuals' digital assets to an heir, they have no control of what the heir will do with that information. The security of the deceased user's information needs to be considered by the individual before they die and the services within their terms of service as well and the digital executor and heirs of the deceased individual [5, 12].

### 2.4 Privacy Concerns

The privacy of a deceased user also brings about a number of legal issues that surround the level and type of access that an heir can gain to a deceased loved one's account. From a legal perspective, there are three parties that are impacted by the content of digital legacies. One party is the deceased individual, they were the ones that signed

up for the accounts and agreed to the terms of service set forth. The services that the deceased individual signed up are another party to be considered from a legal perspective because they ultimately set forth the rules for who can do what to a deceased individuals account. For example, Twitter allows for the estate executor to contact them to get the account closed but can never gain access to data within the account [7]. The third party or in some cases parties are the digital executor and heirs of the deceased user. If they are left with the account credentials to the various accounts, they can essentially act as the deceased user. This could invade the deceased user's privacy which brings back the ethical question, should heirs be given access to a deceased loved one's online accounts?

### 3 Conclusion and Future Work

The information age has brought about a shift in physical items such as pictures, letters, and documents being replaced by digital equivalents [2]. These digital files can be shared or stored in the same way as physical items based on the context. The fundamental shift to digital representations of physical items has not been paired with a fundamental shift in how these digital representations are passed down to the next generation.

A digital legacy can be compared to a legacy in real life and should be handled in a similar fashion. Because of the inherent difference between digital assets and physical assets there are different steps that need to be taken in order to properly protect a digital legacy for the next generation. Ethical and legal ramifications that are associated with the creation of various kinds of digital assets need to be understood better by all individuals that are creating digital content.

The authors assessed from this study that more resources should be made available for people creating and leaving content on various online and social media accounts. These resources would include information and guidance in understanding how various terms of service agreements impact how and to who users can leave their accounts to after death. In addition, further research will need to be conducted from several social media sites in order to reach a better understanding of the content dynamics of digital legacies and how these legacies are impacted from both legal and ethical perspectives.

### References

1. Bell, G.: *Building Social Web Applications: Establishing Community at the Heart of Your Site*, 1st edn. O'Reilly Media, Sebastopol (2011). ISBN 0596518757
2. Braman, J. Dudley, A. Vincenti, G.: *Death, social networks and virtual world: a look into the digital afterlife*. In: *Proceedings of the 9th International Conference on Software Engineering Research, Management and Applications*, Baltimore, MD, USA (2011)
3. Carroll, E. Romano J.: *Your Digital Afterlife: When Facebook, Flickr and Twitter are Your Estate, What's Your Legacy?* New Riders Press, Berkeley (2010)
4. Codjia, M.: A definition of financial information. <https://bizfluent.com/info-8333182-definition-financial-information.html>. Accessed 11 June 2019
5. Cross, M.: *Social Media Security: Leveraging Social Networking While Mitigating Risk*, 1st edn. Syngress Press, Rockland (2013). ISBN 1597499862
6. Digital legacy. <https://www.ionos.com/digitalguide/websites/digital-law/digital-legacy/>. Accessed 1 June 2019






7. How to contact Twitter about a deceased family member's account (n.d.). <https://help.twitter.com/en/rules-and-policies/contact-twitter-about-a-deceased-family-members-account>
8. Legal - iCloud - Apple (n.d.). <https://www.apple.com/legal/internet-services/icloud/en/terms.html>. Accessed 18 June 2019
9. Location data (n.d.). <https://ico.org.uk/for-organisations/guide-to-pecr/communications-net-works-and-services/location-data/>
10. Monetary value (n.d.). <https://www.thefreedictionary.com/monetaryvalue>
11. New York State Office of Information Technology Services: Digital Object (n.d.). <https://its.ny.gov/glossary>. Accessed 11 June 2019
12. Schalit, E.: Why we must care about our digital legacy (2015). <https://blog.dashlane.com/why-we-must-care-about-our-digital-legacy/>
13. Sentimental value (n.d.). <https://www.merriam-webster.com/dictionary/sentimentalvalue>. Accessed 11 June 2019
14. Terms of service (n.d.). <https://www.facebook.com/terms.php>
15. What is the role of a digital executor? <https://www.mylennium.com/content/what-role-digital-executor>. Accessed 18 June 2019



# **User Behavior and Social Network Analysis**



# When Emotions Grow: Cross-Cultural Differences in the Role of Emotions in the Dynamics of Conflictual Discussions on Social Media

Svetlana S. Bodrunova<sup>(✉)</sup> , Kamilla Nigmatullina , Ivan S. Blekanov ,  
Anna Smoliarova , Nina Zhuravleva , and Yulia Danilova

Saint Petersburg State University, Saint Petersburg 199004, Russia  
s.bodrunova@spbu.ru

**Abstract.** *Background.* The spread of affective content on social media, as well as user grouping based on affect [1], has been a focus of scholarly attention for over a decade. But, despite this, we lack evidence on what roles various particular emotions play in the dynamics of discussions on social media. Emotional contagion theory (Hatfield et al. 2014) adapted for social media suggests that diffusion of emotions happens on individual level, via direct one-time contact with emotionalized content [2]. Other theories, like theories of social influence or social learning [3], thought, suggest multiple, hierarchical, and/or topically-restricted contacts. The idea of affective agenda [4] implies that the dynamics of an emotional discussion needs to be assessed on the aggregate level. The question remains – what role the emotions taken on aggregate level play in the discussion dynamics, being either catalyzers or inhibitors of the discussions. One may suggest that emotions of different stance (positive/negative) may spur/slow down the discussions in various ways. *Objectives.* We analyze the spread of two polar emotions – anger and compassion – in three Twitter discussions on inter-ethnic conflicts, namely Ferguson protests (the USA, 2014), Charlie Hebdo massacre (France, 2015), and mass harassment in Cologne (Germany, 2015–2016). By analyzing the co-dynamics of the overall discussions and these two emotions we can conclude whether the pattern of the spread of emotions and its link with the discussion dynamics is the same in various language segments of Twitter. *Data collection and methods.* The data we use were collected by our patented Twitter crawler in the aftermath of the conflicts and include altogether over 2,5 M tweets. We used manual coding by native speakers and machine learning to detect the emotions; then, we visualized the dynamics of growth of the emotional content of the discussions and used Granger test to see whether anger or compassion gave a spur to the discussions. *Results.* We have received moderate results in terms of the dependence of the number of neutral users upon that of emotional users, but have spotted that the beginnings of the discussions, as well as the discussion outbursts, depend more on compassion, not on angry users, which needs more exploration. We have also shown that the hourly dynamics of emotions replicates that of the larger discussion, and the numbers of angry and compassionate users per hour highly correlate in all the cases.

**Keywords:** Twitter · Inter-ethnic discussions · Emotion detection · Discussion structure · Machine learning · Emotional agendas

## 1 Introduction

The spread of affective content on social media, as well as user grouping based on affect [1], has been a focus of scholarly attention due to multiple reasons. Emotional content is seen mostly as a threat to the democratic quality of public discussions [5], as it undermines its integrity and rationality [6], as well as its civility and constructive character [7], especially when one deals with hate speech and toxic expressions [8, 9].

Despite the evident necessity to learn how emotions work within the discussion, we lack evidence on what roles various particular emotions play in the dynamics of discussions on social media. Detection of particular emotions has been mostly linked to political emotionality and political functions of emotions [10–13], as well as their linkage to other features of content like fake news [14].

The substantially negative role of hate speech is practically non-questioned today; but several works still draw attention to controversial relations between freedom of speech and hate speech [15, 16] and to the possible use of offensive language in positive sense in discriminated communities like LGBTQ [17]. Another potentially positive role of emotional content (negative just as well as positive) might lie in spurring the discussions, providing them flame and substance and not letting them ebb.

In this paper, we explore just two aspects of this issue, namely – the growth of emotional load in conflictual discussions and the relations between the number of overall users in the discussion and the angry and compassionate ones.

For this, we use the data from three highly conflictual and emotionally polar discussions, more or less territorially localized, despite potential global participation on Twitter. Use of three discussions from different communicative cultures will allow us generalize on the nature of the relations between the overall growth of the discussion and its highly emotional content.

To explore the discussion bulk, we collected the data via Twitter crawling, then coded the emotions with the help of native speakers, then taught the machine to recognize the emotions, and then detected the users who used emotional speech and reconstructed the hourly graphs of the discussions, as well as time series. We have also performed Granger test to spot the relations between the number of emotional and neutral users in the discussions.

The remainder of the paper is organized as follows. Section 2 discusses our idea of the emotions as discussion drivers. Section 3 poses the hypotheses, and Sect. 4 describes the data collection procedures, sampling, and the method of detecting emotions. Section 5 provides and discusses the results.

## 2 Emotions Aggregated: What Can They Tell?

Emotional contagion theory [18] adapted for social media suggests that diffusion of emotions happens on individual level, via direct one-time contact with emotionalized

content [2]. Individual emotions, though, play a key role in group behavior online, as they form the so-called *ad hoc* discussions [19] of affective publics [1] and, thus, matter more on the aggregate level and might work differently from what is expected, especially in fast-growing conflictual discourse online. Other theories, like theories of social influence or social learning [3], suggest multiple, hierarchical, and/or topically-restricted contacts of users with emotional content; whether these limitations are also true for how the emotions grow and spread collectively, still remains unclear.

The idea of affective agenda [4] implies that the dynamics of an emotional discussion might be dependent on mood changes or spillovers from group to group, language to language, or topic to topic. Also, the relations between various emotions might be a competitive one, since they work as the polarizing agents within the discussion: anger may overcome fear, and denial might beat compassion. A shockingly strong example of this is, of course, the case of *Charlie Hebdo* which polarized the discussion participants to the extent of creating formally opposing hashtags ('jesuischarlie' and 'jenesuispascharlie', representing compassion and its rejection, anger, and discontent).

The change of affective agenda or mood within the discussion might lead to its growth (or even outburst) or ebbing, and this question remains under-explored. We suggest that emotions of different stance (positive/negative) may spur/slow down the discussions in various ways.

### 3 The Research Hypotheses

In this paper, we have posed three hypotheses:

**H1.** In all the three conflictual discussions under our scrutiny, the dynamics of emotions will go ahead of the general discussion dynamics: the number of emotional users will grow and diminish faster than the number of emotional tweets.

**H2.** In all the three discussions, the dynamics of the number of involved users will depend on the dynamics of the number of emotional users.

**H3.** The dynamics of compassion and anger will differ in all the three discussions and will not correlate within the cases.

We will also qualitatively describe the picture we see on the web graphs and on time series graphs, to contextualize and reflect more on the received data. By analyzing the co-dynamics of the overall discussions and these two emotions we can conclude whether the pattern of the spread of emotions and its link with the discussion dynamics is the same in various language segments of Twitter.

## 4 Data Collection and Methods of Research

### 4.1 The Cases Under Scrutiny

We analyze the spread of two polar emotions – anger and compassion – in three Twitter discussions on inter-ethnic conflicts, namely:

- The Ferguson unrest, USA, 2014: number of users who published tweets – 70018, total number of messages – 193812;
- The *Charlie Hebdo* shooting, France, 2015: number of users who published tweets – 238491, total number of messages – 505069;
- The mass harassment of women by the re-settlers from the Arab world in Cologne, Germany, in the New Year night of 2015–2016; number of users who published tweets – 99981, total number of messages – 215253.

The overall number of tweets collected based on the relevant hashtags and keywords is actually much bigger. The data we use were collected by our Twitter crawler in the aftermath of the conflicts and include altogether over 2, 5 M tweets. The samples described above resulted from the fact that we selected the mono-language parts of the discussions only – German for Cologne, French for *Charlie Hebdo*, and English for Ferguson. We have also cleaned spam, non-relevant tweets, and hashtag-only tweets.

## 4.2 Emotion Detection

To be able to detect emotions, we have used an enhanced SVM-based approach with the use of Numpy and scikit-learn libraries. As we have been working with multi-lingual sentiment analysis for several years, we have modified the approach with two features. First, the experiments have shown that elimination of stop words from tweets actually diminishes the quality of the classifier. Thus, the stop words have been returned to the initial dataset, which has raised the quality by 4–5%. Second, we have experimentally found the streaming hyper-parameters and adjusted them: the regularization parameter was 0.9, and we have also used the linear nuclear function and the OVR function. We have also spotted that, in several cases, the machine tended to over-learn; thus, we have balanced the samples to reach better results.

Tweets from real-world discussions, despite we have used the well-preprocessed datasets, remain one of the ‘noisiest’ types of textual data; this has cast some negative impact upon the quality metrics of emotion detection. This is why we have used quite large collections for hand coding. We have selected the tweets for hand coding by the procedure described in our earlier works [20, 21], by selecting influential users by nine parameters – four absolute ones (the number of posts, retweets, likes, and comments) and five graph-dependent ones (in-degree, out-degree, degree, betweenness and pagerank centralities). The top users’ tweets constitute the samples for coding. Altogether, we have coded: for Germany – 13,620 tweets, for the USA – 7200 tweets, and for France – 7433 tweets.

We have received the following best levels for the quality metrics (see Table 1).

**Table 1.** Quality metrics for emotion detection for the three cases

|            | Germany  |           | The USA  |           | France   |           |
|------------|----------|-----------|----------|-----------|----------|-----------|
|            | Accuracy | F-measure | Accuracy | F-measure | Accuracy | F-measure |
| Anger      | 0.7      | 0.7       | 0.77     | 0.77      | 0.78     | 0.77      |
| Compassion | 0.75     | 0.74      | 0.8      | 0.8       | 0.84     | 0.84      |

### 4.3 Data Analysis and Visualization

We have marked the tweet datasets by emotion; then we have aggregated the tweets per user and have calculated hourly numbers of neutral, angry, and compassionate users, according on whether they have expressed these emotions within a given hour. Thus, we have received the hourly numbers of users who were emotional, either compassionate or angry. Of course, in a small number of cases, a user could express both compassion and anger within one hour; but this mattered only within the first couple of hours, while in the course of the discussion the number of such users became negligibly small.

To visualize the results of the emotion detection, we have constructed the hourly web graphs for three days using the YifanHu algorithm from the Gephi library, as well as the time series graphs with the hourly lag. For Cologne and *Charie Hebdo*, we have taken the first three days of after the trigger events; for Ferguson, we did not have such data and, thus, the three days were taken from the middle of the discussion.

To assess the inter-emotional relations (in terms of the number of users), we have used the Spearman's  $\rho$ .

To test the hypotheses, we have performed the Granger test with the hourly data and have also qualitatively assessed the graphs [22]. The results are presented below.

## 5 Results and Discussion

**H1 – H2.** When assessing the cases, we have seen that the basic pattern of the discussions is different from the expected: in general, the lines indicating the number of neutral users are generally followed by those of anger and compassion (see Fig. 1a–c). But, if we look at details and the results of the Granger test, the three start to diverge.

First of all, the three cases differ in whether the growth of the number of angry or compassionate users affects the rest of the speech in the case (see Table 2). In some sense, Cologne and Ferguson ‘mirror’ each other in terms of when exactly the emotional users spur the number of the neutral users. We can see from the Table 2 that it is compassion, not anger, that plays a bigger role in predicting the number of neutral users – that is, the very growth or ebbing of the discussions. As we see from the time series graphs, the regression shows the dependency when the discussion starts to grow. This happens with the first two days of Cologne where the journalists were reluctant to cover the mass harassment story as they were not sure how to report on the nationality of the harassers. It also happens during an outburst of emotions in the Ferguson case, on Day 3. Thus, the role of emotional content in the early-stage discussions needs to be well researched upon.

In the American case, there is also a smaller trend covered by a more general one. There are several moments in the course of the discussion when anger starts to diminish earlier than the number of the neutral tweets, and the blue line follows.

Another important conclusion is that the emotions do not grow/diminish within the discussions but have intense hourly dynamics. The ‘jumps’ of the number and connectivity of users who express anger and compassion (see Fig. 2a–b, the example of Cologne) need to be further explored within the aforementioned theories of emotional contagion and/or social learning.

**Table 2.** The results of the Granger test

|                        | Days 1-3    | Days 2-3    | Day 3          |
|------------------------|-------------|-------------|----------------|
| Cologne anger          | No No       | No No       | 11,17** 13,3** |
| Cologne compassion     | 13,1*** No  | 8,71** No   | 5* 9,8**       |
| Ferguson anger         | No 4,13*    | No No       | 4,76* No       |
| Ferguson compassion    | 5,53* 8,1** | 4,68* 6,52* | 5,89* No       |
| Charlie Hebdo anger    | No No       | No No       | Нет нет        |
| Chare Hebdo compassion | No No       | No 10,2**   | No 5,65*       |

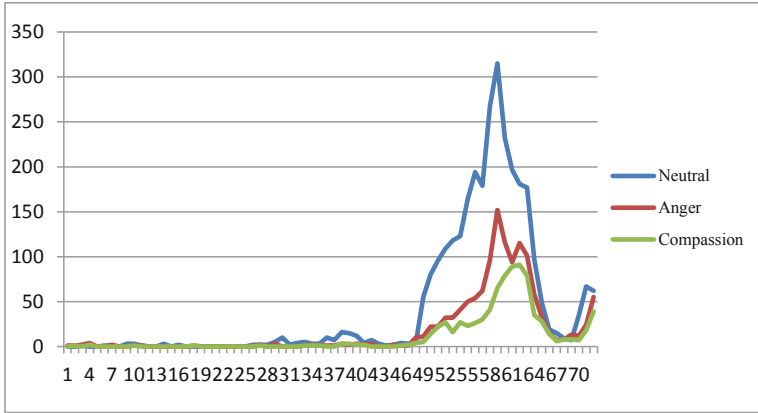
*Note.* Green: the test shows the dependency of the overall discussion upon these users and authors. Yellow: the test shows that the numbers of the neutral users and emotional users correlate due to a third factor. Maroon: The linkage between emotional and neutral users shows the dependency of the other direction: the higher the number of neutral tweets, the higher the number of the angry and/or compassionate tweets. That is, when the discussion is boiling itself, the number of angry users also grows without any coordinated effort.

Thus, **H1 and H2**, in how they were formulated, needs to be rejected: the number of the neutral users does not grow depending of the spread of emotions between emotional users. But what is also important is that the dynamics of the emotions is intensive even within the hourly perspective and that the moments when the discussion grows might depend on compassion, rather than anger. The emotions matter in the very beginning of the discussions or in the moments of discussion outbursts.

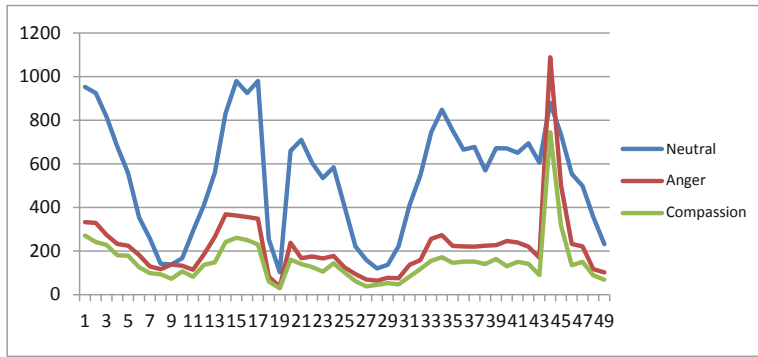
**H3.** Our third hypothesis was created to be rejected, but, again, this is not what happens in the datasets. Thus, the correlations between the two hourly spreads of people are: for Cologne – 0,950\*\*\*\*, for Ferguson – 0,988\*\*\*, for *Charlie Hebdo* – 0,937\*\*. This adds to the evidence that emotional users do not lead the way and do not cause excessive discussion around them but appear within the discussion the same way as neutral users. But, when on an emotional peak, such users might create discussion outbursts.

\* \* \*

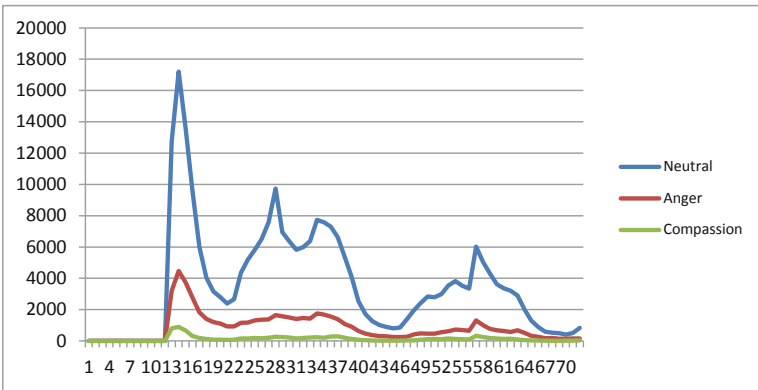
We have tested the dependence of the discussion dynamics on the number of emotional (angry and compassionate) users of Twitter. Within the three cases, there were no similar patterns of such dependence, which might be explained by socio-cultural differences or the structure of the sample. We have also detected a special role of emotional tweets in the beginnings of the discussions and the discussion outbursts; we suggest that, in future, not users but the tweets themselves need to become the unit of analysis.



(a)



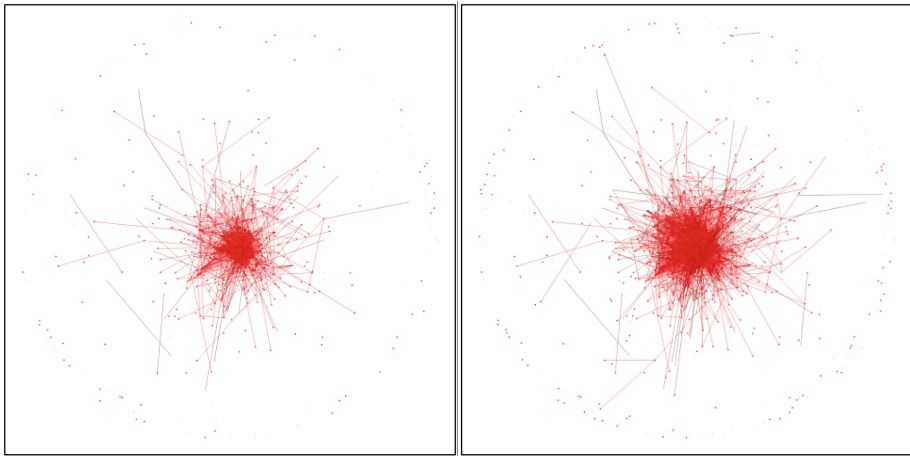
(b)



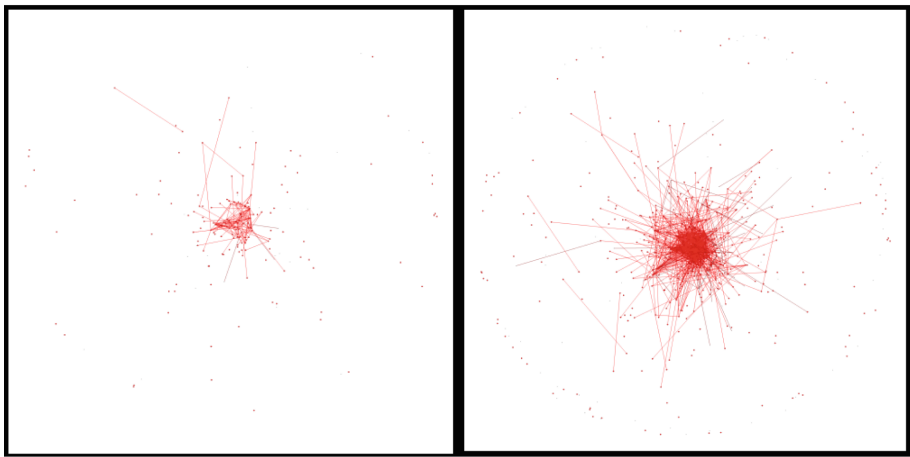
(c)

**Fig. 1.** a. Time series graphs, the Cologne case. b. Time series graphs, the Ferguson case. c. Time series graphs, the Charlie Hebdo case. (Color figure online)





(a)



(b)

**Fig. 2.** a. Fast dynamics of growth of the number of angry users within 3 h (a ‘jump’ of anger), Cologne, January 4, 2016. b. Fast dynamics of growth of the number of compassionate users within 3 h (a ‘jump’ of compassion), Cologne, January 4, 2016

**Acknowledgements.** This research has been supported in full by the Presidential Grant of the Russian Federation to young Doctors, grant MD-6259.2018.6 (2018–2019).

## References

1. Papacharissi, Z.: *Affective publics: Sentiment, Technology, and Politics*. Oxford University Press, Oxford (2015)
2. Coviello, L., et al.: Detecting emotional contagion in massive social networks. *PLoS One* **9**(3), e90315 (2014)

3. Young, R.: *Discursive Practice in Language Learning and Teaching*, vol. 58. Wiley-Blackwell, Malden (2009)
4. Coleman, R., Wu, H.D.: Proposing emotion as a dimension of affective agenda setting: separating affect into two components and comparing their second-level effects. *Journalism Mass Commun. Q.* **87**(2), 315–327 (2010)
5. Cortese, A.J.P.: *Opposing Hate Speech*. Greenwood Publishing Group, Santa Barbara (2006)
6. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: *Proceedings of the 26<sup>th</sup> International Conference on World Wide Web Companion*, pp. 759–760 (2017)
7. Burnap, P., Williams, M.L.: Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. *Policy Internet* **7**(2), 223–242 (2015)
8. Park, J.H., Fung, P.: One-step and two-step classification for abusive language detection on twitter (2017). arXiv preprint arXiv:1706.01206
9. Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G., Plagianakos, V.P.: Convolutional neural networks for toxic comment classification. In *Proceedings of the 10<sup>th</sup> Hellenic Conference on Artificial Intelligence*, pp. 1–6 (2018)
10. Lyman, P.: The domestication of anger: the use and abuse of anger in politics. *Eur. J. Soc. Theory* **7**(2), 133–147 (2004)
11. Ticineto Clough, P., Halley, J.: *The Affective Turn: Theorizing the Social*. Duke University, Durham (2007)
12. Heaney, J.G., Flam, H. (eds.) *Power and Emotion*. Routledge, London (2015)
13. Wahl-Jorgensen, K.: *Emotions, Media and Politics*. Wiley, Hoboken (2019)
14. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. *Science* **359**(6380), 1146–1151 (2018)
15. Dorsett, D.M.: Hate speech debate and free expression. *S. Cal. Interdisc. LJ* **5**, 259 (1996)
16. Cammaerts, B.: Radical pluralism and free speech in online public spaces: the case of North Belgian extreme right discourses. *Int. J. Cult. Stud* **12**(6), 555–575 (2009)
17. Waseem, Z., Davidson, T., Warmusley, D., Weber, I.: Understanding abuse: a typology of abusive language detection subtasks. arXiv preprint arXiv:1705.09899 (2017)
18. Hatfield, E., Bensman, L., Thornton, P.D., Rapson, R.L.: *New perspectives on emotional contagion: a review of classic and recent research on facial mimicry and contagion* (2014)
19. Bruns, A., Burgess, J.E.: The use of Twitter hashtags in the formation of ad hoc publics. In: *Proceedings of the 6<sup>th</sup> European Consortium for Political Research (ECPR) General Conference* (2011)
20. Bodrunova, S.S., Litvinenko, A.A., Blekanov, I.S.: Comparing influencers: activity vs. connectivity measures in defining key actors in twitter *Ad hoc* discussions on migrants in Germany and Russia. In: Ciampaglia, G.L., Mashhadi, A., Yasseri, T. (eds.) *SocInfo 2017*. LNCS, vol. 10539, pp. 360–376. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-67217-5\\_22](https://doi.org/10.1007/978-3-319-67217-5_22)
21. Bodrunova, S.S., Blekanov, I., Smoliarova, A., Litvinenko, A.: Beyond left and right: real-world political polarization in Twitter discussions on inter-ethnic conflicts. *Media Commun.* **7**, 119–132 (2019)
22. Wessa, P.: *Bivariate Granger Causality version 1.0.4 in Free Statistics Software version 1.2.1*, Office for Research Development and Education (2016). [http://www.wessa.net/rwasp\\_grange\\_rcausality.wasp/](http://www.wessa.net/rwasp_grange_rcausality.wasp/)



# The World of Museums and Web 2.0: Links Between Social Media and the Number of Visitors in Museums

Adela Coman<sup>(✉)</sup>, Ana-Maria Grigore, Andreea Ardelean, and Robert Maracine

The University of Bucharest, Bucharest, Romania  
{adela.coman, andreea.ardelean}@faa.unibuc.ro,  
anagrig27@gmail.com, robrtmara@gmail.com

**Abstract.** With the advent of Web 2.0, the life of museums has undergone a profound change: today, almost all museums in the world use social media as a communication strategy with their visitors. However, only a few papers have analyzed the role of social media in attracting a greater number of visitors to museums.

For this reason, the aim of this paper is to analyze how museums use social media to improve their relationship with visitors and, at the same time, to ensure a better positioning of museums on the market by increasing their number.

We tried to find some answers in terms of checking whether a direct connection exists between the number of social media subscribers and the number of visitors to the museums. Secondly, we looked into the rating of museums on social media and analyzed whether ratings do indeed contribute to increasing the number of visitors.

To answer these questions, we conducted a research on 14 museums which are considered to be “popular” – seven located in Bucharest, Romania and seven located in Paris, France. The reasons why we chose this mix of museums are the following: the chosen museums are representative for the two European capitals; there is a variety of social networks used by these museums, which indicates an ongoing interest (in the case of the French museum) or a more recent interest (in the case of the Romanian museums) for personalizing the relationship with visitors through social media. The qualitative and quantitative analysis carried out refers to the year 2018 and the data was collected from the selected museums’ official websites and their respective Facebook, Twitter, YouTube, TripAdvisor and Google pages.

Also, as an added benefit from processing the information gathered through the specialized studies and from the statistical data regarding the museums, a “sketch” of both the typical French and the typical Romanian museum visitor was obtained. As one can expect, there are both similarities and *differences* between the two portraits, despite the fact that France and Romania have a similar cultural background. If the quantitative analysis suggests that the investment in social media is desirable, regardless of the size and notoriety of the museum, the qualitative analysis leads us to the conclusion that personalizing the relationship with visitors becomes a “must” for any museum regardless of its country of origin, in the sense of creating differentiated strategies for each type of visitor and, in particular, for Millennials, in terms of competitive advantage.

**Keywords:** Social media · Museum visitors · Targeted strategies · Millennials

## 1 Introduction

ICT (Information and communications technology), the web and social media transform the lives of museums by enriching their consecrated activities and functions (Hung et al. 2013). Thus, social media becomes a complementary tool for museums which helps them put into practice the policies associated to traditional missions which are destined for: education, facilitating social connections and spending quality spare time (Srinivasan and Fish 2009; Kidd et al. 2011).

Museums try to keep up the pace with the challenges of the continuously changing environment and as such they use social media for their benefit. They use Facebook, Twitter, YouTube, Instagram and other platforms in order to inform their public regarding their exhibitions and projects, to attract potential visitors, as well as to build and support the communities of interest created around the museum (Kidd et al. 2011; Villaespesa 2013). For instance, Chung et al. (2014) identified three marketing applications for which social media is used: building awareness, engagement with the community, and networking.

Moreover, the economic crisis of the last years affected museums because funding had been cut off from the budget. This put a constant pressure over museums on how to become more appealing to the public and to attract a larger number of visitors (Goulding 2000). Therefore, museums explore new and alternative ways to efficiently communicate with their visitors, to provide low prices, to increase the number of visitors and their level of participation, as well as support themselves with their own incomes. In this context, social media seems to support these efforts and to efficiently respond to the demand (Garibaldi 2015).

Social media offers museums the desired interactivity in the visitors-collections relationship, as well as a flexible and more personalized way of collaborating and making conversation with the public. For museums, this represents an opportunity to become more social and more participative (Simon 2010; Trant and Wyman 2006).

The flexibility and ease with which platforms can be used led to the active involvement of the public and to advent of user generated content (UGC) (Fletcher and Lee 2012). UGC represents a powerful tool that connects visitors with the ideas and the content of collections (Durbin 2016). On the other hand, social media offers visitors the possibility to expose comments, observations, memories or experiences, and to upload their own photos and videos taken throughout their visit. Therefore, social media transforms visitors from passive observers to active participants and content creators (Kidd et al. 2011; Villaespesa 2013; Coman et al. 2019).

The paper is structured, as follows: in the first section, we review the specialized literature about the impact of Web 2.0 on the life of museums. In the second section, we discuss the relationship between museums and visitors by means of social media. In the third section, we present the methodology used in the quantitative and qualitative analysis of the data collected from the 14 museums considered as “popular” in France and Romania and a “sketch” of the French and Romanian visitor. We mention the fact that

the selected museums are situated in Paris (France) and the Municipality of Bucharest (Romania), called in the period of time between the great world wars “Little Paris” due to the numerous similarities – art, culture, customs, architecture, etc. – to the French capital. The final part is dedicated to the conclusions and it includes the social media strategies we think museums should adopt in order to become more competitive.

## 2 What Could Social Media Platforms Represent to Museums? A Literature Review

Web 2.0 is a term that describes web-based applications on which users generate, share and curate the content (O’Reilly 2005). Over the last years, Web 2.0 sites, from blogs to YouTube to Wikipedia have transformed the ways that web users interact with content and with each other on the web.

While the more familiar Web 1.0 encompasses standard content providers, websites that give us information that place us in a rather passive position, as viewers in a museum, Web 2.0 removes the authority from the content provider and places it in the hands of the user. Nowadays, the user is the (active) participant who determines what is on the site and who judges which content is more valuable.

According to Simon (2010), Web 2.0 is social and democratic in the process of supporting diverse access paths to museums. And yet, there are some tensions that may turn museums against Web 2.0 such as: museums are strictly designed spaces while Web 2.0 platforms are open to all kind of user designs of varying levels of quality; museums launch exhibits in a “completed” state while Web 2.0 content is always changing; museums rely on authorities (curators and other professionals) while Web 2.0 relies on users who grant each other authority at will.

In this context, the popularity of Web 2.0 platforms such as Facebook, Twitter, Instagram or YouTube is on the rise at a phenomenal pace. For instance, Facebook had an increase of the number of active users in every month of 2019 of 0.13 billion in comparison to 2018 (Facebook 2019). The large number of visitors offers museums great opportunities and the chance to interact and create a personal relationship with them. If in the past, the relationship with the museum was of the one-way kind, due to the improvements in technology it is now a many-to-many type of relationship. This allows visitors to express their ideas, opinions and often cocreate content related to the object of the exhibition, together with professionals and other visitors through and with the help of social media (Russo et al. 2008; Russo 2015; Simon 2010).

Social media networks have become for visitors a new type of platform which is not only social but also intellectual because (Simon 2010): firstly, the people can contribute to the improvement of the exhibition content only by giving feedback or writing a review on the official page on the platform. Secondly, the process of learning in a museum becomes an individualized journey due to the help of social media in which the visitor can scout beforehand their specific interests or needs and therefore focus on them during their visit). Thirdly, social media networks are discussion platforms, i.e. real forums where visitors can exchange general ideas about art or about a specific field promoted by the museum. Thus, the discussed objects become social ones because they unite and coagulate real communities of interested people around them. Consequently, these

active users are those who make the best marketing for a museum, because they promote announcements, or the calendar of events or any other piece of information related to the schedule of the museum (Simon 2010; Kotler et al. 2008).

According to Simon (2010), platforms serve three categories of people: contributors, judges and lurkers. For instance, the majority of users on YouTube are lurkers: they watch video contents, but they do not share them. Then, we have the users who classify and evaluate, tag, comment or recommend the video content, but not even these share it. These people may be considered “judges” because they add metadata to a video by referring to its value and content. The most limited category is that of contributors who actually upload video content created by them. Contributors are also the ones that share existing videos made by others. Therefore, these people contribute and enrich experiences in a museum. At the same time, these become available experiences for all categories of users on the Internet.

There is no doubt that the presence of museums on social media is essential. Nevertheless, we also have to accept the fact that the choice of an appropriate social media channel is equally important. The social media networks we are referring in this study are Facebook, Twitter, YouTube, TripAdvisor and Google; each of these have different functions and different purposes.

In the last quarter of 2019, Facebook registered 2.45 billion active users per month (Facebook 2019). By active users, we refer to people who logged in over the last 30 days. The company declared that, for instance, in the fourth quarter of 2018, 2.7 billion people used at least one of its main products (Facebook, WhatsApp, Instagram or Messenger), each month. According to statistics (Mohsin 2019), Facebook is the social media platform that gets 60.6% of users on the Internet, ending up on the first place in the hierarchy. It is worth mentioning that 1.62 billion users access this platform on a daily basis. Every user, either an individual person or a company has a page on Facebook where they can present their profile. This makes contact between individuals or between consumers and companies easy to attain and manage through the platform.

Statistics also show us this: 57% of users are men whereas 43% are women; 72% of the 50–64 age segment and 62% of seniors over 65 are on Facebook. As for the young users, 88% of the 18–29 age segment and 84% of the 30–49 age segment are present on Facebook. What is equally interesting is the fact that 82% of graduates of higher studies have a Facebook page ([omnicoreagency.com/facebook-statistics](http://omnicoreagency.com/facebook-statistics)). Moreover, 96% of Facebook users have accessed this platform by means of mobile devices – either tablets or smartphones (DataReportal 2019). The data presented can be useful for museums in several ways. Museums can think of methods to adapt their online contents better, so that, for instance, their accessibility on mobile phones can become easier for visitors to view. Museums may also try to attract the group of users which represents an inactive public for the institution, but can possibly become engaged by means of targeted and differentiated strategies; these can be oriented towards the age segmented groups that use Facebook actively, and present potential from the point of view of cultural interests.

Unlike Facebook, Twitter has a smaller number of users per month – 330 millions – and 139 millions are active users every day (2019). In 2018, 80% of Twitter users used this platform on their phones; the total number of tweets sent per day was 500 million (2019). The same source mentions that 34% of Twitter users are women whereas 66%

are men, and 42% of Twitter users are present on the platform on a daily basis. Among the Twitter users, 37% are between 18 and 29, 25% are between 30 and 49, and 80% of users are part of the Millennial generation with a wealthy status ([www.oberlo.com/blog/twitter-statistics](http://www.oberlo.com/blog/twitter-statistics)).

According to Russo (2011), social networks, such as Twitter, help promote and build a loyalty towards the brand. Moreover, as it was also shown by Sarvanakumar and Suganthalakshmi (2012), Twitter is a fast and efficient platform and its use helps gathering more information about customers and augmented sales. Within this context, we suggest that museums could particularly benefit from using this social platform more frequently, with emphasis on creating content destined especially to young women who are interested in participating and getting involved in the world of museums, being able to provide material support for museums as well.

On the other hand, YouTube is a platform that communicates experiences by means of visual materials (Weilenmann et al. 2013). According to statistics (2019), there are 2 billion active users on YouTube per month nowadays, which sets this platform on the second place in the hierarchy of the most visited sites in the world. The total number of active users on YouTube per day is 30 million whereas the number of users who create video contents is approximately 50 million in the present (2018). The same statistics mention the fact that the age groups over 35 and over 55 are those that are increasing most rapidly in comparison to the other users. At the same time, the young people from the Millennials generation prefer watching a video content two times more than watching traditional television (2019). This platform is both a place for content publishing and a social media site; a promoting tool, as well as a place for entertainment and learning. Today, organizations of all kinds engage in the field of content creation that accomplishes a promoting function (for products/services), but this content does not necessarily have to contribute to sales. Museums are situated in an advantageous position from this point of view because they can use content marketing strategy as integrant part of their online presence. Furthermore, an expositional video content adapted to the Millennials generation could attract visitors under 35 to museums, motivated by the desire to explore, cocreate or socialize over the topic of works of art as life experience.

Zeng and Gerritsen (2014) state that social media sites are recognized as being an important source of information for travelers who make holiday plans. This is also the case of TripAdvisor platform. Travelers use TripAdvisor in order to get information regarding the quality of objectives and provided services, offering reviews and recommendations made by tourists who have already visited the objective. Within the context of museums, TripAdvisor constitutes a precious orientation tool used by visitors when they decide what museum is worth seeing and why.

In December 2017, when Google launched Arts & Culture app, the Internet exploded. Social media newsfeeds were full of selfies with people trying to find their double in a work of art within the collection of a museum from around the world. However, Google Arts & Culture app is more than a selfie generator. The application is a mobile version of the website and it combines all the cultural elements with the ability to topics, such as: artists, artistic movements, historical personalities or events, as well as various places.

On the other hand, Google Arts & Culture only sends you to exhibits owned by museums that concluded an online partnership with Google, integrating the Google

Street View function with virtual tours. However, there are great names from the world of museums that entered such partnerships with Google, such as: The Guggenheim Museum in New York, Musée d'Orsay in Paris or Rijksmuseum Museum in Amsterdam.

Google Arts & Culture has several interesting functions, such as Nearby – which provides visitors with information regarding cultural locations and events in the area; or the movement function and the ability to watch the movements in time or alphabetically and to create a gallery with one's favourite works of art. However, there is one disadvantage: if one looks for a specific artwork by using Google search engine, this will send him or her to the Google Arts & Culture app before sending the person to the site of the museum that owns the particular work of art. This means that Internet traffic could avoid the site of the institution, thus affecting statistics referring to online visitors.

If the social media platforms described above help us make an idea regarding their stature and potential, the next step will be to research who the visitors of museums in Paris and, respectively, Bucharest, are from the point of view of statistical data presented in specialized studies, referring to cultural consumption.

### 3 Museum Visitors

Thyne (2001) claims that despite the fact that museums are usually not linked to concepts of profitability and competitiveness, they still have to provide the best customer service. Therefore, they need to understand different segments of visitors regarding their demographic and psychographic traits. To gain a clearer image of museum use, we tried to review a number of findings of visitor surveys.

For example, Macdonald (2011) reported that the typical museum visitor was in the upper education, occupation and income groups, usually looking for opportunities to learn, to experience something new, to feel ease and comfort and also to participate actively. According to Rounds (2004), only a minor part of visitors attend the exhibitions in a through manner. What they attempt to achieve is their “total interest value” of the museum visit by focusing on “those exhibit elements with high interest value and low search costs” (p. 36). The extremely fast development of new technologies further accelerated these processes. Virtual demonstrations and displays and other computer-based interpretations of art have become today almost indispensable parts of exhibitions (Rentschler and Hede 2007).

#### 3.1 Who Are the Visitors of Museums in Romania?

The study of cultural consumption on the Municipality of Bucharest – 2016 – regarding the frequency of Internet use shows that 84% of the respondents use social media, particularly Facebook, Twitter, Google and Instagram whereas 16% of them look for information about digital events and 13% look for touristic information.

41% of the respondents – men – and 44% of the respondents – women – declare that they have not visited a museum in the last 12 months. However, it is also important to keep in mind that only 4% of men and 6% of women visited a museum more than 5 times in the last year. According to this study, the socio-demographic profile of the Romanian



**Table 1.** The frequency of visits to a museum depending on gender

| Gender            | Male | Female |
|-------------------|------|--------|
| Never             | 41%  | 44%    |
| Rarely            | 29%  | 23%    |
| 1–2 times         | 17%  | 21%    |
| More than 5 times | 4%   | 6%     |
| Don't know        | 1%   | 0%     |
| Don't respond     | 1%   | 1%     |
| Total             | 100% | 100%   |

Source: Cultural consumption on the Municipality of Bucharest (2016)

museum visitor is mostly feminine, aged between 36 and 50, higher education graduates (Table 1).

The same study shows us that the young adults up to 35 from Bucharest, as well as the elderly (over 65) in general show very little interest in visiting museums (Table 2).

**Table 2.** The frequency of visits to a museum depending on age

| Age               | 14–20 | 21–27 | 28–35 | 36–50 | 51–65 | Over 65 |
|-------------------|-------|-------|-------|-------|-------|---------|
| Never             | 30%   | 30%   | 27%   | 38%   | 48%   | 68%     |
| Rarely            | 34%   | 37%   | 34%   | 20%   | 24%   | 18%     |
| 1–2 times         | 22%   | 23%   | 21%   | 25%   | 16%   | 9%      |
| 3–5 times         | 7%    | 6%    | 9%    | 9%    | 7%    | 3%      |
| More than 5 times | 7%    | 4%    | 6%    | 8%    | 4%    | 2%      |
| Don't know        | 0%    | 0%    | 2%    | 0%    | 0%    | 0%      |
| Don't respond     | 0%    | 0%    | 1%    | 1%    | 1%    | 1%      |
| Total             | 100%  | 100%  | 100%  | 100%  | 100%  | 100%    |

Source: Cultural consumption on the Municipality of Bucharest (2016)

The study of cultural consumption on the Municipality of Bucharest is a survey made at the request of the Cultural Centre of Bucharest Municipality (ARCUB) in July–August 2015, on a sample of 1068 people over 14, with a  $\pm 3\%$  error and a 95% level of confidence.

### 3.2 Who Are the Visitors of Museums in France?

According to statistics, in France, 42% of the total number of visitors of a museum is male and 58% is female. The visitors' distribution on age segments is represented below (Table 3):

**Table 3.** Distribution of museum visitors: age segments in France

| Age segments | Percentages |
|--------------|-------------|
| – under 25   | 41%         |
| – 25–39      | 35%         |
| – 40–59      | 44%         |
| – 60–69      | 37%         |
| – Over 70    | 32%         |

Source: <https://www.statista.com>

According to the same study, from the total number of museum visitors, 22% visited a museum 1–2 times, 9% visited the chosen museum 3–4 times, and 8% – 5 times or more. We may remark the really high percentage of the public that did not/does not visit museums: 61%.

Therefore, the French museum visitor profile is: mostly female, between 40–59 years of age. Very close, we have the age categories under 25 and 60–69 (young seniors). The majority of visitors are graduates of higher studies. In accordance with the same source (statista.com), 61% of the French people are non-visitors (they have never visited a museum throughout the period of the analysis). The study was realized in June 2017–June 2018 on a sample of 2019 respondents.

We may notice the existence of several common grounds between the Romanian and the French visitor: in both cases, we can speak of a public that is *mostly female* with *higher studies*. On the other hand, we may notice that there are differences regarding the age categories: in Romania, the visiting public that prefers to go and visit a museum is part of the 36–50 age category whereas in France, the museum amateur public is situated with a majority in the 40–59 age category, followed by young adults under 25 and then, young seniors (60–69).

According to the same source with regards to the public that visits *the famous museums in Paris*, we may remark the great number of those who are foreign tourists. For instance, foreigners represented 73% of the visiting public of Louvre in 2018 (Walter 2019). In France, a visit to a museum is associated with a visit in the cultural world that belongs to the elite group. This takes on an exceptional character: as little over a fifth from the total number of visitors declares that they have been to the museum 1–2 times throughout the year. Only 8% of the respondents have chosen to visit a museum more than 5 times a year. The majority of visits were made together with families. Only 4% of the French people visit a museum on their own (Cultural Practices of the French).

The “Cultural Practices of the French” inquiry also shows that, if art museums occupy the first place on the national hierarchy, almost half of their visitors have exclusively visited another type of museum: specialized, scientific or ecomuseum. From here, we may deduce the fact that people who frequently go to art museums, visit the other types of museums as well. Thus, according to the mentioned study, 68% of art museum visitors have visited a scientific museum, 65% an ecomuseum and 71% a specialized museum.

### *Why are Millennials important for museums in the first place?*

According to Howe and Strauss (2000, p. 4), “as a group, Millennials are unlike any other youth generation in living memory. They are more numerous, more affluent, better educated and more ethnically diverse. More importantly, studies indicate that Millennials are gravitating towards group activity and believe in their collective power”. Each generation has a unique personality and preferences and our research reveals the fact that the largest category to visit French and Romanian museums are Millennials.

The Pew Research Center’s 2010 Report: “Millennials: Confident, Connected, open to Change” identifies the Millennials (generation Y) as the largest and most diverse in history. Typically defined by individuals born between 1980 and 2000, Millennials will soon step into the role of being a primary audience for cultural institutions in the 21<sup>st</sup> century. While not born into a world with this advanced mobile technology and social media, they are the first to grow up with it, due to the exponential growth of the communication landscape with the introduction of the internet in the early 1990s. They are the generation that has grown up with the internet, texting and social media, hence being called “digital natives”. This generation is also anticipated to be the *most educated* in history, *wealthy* enough to spend their money on culture or whatever *adds value* to their lifestyle: yet there is no way to know if Millennials will choose to support museums as they age.

Social media and mobile phones are more than a source of information and entertainment for Millennials: they have become essential to their self-expression and sense of connectedness. They are not the first generation to site technology as generation defining, however the way they fused it to their social lives is quite unique. By a majority (54%) Millennials think that new technology makes people closer to each other rather than more isolated. Studies also show that Millennials are *much more likely to contribute*: one in five Millennials (20%) have posted a video online, compared with only 6% of Gen X-ers, 2% of Boomers and 1% of those in the Silent generation or those over 65 years of age (Guasy 2012).

In other words, the digital age has created a learning environment that relies on Millennials’ abilities to critically navigate as well as interact with various bodies of knowledge. Millennials expect to shape content and not just consume the content of an exhibition (Adair et al. 2015). They are educated and technologically skilled, willing to contribute (in a creative way) and support (financially) the “inner life” of museums. Therefore, museums need to understand the important skills that this generation possesses and invest time and energy into designing and personalizing content for them.

## **4 Data and Methodology**

Our analysis of museum visitors was carried out on 14 museums in Bucharest and Paris (The National Museum of Art of Romania, “Grigore Antipa” National Museum of Natural History, Peasant Museum, National Museum of History of Romania, National Museum of Contemporary Art, Museum of Bucharest, “Dimitrie Gusti” National Village Museum, The Louvre, The Centre Pompidou, Musée d’Orsay, Musée des Arts Décoratifs, Musée de l’Orangerie, Musée de l’Armée, Cité des sciences et de l’industrie). The data was collected from the official sites of these museums, to which their official

pages on Facebook, Twitter, YouTube, TripAdvisor and Google are added. The data were processed in SPSS and Excel, the year being 2018.

### *The Popularity Concept*

We estimated the popularity level of a museum according to 4 indicators: number of visitors; Facebook rating; TripAdvisor rating; Google rating. At the same time, we considered as reference museums Louvre (Paris) and the Village Museum (Bucharest) which enjoy the best ratings on the 3 platforms – Facebook, TripAdvisor and Google. We included in the category of museums only those that have a rating of at least 3.5 on all 3 platforms. Thus, in the case of the 14 chosen museums, the situation is presented, as follows (Table 4):

**Table 4.** Museums' ratings on Facebook, TripAdvisor and Google

| Museums      | Visitors | Facebook rating | TripAdvisor rating | Google rating |
|--------------|----------|-----------------|--------------------|---------------|
| MNAR         | 127707   | 4.7             | 4                  | 4.6           |
| Louvre       | 10200000 | 4.7             | 4.5                | 4.7           |
| MNIR         | 52564    | 4.8             | 3.5                | 4.3           |
| MS           | 880000   | 4.8             | 4.5                | 4.6           |
| L'Armee      | 1208199  | 4.6             | 4.5                | 4.6           |
| D'Orsay      | 3286224  | 4.7             | 4.5                | 4.7           |
| L'Orangerie  | 1004287  | 4.8             | 4.5                | 4.6           |
| MNINGA       | 311639   | 4.6             | 4.5                | 4.7           |
| Pompidou     | 3551544  | 4.5             | 4                  | 4.4           |
| Cite_Science | 2231000  | 4               | 3.5                | 4             |
| duMAD        | 283959   | 4.4             | 4                  | 4.5           |
| MNTR         | 90849    | 4.2             | 4                  | 4.3           |
| MNAC         | 21093    | 3.9             | 3.5                | 4.3           |
| MMB          | 28000    | 4.8             | 4.5                | 4.5           |

Source: Facebook, TripAdvisor, Google

The formulated hypotheses are the following:

1. There is a direct and positive connection between the number of visitors of a museum and the number of their subscribers on social media platforms.
2. The bigger the rating of the museum is on social media, the more powerful is the rating influence on the number of visitors.

For the two hypotheses of the analysis, linear regressions will be applied. It is the most appropriate method when determining a relationship between a predictor variable and the response variable (Isaic-Maniu et al. 2003). In all the models presented below,

the verification of the fulfillment of the main conditions when applying regression was taken into account – normality, linearity, homoscedasticity, independency of observations (Andrei and Bourbonnais 2008).

The first hypothesis was to verify the existence of a positive link between the number of visitors and the number of subscribers on social media. We took into account the number of followers on Facebook, Twitter and YouTube (Table 5).

**Table 5.** Museums and their followers on social media in 2018

| Museums      | Facebook followers | Twitter followers | YouTube followers |
|--------------|--------------------|-------------------|-------------------|
| MNAR         | 18801              | 1534              | 89                |
| Louvre       | 2472671            | 1477115           | 39500             |
| MNIR         | 12022              | 731               | 34                |
| MS           | 30838              | 91                | 1                 |
| L'Armee      | 24412              | 11800             | 12100             |
| D'Orsay      | 835950             | 704200            | 16200             |
| L'Orangerie  | 92117              | 72971             | 656               |
| MNINGA       | 41597              | 123               | 29                |
| Pompidou     | 720109             | 1069041           | 7820              |
| Cite_Science | 109209             | 608182            | 5050              |
| duMAD        | 136284             | 55575             | 538               |
| MNTR         | 99758              | 4524              | 84                |
| MNAC         | 24242              | 0                 | 76                |
| MMB          | 16245              | 34                | 290               |

As expected, all three models have a strong coefficient of determination (R squared). The validity is achieved with low p-values ( $<0.05$ ) for the F-tests and the terms are statistically significant also with low p-values ( $<0.05$ ) for t-tests. The variables are highly correlated, fitting the line condition that is obtained.

The first scatter plot (Fig. 1) shows the strong positive relationship between the number of visitors and the Facebook followers, while the equation indicates the fact that an increase only by one unit in the number of Facebook followers produces an increase of 3.96 in the number of visitors.

The second scatter plot (Fig. 2) also shows a strong positive relationship between the number of visitors and the Twitter followers, with a 95% confidence interval, while the equation indicates the fact that if the number of Twitter followers increases by one unit, the number of visitors increases by 5.23.

The third scatter plot (Fig. 3) again shows a strong positive relationship between the number of visitors and the YouTube followers, with a 95% confidence interval, while the equation highlights the fact that, if the number of YouTube followers increases by one unit, the number of visitors also increases by 237.797.

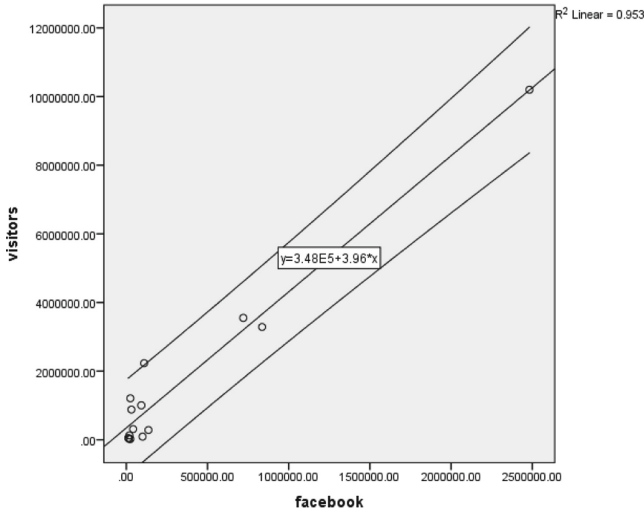


Fig. 1. The relationship between the number of visitors and Facebook followers

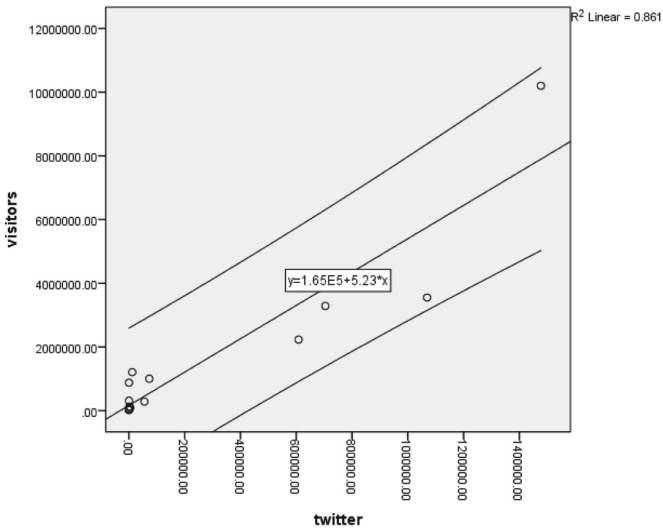
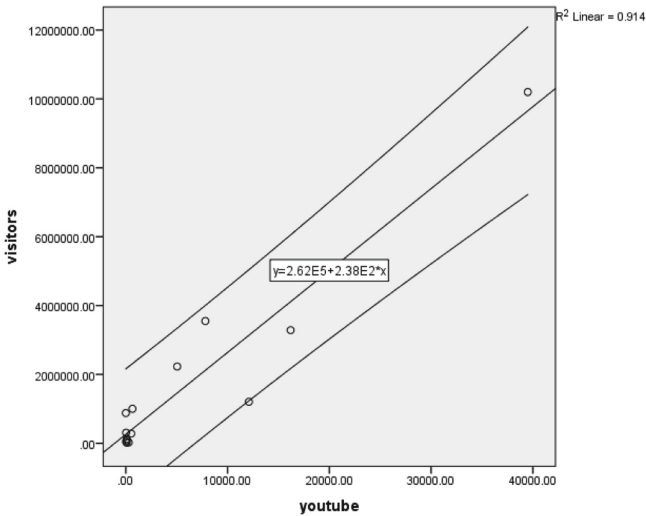


Fig. 2. The relationship between the number of visitors and Twitter followers

By conducting this research, the importance of social media in attracting visitors can be easily deduced. There is a direct link between those two variables and the models above mainly indicate the popularity of a museum. Many followers in any social media platform would definitely signify having many visitors. Another interesting fact that can be drawn from this analysis is that of the three social networks included here. It seems that the number of museum visitors is more sensitive to a change in the YouTube followers. Once more it is confirmed that people are more interested and engaged when



**Fig. 3.** The relationship between the number of visitors and YouTube followers

seeing videos (Reino and Hay 2016) and that YouTube plays an important role in the travel industry (Gale 2011) – thus concluding that one should definitely invest here when promoting the image of museums on social media.

The second hypothesis was to verify the influence of an increased rating on the number of visitors. Here, we took into account the ratings that were given on Facebook, TripAdvisor and Google.

In this case, the validity of the three models and the statistical significance of the terms included in the analysis were not fulfilled, the p-values of the tests being smaller than 0.05 (as seen in Table 6).

**Table 6.** Models summary

| Model | Independent variable | F-test | P-value | T-test | P-value |
|-------|----------------------|--------|---------|--------|---------|
| 1     | Facebook ratings     | 0.222  | 0.646   | 0.471  | 0.646   |
| 2     | TripAdvisor          | 0.954  | 0.348   | 0.977  | 0.348   |
| 3     | Google ratings       | 0.902  | 0.361   | 0.95   | 0.361   |

This might indicate the fact that an increased rating in the online environment does not necessarily influence the number of visitors. However, further investigations are recommended in the near future, when the methodologies related to the construction of more appropriate indicators for the social media environment will be developed.

## 5 Conclusions

1. The first hypothesis of our study according to which the use of social media by museums in their relationship with actual and potential visitors is reflected directly and positively on the number of visitors has been confirmed. Thus, by processing the data collected from the websites of the selected museums, we have the following result: when the number of followers increase in the case of the 3 platforms – Facebook, Twitter and YouTube – the number of visitors also increases. A higher value in the third model (Fig. 3) compared to the other two models indicates a potential for exploration here in future research.
2. On the other hand, the second hypothesis has not been validated. From our calculations, it resulted that the number of visitors is not influenced by the rating of a museum on Facebook, Google or TripAdvisor. Consequently, reviews and recommendations made on these platforms have an informing and documenting role rather than a role of awakening curiosity in the public. It is desirable to have a good rating on these platforms, but for museums, this rating rather tells the public that there are *exhibits that must be seen*; there are good *conditions* that make the visit a remarkable one. However, from the intellectual point of view, such information does not represent a triggering factor of decision in order to make a visit to a museum in reality.
3. For each of the two categories of museums – French and Romanian – we must think of differentiated strategies of personalization for their relationship with the visitors. Therefore, in the case of Romanian museums, attention has to be oriented especially towards the young public under 35, active on Facebook. At the same time, a challenge for museums is represented by the elderly (60–69 or over 69) also active on *Facebook* – who can be attracted to museums by uploading a video content on this platform.

Regarding the French museums, the amateur museum public is situated mostly in the 40–59 age category, followed by young adults under 25 and then, young seniors (60–69). We believe that the mature and young public should constitute the topic of some aggressive campaigns of attracting people to museums, carried out on *all 3 platforms* – Facebook, Twitter and YouTube. In this case, the only exception is Louvre: this museum has been carrying out specific events on social media with a definite target for every category of visitors for at least 10 years now.

Both categories of museums – Romanian and French – should also think of strategies to attract male public to the world of museums, the video content being recommended for them as well. Moreover, introducing some elements of “gaming” and interactive games could contribute to the increase of the number of male visitors (Bem-Neamu 2011).

4. A focus on attracting Millennials to museums  
According to our research, Millennials represent the largest group of museum visitors. Since they are “confident, self-expressive, liberal, upbeat and open to change” (Pew Social and Demographic Trends 2010) as well as technologically proficient, these people have the potential to challenge and transform museums into vibrant and very- much-alive institutions. A unique combination of generational characteristics,



technology and creativity makes Generation Y fit for sustaining museums' public mission. But, in order to appeal to younger audiences, museums need to develop new ways of working. Millennials' needs change quickly. Consumer markets change quickly as well. Rather than guessing what Millennials want, museums could use social media's flexibility and feedback as tools meant to allow them to know what the public really needs.

But what social media platform is best for attracting Millennials to the French and Romanian museums as well? Whether the platform is tried and true such as Facebook or trendy like Twitter and Instagram, the purpose of using social media is to help people connect to a museum's collection. Maintaining an active, friendly and engaging social media account shows the public that they are invited and encouraged to come into the physical space of museum.

According to the Pew Research Center Report (2010), 95% of participants in this study considered following museums on social media as a means of staying informed about events and 81% of all surveyed cited Facebook, Instagram and Twitter as their preferred platforms to follow museums. These numbers are significant, showing that young adults maintain an active presence on social media and will use these platforms to learn about museums. Considering this is a free resource, it makes sense to capitalize on this opportunity to reach more people.

To invest in social media is a long-term process: it takes time and constant attention in order to have results. Even if museums are fighting against a decrease in their budgets, it is obvious that any amount of money invested in activities/programmes on social media brings benefits, namely a greater number of visitors.

## 6 Limitations and Future Directions of Research

As data is difficult to obtain from all the museums over several years, the analysis had to be limited to the most recent year (2018), but as a future research direction, the time component is also recommended to better capture the influence of social media factors. Another limitation is that the impact of social media has increased in the last couple of years, as such any analysis would include only a few years, thus it is recommended to focus the research direction on panel data analysis, if possible.

There are at least two directions of research that are worth approaching in the following period: one of these is connected to the non-visitor public of museums. Taken the conditions in which the non-visitor public also uses social platforms, we believe that influencers could play the main role in attracting non-visitors to museums. At the same time, we believe that it would be interesting and useful to study how interactive gaming activities at the museum could be suitable to attract more visitors, so that their potential can be better valued in the benefit of individuals and communities.

## References

- Adair, B., Filene, B., Koloski, L.: Introduction. In: *Letting go? Sharing Historical Authority in a User-Generated World*. Left Coast Press, Philadelphia (2015)

- Andrei, T., Bourbonnais, R.: *Econometrie*. Editura Economica, Bucuresti (2008)
- Bem-Neamu, R.: Copiii in muzeu. Dilema veche (2011). <https://dilemaveche.ro/sectiune/dileme-on-line/articol/copiii-in-muzeu>
- Chung, T.-L., Marchetti, S., Fiore, A.M.: Use of social networking services for marketing art museums. *Museum management and Curatorship* **29**(2), 188–205 (2014)
- Coman, A., Grigore, A.-M., Ardelean, A.: The digital tools: supporting the “inner lives” of customers/visitors in museums. In: Meiselwitz, G. (ed.) *HCI 2019*. LNCS, vol. 11578, pp. 182–201. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-21902-4\\_14](https://doi.org/10.1007/978-3-030-21902-4_14)
- Durbin, G.: User-generated content on museum websites. *Museum ID* (2016). <https://www.museum-id.com>
- Fletcher, A., Lee, M.J.: Current social media uses and evaluations in American museums. *Mus. Manag. Curatorship* **27**(5), 505–521 (2012)
- Gale, A.: YouTube and the Tourism Industry (2011). <http://digitalresources.nz/article/m6ZktMh>
- Garibaldi, R.: The use of Web 2.0 tools by Italian contemporary art museums. *Mus. Manag. Curatorship J.* **30**(3), 230–243 (2015)
- Guasy, L.: Baby Boomers & Generation X-ers describe Millennials as lazy come recommended. *Content Marketing & Digital PR Come Recommended* (2012). <https://www.comerecommended.com>
- Goulding, C.: The museum environment and the visitor experience. *Eur. J. Mark.* **34**, 261–278 (2000)
- Howe, N., Strauss, W.: *The next great generation*. In: *Millennials Rising: The Next Great Generation*. Vintage Books, New York (2000)
- Hung, S.Y., Chen, C.C., Hung, H.M., Ho, W.W.: Critical factors predicting the acceptance of digital museums: user and system perspectives. *J. Electron. Commer. Res.* **14**(3), 231–243 (2013)
- Isaic-Maniu, A., Mitrut, C., Voineagu, V.: *Statistica*. Editura Universitara, Bucuresti (2003)
- Kidd, J., Ntalla, I., Lyons, W.: Sensing the social museum. In: *International Conference “Rethinking Technologies in Museums”*, Ireland (2011)
- Kotler, G.N., Kotler, P., Kotler, W.I.: *Museum Marketing and Strategy*. Jossey-Bass, San Francisco (2008)
- Macdonald, S.: *A Companion to Museum Studies*. Wiley-Blackwell, New York (2011)
- Mohsin, M.: 10 Facebook Stats Every Marketer Should Know in 2020 (2019). [oberlo.com/blog/facebook-statistics](https://oberlo.com/blog/facebook-statistics)
- O’Reilly, T.: What is Web2.0? (2005). <https://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>
- Pew Research Center: Millennials: confident, connected, open to change (2010). <https://www.socialtrends.org/2010/02/24/millennials-confident-connected-open-to-change>
- Reino, S., Hay, B.: “The use of YouTube as a tourism marketing tool”. *Travel and Tourism Research Association: Advancing Tourism Research Globally*, vol. 69 (2016). <https://scholarworks.umass.edu/ttra/2011/Visual/69>
- Rentschler, R., Hede, A.M.: *Museum Marketing: Competing in the Global Marketplace*. Butterworth-Heinemann, Oxford (2007)
- Rounds, J.: Strategies for the curiosity-driven museum visitor. *Curator: Mus. J.* **47**(4), 389–412 (2004)
- Russo, A., Watkins, J., Kelly, L., Chan, S.: Participatory communication with social media. *Curator: Mus. J.* **51**(1), 21–31 (2008)
- Russo, A.: Transformations in cultural communication: social media, cultural exchange and creative connections. *Curator: Mus. J.* **54**(3), 327–346 (2011)
- Russo, A.: Museums as creative incubators. In: *International Research Conference: Museum Communication: Practices and Perspectives*. Danish Royal Academy of Sciences and Letters, Copenhagen, Denmark, August 2015

- Saravanakumar, M., SuganthaLakshmi, T.: Social media marketing. *Life Sci. J.* **9**(4), 4444–4451 (2012)
- Simon, N.: The participatory museum (2010). <https://www.participatorymuseum.org/read/>
- Srinivasan, R., Fish, A.: Internet authorship: social and political implications within Kyrgyzstan. *J. Comput.-Mediat. Commun.* **14**(3), 559–580 (2009)
- Thyne, M.: The importance of values research for nonprofit organizations: the motivation-based values of museum visitors. *Int. J. Nonprofit Volunt. Sect. Mark.* **6**(2), 116–130 (2001)
- Trant, J., Wyman, B.: Investigating social tagging and folksonomy in art museums with Steve.museum (2006). <https://www.museumsandtheweb.com/mw2006/papers/wyman/wyman.html>
- Villaespesa, E.: Diving into the museum's social media stream. Analysis of the visitor experience in 140 characters. In: *MW 2013: Museums and the Web 2013*, Portland, USA (2013). <https://mw2013.museumsandtheweb.com/paper/diving-into-the-museums-social-media-stream/>
- Walter, S.: Personnalisation de la relation avec les visiteurs au musée du Louvre. In: *Museums Meet Museums Conference*, Bucharest, 11 October 2019, 3rd edn. (2019)
- Weilenmann, A., Hillman, T., Jungselius, B.: Instagram and the museum: communicating the museum experience through social photo sharing. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris (2013). <https://doi.org/10.1145/2470654.2466243>
- Zeng, B., Gerritsen, R.: What do we know about social media in tourism? A review. *Tourism Manag. Perspect.* **10**, 27–36 (2014)
- Studiu de consum cultural la nivelul orasului Bucuresti (2016). [https://issuu.com/incfculturadata/docs/2016\\_studiu\\_de\\_consum\\_cultural\\_la\\_n](https://issuu.com/incfculturadata/docs/2016_studiu_de_consum_cultural_la_n)
- Cultural Practices of the French. <http://www.pratiquesculturelles.culture.gouv.fr/doc/08synthese.pdf>
- DataReportal (2019). <https://datareportal.com/>
- Facebook (2019). <https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>
- [www.oberlo.com/blog/facebook-statistics](http://www.oberlo.com/blog/facebook-statistics)
- [www.oberlo.com/blog/twitter-statistics](http://www.oberlo.com/blog/twitter-statistics)
- [www.omnicoreagency.com/facebook-statistics](http://www.omnicoreagency.com/facebook-statistics)
- [www.statista.com](http://www.statista.com)
- <https://www.omnicoreagency.com/digital-marketing-statistics/>
- <https://muzeulbucurestiului.ro/>
- <http://www.mnac.ro/>
- <http://www.muzeultaranuluiroman.ro/>
- <https://madparis.fr/>
- <http://www.cite-sciences.fr/fr/accueil/>
- <https://www.centrepompidou.fr/>
- <https://antipa.ro>
- <https://www.musee-orangerie.fr>
- <https://www.musee-orsay.fr>
- <https://www.musee-armee.fr>
- <https://www.louvre.fr>
- <http://muzeul-satului.ro/>
- <https://www.mnar.arts.ro>
- <https://www.mnir.ro>



# Virtual Fitness Community: Online Behavior on a Croatian Fitness Forum

Kristina Feldvari<sup>(✉)</sup> , Anita Dremel , and Snježana Stanarević Katavić 

Faculty of Humanities and Social Sciences, Josip Juraj Strossmayer University of Osijek,  
Osijek, Croatia

{kfeldvari, adremel, sstanare}@ffos.hr

**Abstract.** Digital communication has influenced our lives in a number of ways, and the focus in this paper is on how people obtain, share and interpret health and fitness related information in the context of a virtual community on a Croatian fitness forum. The Internet is proven to be the leading source of information in Croatia and health-related content is among the most represented coverage people seek online. The purpose in this paper is therefore to analyze information needs expressed by users of the biggest fitness forum in Croatia and to look into the topics represented on it. Also, the aim is to gain insight based on empirical results into the use of information communication technology to create and sustain a sense of belonging and mutual support in this virtual community. For this purpose, we conducted a qualitative subject and content analysis of posts on the most active place on the [fitness.com.hr](https://fitness.com.hr) forum in 2019 (subcategory *How to lose weight*) and interviewed forum administrators and the most active forum members. Our results show that information needs of fitness forum users fall into six broader facets, all related to weight loss: nutrition, physical activity, psychological and health issues, personalized initial status, reporting results and other. Analysis also showed that a sense of belonging and emotional and peer support can in some ways be recognized in this fitness virtual community.

**Keywords:** Virtual community · Fitness literacy · Forum · Online behavior · Information needs · Croatia

## 1 Introduction

The Internet as a communication tool for seeking, searching and exchanging information has also become a widely used source of health-related information in the general population [1]. According to a WHO study, using Internet for seeking health related information has been constantly growing [2]. The newest results for Croatia from 2019 confirm that the Internet is the leading source of information for 78% respondents, and 87.7% respondents claim that the Internet gives them information not available elsewhere. When it comes to media content, coverage on health (28.8%) and sports (28.4%) are on the third and fourth place, after news from the world (51.7%) and local news (40%) [3]. We can therefore conclude that health and sports related information is among the most frequently sought information by Internet users in Croatia. Evidence also shows

that adults use the Internet to seek diagnosis or learn about a health problem or condition [4] and that the Internet serves as the primary source of information for health-related issues for young people [5], whereby nutrition and exercise are the topics most frequently searched for by the young [6]. Research conducted in Croatia has also shown that almost half of respondents obtain nutrition-related information from personal sources, including parents, trainers and fitness instructors, experienced individuals, nutritionists etc. Furthermore, 42.3% of the young state that their search for information relies on the Internet (web sites, forums, services and sources like Wikipedia etc.), while only 8.2% use other sources of information (books, journals, libraries) [7, 8]. The highest percentage of all health-related topics searched for thereby goes to nutrition [8]. Online health information seeking behavior (HISB) and health literacy, including fitness related themes such as nutrition and exercise, are evidently gaining ever greater worldwide attention, and we can see that Croatia is no exception to this.

In turn, contemporary social media and various virtual communities have the potential to impact health literacy of these online participants. Virtual communities are defined simply as social networks of individuals who interact through specific social media in order to pursue mutual interests or goals [9]. These virtual communities serve different users with different tasks, situations, and information needs, they are platforms that provide basic functionality in information exchange and sharing [10]. Fitness virtual communities are altering the way people are sourcing health and fitness information relating to dieting and exercise. Although the role of online fitness community is not specifically defined, it encompasses the concept of fitness culture where information linked to key concepts of health (like exercise and diet) is produced and distributed using typed and photographic or video communication. As SNSs revolutionize the concept of the audience into participants and users of information and communication technologies (ICT), new media technologies shape audience practice in new ways. This ability to create and exchange user generated content empowers online community members to produce themselves health related posts which can be easily and readily disseminated [11].

The broad motivation for this study is founded in our interest in complex nuances of social changes taking place in the face of new media and technologies, especially regarding the formation of communities online and the way they communicate about health and fitness. The purpose of this study is thus to explore information needs and fitness (health) related information behavior of a virtual fitness community on the Croatian forum [fitness.com.hr](http://fitness.com.hr), to determine the ways in which users and members of this fitness community share, present and interpret information related to fitness and health in the context of computer mediated text communication. However, the authors of this paper also wish to discuss far more complex phenomenology of this social change, pointing to the formation of virtual communities, which have recently excited much debate and research on whether and how communities are built on the Internet [12] and on the so-called social life of information and how communities form around fields of knowledge [13]. Focus is therefore evidently put on studying the modes and effects of online computer mediated communication in a specific field of interest (health and fitness) and on whether there are indicators of the formation of virtual community, with the sense of belonging and mutual support it provides. The interdisciplinary approach was therefore chosen in this paper, from information and communication science and sociology,

to analyze three key concepts: health and fitness related information behavior, health (fitness) literacy, and virtual community.

## 2 Theoretical Background and Literature Review

Information needs have for a long period been among central theoretical and empirical research interests. Widely, an information need is the state in which an individual or a community recognizes that their knowledge about a certain subject is insufficient and therefore desire, require or expect additional information in order to decide or act in a certain way. One of the divisions of information needs is into general information needs and information related to life needs such as health, nutrition, safety, emotional stability and intellectual advancement [14]. It is precisely based on health information that an individual understands and makes decisions about health [15] and not only in terms of healthcare but also concerning health related decisions we make in our everyday lives (e.g. in relation to obesity, nutrition, exercise etc.) [6, 16].

Health information needs are inseparably linked to HISB and health literacy. According to Niederdeppe et al., HISB is an activity that includes gathering information on health treatments, alternative medicine, nutrition and physical exercise [17]. Health literacy on the other hand is the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions [18]. Kickbusch states that health literacy is frequently wrongly interpreted and confused with medical literacy, and proposes an active, dynamic and empowering understanding of health literacy, which is an important life skill required in general population to navigate the choices in everyday life that influence health and well-being [16].

As HISB includes seeking information on health, nutrition and physical activity, the connection between nutrition and physical activity and their direct influence on health is revealed. The connection between health, nutrition and physical activity is also confirmed by two highly relevant documents - WHO Global Strategy on Diet, Physical Activity and Health from 2004 [19] and Global Recommendations on Physical Activity for Health from 2010 [20]. Another relevant contribution on health literacy from the point of view of skills related to decisions made about health in everyday life in connection with virtual communities is given by Stephanie T. Jong and Murray J. N. Drummond, who explored the nature of socially constructing healthy ideals online and how this potentially impacts health literacy (health knowledge and health practices) [21]. In addition to previously cited research in Croatia, the only paper on searching fitness related information was written in 2019 by Feldvari, Petr Balog and Faletar Tanacković, who compare different levels of information (fitness) literacy and fitness information seeking behavior of personal trainers with various levels of kinesiology education [22].

The definition of fitness on the other hand, according to Sports lexicon, implies ability, health and good physical condition [23]. Professional kinesiology literature defines fitness as a specific ability to utilize an organism's working capacity to perform a certain task under given conditions. Thus, fitness as a condition refers to the optimum quality of mental and physical systems and the ability to perform everyday tasks, with the purpose of reducing the risk of disease and increasing life quality [24].

We can therefore conclude that the definition of fitness is in accordance with the constitution, recommendations and strategies of WHO and that key concepts (health, nutrition, physical activity) relating to health literacy and fitness are the same. It is for this reason that we rightfully may speak of “fitness literacy” as a term narrower than “health literacy” when it comes to health-related information seeking and sharing. The definition of fitness literacy in this article partly overlaps with the skills and competencies comprising health literacy according to Zarcadoolas et al., and they include a wide range of skills, and competencies that people develop to seek out, comprehend, evaluate and use health (fitness) information and concepts to make informed choices, reduce health risks and increase quality of life [25]. However, the authors of this paper think that in defining fitness literacy the accent should be put on skills and competencies pertaining to two domains: nutrition and physical activity and exercise. The skills and competencies developed in these two areas help people make decisions about their health in everyday life (outside of the healthcare context).

Such personalized non-professional use of social media for health and fitness purposes is particularly on the rise because social media can provide a platform to not only gather information, but also seek support and share experiences [26]. Information is not the only social resource exchanged on the Net. Information behavior and the formation of virtual communities are significantly connected, in the sense of both specific (e.g. health and fitness related) information-oriented activities and emotional and peer-group support and other types of social interactions [27]. Such arguments valorize, above all, the interpersonal aspects of online interaction as opposed to the informational content that may be exchanged through communicative interactions [28].

We are today witnessing the development of community studies [29], but without grand theorizing and narratives, seeking instead to tell thicker stories in local settings [30]. The questions appearing in theory are on the formation and maintenance of virtual communities in relation to interaction and communication online, whether online communication facilitates social construction and transmission of knowledge, whether communities are being abolished, rebuilt in new ways or lead a double life [31]. Relevant theoretical claims for this study suggest that in virtual spaces new but still satisfactory forms of community are replacing old ones [27] and require distancing from the discourse of counterfeit versus genuine community [32]. There is today heightened awareness in scholarship that community is a cultural construct and Benkler [33] suggests that virtual communities would come to represent a new form of human communal existence, providing new scope for building a shared experience of human interaction. It is clear that virtual interaction is extremely complex, including thickening of pre-existing relations and the emergence of greater scope for limited-purpose relationships (such as fitness for example) that are loose but remain meaningful.

In the aim to find indicators in our material to study community formation, a sense of belonging and emotional and peer support on the Croatian fitness forum, virtual community was treated as interaction, communication or socializing of members in an online space dedicated to a specific purpose (health and fitness). According to Colachico [34], the authors presumed that more frequent interaction and the sense that members matter build a stronger sense of community, although the members might be relative strangers and their bonds not strong in the traditional sense. People join formed and structured

virtual communities to meet specific needs, for commercial interests, to find similar-minded people and to get support from them or for combined reasons. This means that virtual communities are not entities or products but present fluid processes that require interaction in which also social responsibility is developed and a virtual identity negotiated [35]. Also, supportive environment is built where the members are accommodated and thus empowered, especially through honoring their contributions [34]. Trust, satisfaction and communication are crucial in development of participation in a virtual community [36]. Availability of support, commitment to goals, cooperation and satisfaction grow only with group efforts. Members benefit from community membership by experiencing a greater sense of well-being and having more willing individuals to call upon. Reciprocally, the stronger feeling of community increases the flow of information; united around a central topic, virtual communities become important factors in creating and sharing knowledge. Conclusively, virtual communities can be from this lens seen as extensions of social environment that enrich users' experience. Although these theoretical points can be applied to various communities, the relevance of the sense of belonging and emotional and peer support operationalized in the way described above is particularly relevant and consequential when it comes to people's health and how they seek and share health-related information online forming thereby virtual communities around shared interests and needs.

### 3 Research

#### 3.1 Goals and Research Questions

This paper focuses on fitness (health) related information behavior of a virtual fitness community. The purpose is to determine the ways in which users and members of the forum [fitness.com.hr](http://fitness.com.hr) seek, share, present and interpret information related to fitness and health in the context of computer mediated text communication in a virtual forum community. Also, the purpose is to study the formation of the virtual community and a sense of belonging and emotional and peer support in it. This motivation led to interdisciplinary cooperation between the authors, who are researchers in the fields of information science and sociology. This research looks into the posts on the forum [fitness.com.hr](http://fitness.com.hr) in 2019 and analyses interviews with moderators and active members of the forum, focusing thereby on several research questions: 1) What information needs do forum users express? 2) What topics are represented on the forum and what experiences and feelings do users typically seek and share? 3) Are there indicators of the formation of a virtual community on this online fitness forum?

#### 3.2 Methodology

Fitness.com.hr ([www.fitness.com.hr](http://www.fitness.com.hr)) forum was chosen because it represents the largest fitness community in Croatia, with over 750 000 views per month, 50 000 registered members and 100 000 Facebook fans. The forum offers the following eight categories: Intro, Beginners, Fitness/Aerobic, Nutrition, Training, Challenge, Comments on articles and Miscellaneous. Within the main categories there are numerous subcategories, like



How to lose weight, How to increase muscle mass, Health, Nutrition, Exercise etc. [37]. In order to find answers to the research questions, the authors used two qualitative methods: subject and content analysis of posts and online interviews, described in the following section.

**Subject and Content Analysis.** The first method used in this research is subject and content analysis of posts on forum [fitness.com.hr](https://www.fitness.com.hr). We selected the subcategory How to lose weight for analysis, because the statistics showed that the biggest number of posts was published in this category. We analyzed the posts from this subcategory from 2019, with 20 December 2019 as the date when the last analyzed post was posted. The most frequently posted category is justified not only statistically, but also from the point of view of the connection between the frequency of interaction and the building of a sense of belonging to and support sought and provided in the virtual community.

There were seven topics opened in 2019 in the subcategory How to lose weight, five out of which were newly opened topics (Help with definition, Specific situation – help, Ketogenic diet – yes or no, Serious help, Help), and two were older topics reposted in 2019 (The quickest way to lose weight, Corner for middle-aged ladies). The authors analyzed a set of 128 posts and this dataset is not a sample ( $N = 128$ ) but is comprised of all posts from 2019 that appeared on the forum in the said subcategory.

[fitness.com.hr](https://www.fitness.com.hr) is available 24 h a day and is anonymous and asynchronous. This means that users do not use their real names and they do not need to be online at the same time in order to exchange information and communicate. Also, this study, although qualitative in nature, uses numbers and percentages, which bear no statistical importance, but are used to illustrate the obtained results.

A multi-stage qualitative subject and content analysis was conducted. The author who conducted the subject and content analysis wrote her doctoral dissertation in this topic and her narrow field of teaching and research pertains to subject indexing, information search and organization, and creation of subject-specific dictionaries and indexes [38]. In addition to this, the author is a fitness trainer with relevant formal and informal fitness education and 13-year experience in the field of fitness. For this reason, this author alone developed the coding scheme and the protocol for categories and topics. In the first stage of the analysis, prior to the analysis of topics, all posts were categorized into broader initial formal categories. A set of initial codes was derived based on existing literature [39], ISO standard [40] and literature where examples of coding schemes in relation to online weight loss and fitness in general can be found [41]. In our study the total population ( $N = 128$ ) of posts in the subcategory *How to lose weight* were coded using the initial coding scheme both to determine their value and to identify missing codes. In order to carry out subject and content analysis and code the post, the author determined under which topic(s) the post belonged. Many posts discussed multiple topics, and one individual post could be classified under more than one coded category. In this process, some codes were identified as irrelevant and therefore abandoned, and some new ones were added. Once the coding scheme was established, all posts were coded again by the same author in order to see whether established categories and topics could be grouped in a broader facet and to avoid possible errors. Eventually, the whole dataset was coded against two distinctive sets of codes: one referred to the loose initial formal category

of the type of posts and the second to the topic/theme of the post. Finally, the authors looked into emotions that were shared or asked about by posters.

**Online Interviews.** In the second part of research, the authors aimed at acquiring data on personal opinions and experiences of moderators and active users of [fitness.com.hr](http://fitness.com.hr) forum as well as their interpretation and understanding of the research topic. For this reason, the method of structured online interview was chosen. Asynchronous structured online interviews were conducted via e-mail with four forum moderators and the most active forum members, who are very good informants about information behavior and interaction on the forum. This sample was selected in a deliberate non-random way [42]. This type of interviewing is very useful for the reflective process, which helps to assure rigor. The respondents are from different towns in Croatia, which was another reason for choosing online interviews. This type of interview enabled maximum flexibility in obtaining detailed answers to the posed questions. One of the interviewers was herself the moderator of the forum and the member of this fitness community, which facilitated the communication with interviewees and the gathering of data. All interviewees were informed about the research goals, the protocol and all interview questions beforehand and provided their informed consent. They are all made anonymous in this study.

The online interview consisted of seventeen main questions with adjoining sub questions. The interviews were conducted in November 2019. Two members were selected based on their function as moderators (one active and one inactive member) and additional two forum members were asked to be interviewed based on their frequent responses to questions by forum users and their active information sharing (one yearlong and one more recent forum member).

### 3.3 Findings and Discussion

**Virtual Fitness Community Members Information Needs and Online Behavior.** Distribution of posts indicates that all analyzed posts fall into one of seven formal categories. Under the formal categories of posts, the following codes were distinguished (Table 1).

**Table 1.** Formal categories of posts (Post type)

| Post type                           | N  | %    |
|-------------------------------------|----|------|
| Questions by users                  | 52 | 40,6 |
| Current condition (give or seek)    | 71 | 55   |
| Planning, diet or training (give)   | 23 | 18   |
| Sharing experience by user or admin | 17 | 13   |
| Reply by user                       | 32 | 25   |
| Reply by admin                      | 33 | 26   |
| Giving advice by user or admin      | 34 | 27   |

*Formal Categories of Posts.*

1. Questions by users (poster seeks general or distinct piece of information or advice; e.g., “Do I have to stick to the macronutrients ratio strictly?”)
2. Current condition, diet or training (poster either shares or seeks general and specific information about current habits in nutrition and exercise; e.g. “I am 15 years old, weigh 111 kg, 195 cm tall and train box twice a day. I would like to lose 10 kg.”; “What are you currently eating and what is your physical activity level?”)
3. Planning, diet or training (poster seeks or shares general or specific information about making plans for nutrition and/or training, e.g. “Please write down at least for 2 or 3 days everything you eat, meal by meal. And kick junk food out and see where you stand, then plan your diet.”)
4. Sharing experience by user or admin (poster describes one’s situation without stating an explicit question, e.g., “Long-term keto did not work for me precisely because of the lack of energy during workout, already with 15 repetitions. It was easier to lose weight while on keto diet because you feel full and some people like fat, but I need carbs for training.”)
5. Reply by user (poster replies to a question and is not a moderator, e.g. “I think diet should be sustainable long-term and adjusted to your goal, you should not eat same things every day but have a balanced diet. You can look into carb cycling.”)
6. Reply by admin (moderator is the poster of reply to a particular question)
7. Giving advice (poster explicitly gives a piece of advice, e.g. “Make yourself go jogging at least 3 times a week, and eat vegetables plus do exercises for one body part every day.”)

Numbers and percentages here bear no statistical importance.

Posters most commonly seek or give information about their current condition (55%) and explicitly seek general or specific information about making plans for nutrition and/or training (40.6%). The following most represented category is giving advice by user or administrator (27%). The results show that administrators (26%) and other users (25%) equally reply to questions. To a lesser degree, users give specific plan of their diet or training (18%) in advance and share their own experience (13%). As far as the formal category of giving information about diet or training plan is concerned, it is clear that few users have a definite plan prior to posting a question, which indicates that their information needs on diet and nutrition were either not met or not clearly defined.

In 71 (55%) out of 128 posts there was a discussion about what we formally termed “current condition” and what refers to seeking (36 posts) and giving (35 posts) of information about detailed and specific habits of daily life regarding nutrition, exercise, level of activity, job, health etc. This indicates that moderators and active users are familiar with the basic conditions of weight loss and fat loss, which is reflected in the determination of TDEE<sup>1</sup> (total daily energy expenditure) [43] based on insight into users’ initial

<sup>1</sup> TDEE (total daily energy expenditure) is the most important information available to us when trying to burn fat and lose weight. It is the number of calories one burns in a day. This number is important to know as it gives one a baseline to compare current consumption and then adjust as needed to goals.

status and comprises an individualized approach. Administrators often ask about specific daily habits and seek additional personal information about question posters. The smallest number of posts contains personal experiences shared by moderators and users who reply to questions. The first likely reason for this lies in their awareness that not every individual experience can be applied to others, precisely because every user has a different initial status and TDEE, based on which calorie deficit needed for weight and fat loss is determined. The second reason is that posters are more inclined to propagate objective science-based knowledge and see themselves as its mediators.

*Post Topics.* Every topic appearing in all 128 posts was singled out after which 25 topics were extracted based on the qualitative principle of emergent research based in the gathered empirical material. After that, the author who performed the coding organized these 25 topics into six broader facets: Diet and weight loss (9 topics), Physical activity, training and weight loss (8 topics), Personalized initial status and weight loss (2 topics), Psychological and health issues and weight loss (2 topics), Reporting results (single facet), Other (3 topics) (Table 2).

Results show that topics in the subcategory on How to lose weight predominantly pertain to nutrition/diet (9 topics) and physical activity and training (8 topics), which is in line with the definition of fitness literacy given in this paper, where accent is put on skills and competencies pertaining to two domains: diet (nutrition science) and physical activity and exercise (kinesiology science).

The first facet mostly contains the posts discussing nutrition plans (72), specific food (59) and calorie deficit (52). This is congruent with scientific literature stating that despite the ability of physical activity and exercise to create calorie deficit, the actual impact of exercise alone on weight loss has often been found to be minimal [43].

Specific food is mostly mentioned in the sense of food that should be avoided in the process of losing weight, whereby administrators and active users point out that calorie deficit is the main premise of weight loss and that no specific food should be avoided if a person has no medical condition. This is the reason why these two topics have a very similar number of posts. It is interesting that although calorie deficit is indicated as the basis for weight loss, fewest posts in this facet are related to measuring and tracking of food and calories and to specific number of lost kilograms, volume centimeters and body fat. This shows that forum users are aware that calorie deficit is the foundation, but nonetheless do not take care of determining this process by monitoring calories or progress (kg, cm, bf). It should be pointed out that some moderators think it is not necessary to track calories through applications because they often give wrong information (e.g. "Don't worry about the calories and application tracking because you cannot know if the app miscalculated").

The Physical activity, training and weight loss facet contains the second largest number of posts (8 topics), with Weight training and Overtraining being the two most represented topics. This indicates that the majority of users get information and advice that not only cardio is sufficient for weight and fat loss, but that weight training is also crucial in the realization of the said goals. Fewest posts discussed at home exercise, which might be due to the fact that it is hard to find motivation to exercise alone or prefer to do it in company or fitness center. This also suggests that community-based aspects of support are important in achieving goals.

**Table 2.** Topics appearing in posts

| <b>Topics</b>   | <b>No. of posts</b> |
|---|---------------------|
| <i>Diet and Weight Loss</i>   |                     |
| Diet plan   | 72                  |
| Specific diet (keto, low carb, paleo etc.)  | 26                  |
| Specific no. of kg, cm or bf  | 20                  |
| Meals (breakfast, snacks, lunch, dinner, cheat meal)                                    | 44                  |
| Specific food questions   | 59                  |
| Macronutrients (carbs, fats, protein)   | 37                  |
| Calorie deficit   | 52                  |
| Supplement consumption  | 27                  |
| Calorie/food/macros tracking  | 22                  |
| <i>Physical Activity, Training and Weight Loss</i>                                      |                     |
| At home exercise  | 2                   |
| Gym exercise  | 27                  |
| Weight training   | 35                  |
| Cardio training   | 15                  |
| Specific training/exercises (body parts)  | 18                  |
| Overtraining  | 30                  |
| Training frequency (no. of trainings per week)  | 13                  |
| Activity level  | 17                  |
| <i>Personalised Initial Status and Weight Loss</i>                                      |                     |
| Diet, training history and lifestyle (stress, sleep, hydration etc.)                    | 60                  |
| Morphology status and body composition (height, age, weight, bodyfat, muscle mass, BMI) | 46                  |
| <i>Reporting Results</i>  | 25                  |
| <i>Psychological and Health Issues and Weight Loss</i>                                  |                     |
| Difficulties with diet  | 54                  |
| Health problems (hormones etc.)   | 32                  |
| <i>Other</i>  |                     |
| Weight gain   | 8                   |
| Muscle growth   | 12                  |
| Social media  | 4                   |

We can conclude that 4 topics (Diet plan; Diet, training history and lifestyle - stress, sleep, hydration etc., Calorie deficit and Morphology status and body composition - height, age, weight, bodyfat, muscle mass, BMI) refer to the initial status of user. This

means giving (by the poster asking a question) and receiving (by the moderator or user replying) information necessary for determining TDEE and calorie deficit needed for weight and fat loss based on it. Information from these subtopics can be considered general and basic information related to nutrition. Taking into consideration and calculating all this information, a user should be able to accomplish the goal of weight and fat loss. Other two most represented topics, Specific food questions and Difficulties with diet, together with topic Specific no. of kg, cm and bf and Health problems (hormones etc.) do not belong to basic but to specific information that mostly refer to specific health issues of an individual.

Considering the fact that psychological, emotional and potential health problems of an individual have a big impact on weight loss, they need to be taken into consideration together with general information (initial status of user) when striving to lose weight and fat. This is congruent with the results showing that posts regularly mention doctor's advice and a significant portion of posts (32) discuss health problems. One administrator often accentuated the psychological aspect as important in the process of losing weight (e.g. "How did you feel when you started to eat like this? Is it very different from your habits before? When were you at your ideal weight, when did you feel best?"). There are cases when the moderator gets in the conversation between two users after which users no longer interject and are thankful, which shows respect for the moderator (e.g. "Listen to her, she is an experienced moderator!").

It is important to point out that expert terminology and information needed for calculating TDEE (TDEE itself, NEAT or Non-Exercise Activity Thermogenesis, Exercise Activity or calories spent during training and TEF or Thermic Effect of Food) do not appear in any post. The reason for this might be that neither active moderators nor users who replied have formal education in nutrition science or kinesiology, which was confirmed in interviews with them.

It is worth mentioning that 3 out of 7 (43%) posters who initiated a topic about their problem in the subcategory How to lose weight reported later on their progress (e.g. "Hi people, here I am after 2 months, I lost 5 kgs so hit me if you have any diet advice, I train in the gym 4 to 5 times a week, 3 out of that with a trainer") and success that ensued based on information and advice they received on the forum. This shows that information and advice given on the forum satisfied information needs of 43% of users who posted a question in 2019. This is the reason for singling out the facet Reporting results. The final facet, Other, contains three topics: Weight gain, Muscle growth and Social media. The first two topics were probably posted in the How to lose weight subcategory because some users think it is possible to lose weight and gain muscle mass at the same time or because they failed to post in the right subcategory (e.g. How to grow muscle subcategory would be more appropriate). This could indicate the lack of familiarity with the forum organization and searching tools. Social media were discussed in relation to at home exercise in the smallest number of posts and referred to seeking fitness trainer recommendation on Instagram (e.g. "You can find trainers under#onlinecoach. It works, just send them a message and say you want to work with them, that's it! Good luck!!!").

**Virtual Community: Sense of Belonging, Mutual Support and Acceptance.** The results have shown that there are active members on the forum, who socialize and interact

online to meet specific needs. There were indicators suggested in theoretical background found in this forum that testify to the existence of a virtual community. The more members interact the higher sense of community is built, even if they remain relative strangers and do not bond strongly in the traditional sense. Even though it sometimes seems that there are material or even commercial interests involved (to find clients, to get advice free of charge, to advertise), members also express emotions and report on their progress, which contributes to the sense of continued belonging.

Weight-loss is an emotional issue and forum users express emotions and seek emotional support. Also, moderators seek information on emotions from users. Emotions most frequently, though not always verbatim, appearing in posts are: on the negative spectrum guilt and sin as the most frequent (in 5 posts explicitly, all posts longer than the average length, and often implicitly, e.g. "I have been naughty these days, eating sweets."), worry (5 posts, e.g. "I stick to my diet strictly because I do not want my weight to come back", "I worry because I lost my period"), frustration (4 posts, e.g. "I am not being unrealistic, I am just really stuck", "This is so frustrating, but let all evil be there"), sadness (2), fear and confusion; and on the positive spectrum satisfaction (5 posts, e.g. "I can finally fit into my old pants"), support (5 posts, "We are here for you to help you with your motivation"), pride (3 posts, e.g. "I think I got it all now with nutrition, and it works"), hope, joy, gratitude (4 posts, e.g. "I cannot thank you enough for all the help and advice you gave me, they were life changers") and determination (3 posts, e.g. "I so want to get rid of my fat until July").

Trust is crucial in health and fitness virtual communities and interviewees say that there are members with theoretical knowledge you can talk to and learn from, which also shows social responsibility. Interview respondents say they plan to continue their education in the field of fitness and that the aim is not to collect certificates (which serve profit making) or build a career, but that education in the field of fitness arises out of love. Their motivation to participate in the forum is the desire to learn and be informed in the field of fitness, but also the wish to be accepted as members of the fitness community that other members will turn to for relevant knowledge. They also stress that they only reply and give advice regarding health issues when they had such personal experience and otherwise direct users to seek professional medical advice. They give no answer at all if not completely sure it is a good one, because health and safety of people are top priority and it is better to use an individual approach to get more details on health of users seeking a reply. They say it is a great honor to be the moderator considering the time and energy put into the forum.

Interview respondents think that users most frequently share their positive experience to show their knowledge, boast and get motivation, while they share bad experience to seek or provide help and thereby strengthen further motivation. The positive effects singled out are the dispelling of various fitness myths, learning the basics about menus and micro/macronutrient's influence on satiety and hormones, and the support you get when you report on your progress in diet and/or training. The respondents stress there is on the forum no pressure or exposure of identity and agree that network mediated communication is a great relief for people who worry about what other people think about them or their problems. They point out joy and satisfaction when you help someone, and also gratitude and friendship when users share similar opinions with someone on the forum. There is one interview respondent who claims that deeper connection between the

members belongs to the past, primarily because social networking sites are taking over instead of specific forums, but still mentions friendship with older now inactive forum members and personal liking towards forum users with similar interests and opinions. Interview respondents accentuate that people who share similar positive experience tend to connect outside of the forum and mention the positive influence of the people they met over the forum on the realization of their goals.

The results have also shown that there is a supportive environment where the members are accommodated (individualized questions about habits, diagnoses, feelings, desires) and thereby also empowered, motivated and supported. The fact that the forum is asynchronous makes it possible to participate, bond and share even despite busy and conflicting working and life schedules of members. Availability of support, commitment to goals, cooperation and satisfaction grow only with group efforts. Strong sense of community is built by the experience that members matter. Members experience a greater sense of well-being when there are willing individuals to call upon and when they feel they received personal attention, which is frequently demonstrated in the posts and also confirmed by interviews with informants.

The stronger feeling of community reciprocally increases the flow of information and makes this virtual community an important factor in creating and sharing knowledge (e.g. members communicate about world-famous names from the field of kinesiology like Paul Cheka, Juan Carlos Santana (Human Kinetics), Mike Boyle, Stuart McGill, Joel Jamieson, John Berardi, Michol Dalcourt (Institute of Motion), Fabio Coman, Gray Institute etc.). The sense of community is boosted also when thankfulness is expressed, results and progress reported and individual members' contributions honored, which are points present in our results. This indicates that the forum virtual community is a type of extended social environment that enriches users' experience and deepens connection between the members and their similarity in values and reasoning regarding a healthy lifestyle.

## 4 Conclusion

This paper aimed at providing a look into wider social relevance of studies on online behavior and virtual communities, in this case regarding context of health and fitness information needs. Concerning information needs expressed by forum users, the findings show that many posts explicitly seek specific or general information, predominantly on nutrition/diet and/or exercise/training. This is congruent with the definition of fitness literacy given in this paper and with the domination of two domains in users' questions: nutrition science and kinesiology science. In line with previous research results that the young most frequently seek information about nutrition [8], which is a part of fitness literacy as shown in this paper, and that this is the most frequently sought information on the [fitness.com.hr](http://fitness.com.hr) forum within the subcategory How to lose weight, as well as that Croatia is the fourth most overweight country in Europe [44], the authors recommend the introduction of the content teaching fitness literacy into school programs in Croatia. As far as formal categories that posts are divided into are concerned, in over 55% of posts (40.6% of which are questions by users) the moderator or user, when replying to a question, seeks additional information about the initial status and condition of the



user: height, weight, age, activity level – training and NEAT (Non-Exercise Activity Thermogenesis), current nutrition (menus in meals and quantities of food), and general lifestyle (stress level, sleep, hydration etc.). The fact that users and administrators explicitly sought this information shows reluctance to give too general or superficial (default) replies or advice, and indicates that effort is put into an individualized approach in preparing replies. Advice given on forum frequently stimulates consultation with medical doctors, which is consistent with this interpretation. The lowest number of posts (17; 13%) contain shared personal experiences, which can indicate that posters rather propagate scientific knowledge.

It is also recommendable to include into kinesiology and food technology study programs in Croatia more courses that bring nutrition science and kinesiology together, as well as the courses on retrieving, searching and evaluating fitness and health related information. Considering the already stated fact that a 2017 study in Croatia showed that people frequently seek health and sports related information online [7], future research is welcome that would test fitness and health related information literacy in primary, secondary and tertiary education institutions and make it a part of the curriculum.

Conclusion can be made that new technologies and the pervasive use of the Internet are altering the way people are sourcing health and fitness information relating to dieting and exercise, namely online, under a nickname, from the comfort of their home, at their convenience, but not necessarily without common ties and the sense of community, because moderators and active users stress that users are ready to help each other and provide support. Still, they mention that other social networking sites are taking over and that the forum is less active and less relevant than 10 years ago. They recognize the tendency of respect for expertise and professionalism. The results indicate that participants within the online fitness community are overwhelmed with information related to health and fitness practices, which are discussed in great detail. The findings contribute to a broader understanding of complexities in the ways members of fitness communities obtain health and fitness information and emphasize the need for critical e-health and fitness literacy. Because of its potential influence on so many people's beliefs and perceptions, also regarding health, the future of the Net and of community, but also democracy, health, education, science etc. are deeply connected.

## References

1. Lambert, S., Loisel, C.G.: Health information-seeking behavior. *Qual. Health Res.* **17**, 1006–1019 (2007). <https://doi.org/10.1177/1049732307305199>
2. Andreassen, H.K., et al.: European citizens use of e-health services: a study of seven countries. *BMC Public Health* **7**, 1–7 (2007). <https://doi.org/10.1186/1471-2458-7-53>
3. Medijske navike u Republici Hrvatskoj: ožujak (2019). [https://showcase.24sata.hr/2019\\_hosted\\_creatives/medijske-navike-hr-2019.pdf](https://showcase.24sata.hr/2019_hosted_creatives/medijske-navike-hr-2019.pdf)
4. Fox, S., Duggan, M.: Health online 2013. Report, Pew Research Center's Internet & American Life Project (2013). [https://www.pewinternet.org/wp-content/uploads/sites/9/media/Files/Reports/PIP\\_HealthOnline.pdf](https://www.pewinternet.org/wp-content/uploads/sites/9/media/Files/Reports/PIP_HealthOnline.pdf)
5. Berkman, N.D., Davis, T.C., McCormack, L.: Health literacy: what is it? *J. Health Commun.* **15**, 9–19 (2010). <https://doi.org/10.1080/10810730.2010.499985>

6. Velardo, S., Drummond, M.J.N.: Understanding parental health literacy and food related parenting practices. *Health Sociol. Rev.* **22**, 137–150 (2013). <https://doi.org/10.5172/hesr.2013.22.2.137>
7. Peličić, G., et al.: Croatian children's views towards importance of health care information. *Collegium antropologicum* **36**, 543–548 (2012). nema doi
8. Martinović, I., Badurina, B., Bakota, S.: Informacijske potrebe i informacijsko ponašanje učenika i učenica I. gimnazije u Osijeku pri pretraživanju zdravstvenih informacija. *Vjesnik bibliotekara Hrvatske* **61**, 1–27 (2018). <https://doi.org/10.30754/vbh.61.2.691>
9. Rheingold, H.: *The virtual community: homesteading on the electronic frontier*. HarperPerennial, New York (1993)
10. Li, H., He, X., Hu, D.: Information Seeking and Sharing in Virtual Communities: A Case Study of Chinese IT Professionals. *Proc. Assoc. Inf. Sci. Technol.* **52**, 1–10 (2016). <https://doi.org/10.1002/pra2.2015.145052010030>
11. Jong, S.T., Drummond, M.: Online fitness communities and health literacies: critical digital awareness. In: 29th ACHPER International Conference: values into action - a brighter future: edited proceedings, pp. 158 – 168. Achper, South Australia (2015) Nema doi
12. Kim, A.J.: *Community building on the web: Secret strategies for successful online communities*. Peachpit, Berkeley (2000)
13. Brown, J.S., Duguid, P.: *The Social Life of Information*. Harvard Business School Press, Boston (2000)
14. Farrell, G.D.: Library and information service needs of the nation. In: Cuadra C. A., Bates, M. J. (eds). *Proceedings of a conference on the needs of occupational, ethnic, and other groups in the United States*, pp. 153. U.S. Govt. Print. Off., Washington, D.C (1974)
15. Ek, S.: Gender differences in health information behaviour: a Finnish population-based survey. *Health Promot. Int.* **30**, 736–745 (2015). <https://doi.org/10.1093/heapro/dat063>
16. Kickbusch, I.S.: health literacy: addressing the health and education divide. *Health Promot. Int.* **16**, 289–297 (2001). <https://doi.org/10.1093/heapro/16.3.289>
17. Niederdeppe, J., et al.: Examining the dimensions of cancer-related information seeking and scanning behavior. *Health commun.* **22**, 153–167 (2007). <https://doi.org/10.1080/10410230701454189>
18. Sørensen, K., et al.: Health literacy and public health: a systematic review and integration of definitions and models. *BMC Public Health* **12**, 1–13 (2012). <https://doi.org/10.1186/1471-2458-12-80>
19. World Health Organization. *Global Strategy on Diet, Physical Activity and Health*. <https://www.who.int/dietphysicalactivity/strategy/en/>
20. World Health Organization. *Global Recommendations on Physical Activity for Health*. <https://www.who.int/dietphysicalactivity/publications/9789241599979/en/>
21. Jong, S.T., Drummond, M.: Exploring online fitness culture and young females. *Leisure Stud.* **35**, 758–770 (2016). <https://doi.org/10.1080/02614367.2016.1182202>
22. Feldvari, Kristina, Petr Balog, Kornelija, Faletar Tanacković, Sanjica: Workplace information Literacy of Croatian fitness and conditioning personal trainers. In: Kurbanoglu, Serap, et al. (eds.) *ECIL 2018. CCIS*, vol. 989, pp. 191–200. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-13472-3\\_18](https://doi.org/10.1007/978-3-030-13472-3_18)
23. Flander, M. (Ed.): *Sportski leksikon: A-Ž. Jugoslavenski leksikografski zavod "Miroslav Krleža"*, Zagreb (1984)
24. Jurko, D., Čular, D., Badrić, M., Sporiš, G.: *Osnove kineziologije*. Gopal, Split (2015)
25. Zarcadoolas, C., Pleasant, A., Greer, D.S.: Understanding health literacy: an expanded model. *Health Promot. Int.* **20**, 195–203 (2005). <https://doi.org/10.1093/heapro/dah609>
26. Winkler, M. (ed.): *Community organizing and community building for health*. Rutgers University Press, New Brunswick (1999)

27. Wellman, B., Gulia, M.: Netsurfers don't ride alone: virtual communities as communities. In: Kollock, P., Smith, M. (eds.) *Communities in Cyberspace*, pp. 167–194. Routledge, London (1999)
28. Walther, J.B.: Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. *Commun. Res.* **23**, 3–43 (1996). <https://doi.org/10.1177/009365096023001001>
29. Bell, C., Newby, H.: *Community Studies: An Introduction to the Sociology of the Local Community*. Praeger, New York (1973)
30. Kingsley, G.T., McNeeley, J.B., Gibson, J.O.: *Community Building: Coming of Age*. Urban Institute, Washington, DC (1997)
31. Bellah, R.: *Habits of the Heart*. University of California Press, Berkeley (1985)
32. Freie, J.F.: *Counterfeit Community: The Exploitation of our Longings for Connectedness*. Rowman & Littlefield, Lanham (1998)
33. Benkler, Y.: *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, New Haven (2006)
34. Colachico, D.P.: Creating and Sustaining Community in a Virtual World. In: Information Resources Management Association (Ed.) *Virtual Communities: Concepts, Methodologies, Tools and Applications*, vol.1, pp. 1–13. Information Science Reference, Hershey & New York (2011). <https://doi.org/10.4018/978-1-60960-100-3.ch101>
35. McQueen, D.V., Kickbusch, I., Potvin, L., Pelikan, J.M., Balbo, L., Abel, Th: *Health and Modernity: The Role of Theory in Health Promotion*. Springer, New York (2007). <https://doi.org/10.1007/978-0-387-37759-9>
36. Jones, S.G.: Understanding community in the information age. In: Jones, S. (ed.) *CyberSociety: Computer-Mediated Communication and Community*, pp. 10–35. Sage Publications, Thousand oaks (1995)
37. Fitness.com.hr: najbolje mjesto na netu za aktivan i zdrav život. <https://www.fitness.com.hr/forum/>
38. Feldvari, K.: *Okvir za izradu i dizajn tezaurusa za označivanje*. Ph.D. Filozofski fakultet, Zadar (2014)
39. Information Resources Management Association. (Ed.): *Virtual Communities: Concepts, Methodologies, Tools and Applications*. Information Science Reference, Hershey & New York (2011). <https://doi.org/10.4018/978-1-60960-100-3.ch101>
40. ISO 5963:1985. Documentation- Methods for examining documents, determining their subjects, and selecting indexing terms. <https://www.iso.org/obp/ui/#iso:std:iso:5963:ed-1:vi:en>
41. Li, V., et al.: Losing it online: characterizing participation in an online weight loss community. In: *GROUP 2014: Proceedings of the 18th International Conference on Supporting Group Work*, pp. 35–45. Association for Computing Machinery, New York (2014). <https://doi.org/10.1145/2660398.2660416>
42. Salmons, J.: *Qualitative Online Interviews: Strategies, Design, and Skills*. SAGE Publications, Thousand Oaks (2014)
43. Kang, J.: *Bioenergetics Primer for Exercise Science (Primers in Exercise Science)*. Human Kinetics, Champaign (2008)
44. Eurostat.: Overweight and obesity – BMI statistics. [http://ec.europa.eu/eurostat/statistics-explained/index.php?title=Overweight\\_and\\_obesity\\_-\\_BMI\\_statistics](http://ec.europa.eu/eurostat/statistics-explained/index.php?title=Overweight_and_obesity_-_BMI_statistics)



# An Examination of Gaze During Conversation for Designing Culture-Based Robot Behavior

Louisa Hardjasa<sup>(✉)</sup> and Atsushi Nakazawa<sup>(✉)</sup>

Department of Intelligence Science and Technology, Kyoto University, Yoshida Honmachi,  
Sakyo-ku, Kyoto 606-8501, Japan

[louisa@ii.ist.i.kyoto-u.ac.jp](mailto:louisa@ii.ist.i.kyoto-u.ac.jp), [nakazawa.atsushi@i.kyoto-u.ac.jp](mailto:nakazawa.atsushi@i.kyoto-u.ac.jp)

**Abstract.** Gaze behavior, including eye contact and gaze direction, is an essential component of non-verbal communication, helping to facilitate human-to-human conversation in ways that have often been thought of as universal and innate. However, these have been shown to be influenced partially by cultural norms and background, and despite this, the majority of social robots do not have any cultural-based non-verbal behaviors and several lack any directional gaze capabilities at all. This study aims to observe how different gaze behaviors manifest during conversation as a function of culture as well as exposure to other cultures, by examining differences in behaviors such as duration of direct gaze, duration and direction of averted gaze, and average number of shifts in gaze, with the objective of establishing a baseline of Japanese gaze behavior to be implemented into a social robot. Japanese subjects were found to have much more averted gaze during a task that involves thinking as opposed to a task focused on communication. Subjects with significant experience living overseas were found to have different directional gaze patterns from subjects with little to no overseas experience, implying that non-verbal behavior patterns can change with exposure to other cultures.

**Keywords:** Gaze · Eye contact · Culture · Social interaction · Non-verbal communication · Social robotics

## 1 Introduction

Non-verbal behavior plays an important role in facilitating communication between humans, with gaze and eye behavior (including eye contact, eye movements and direction, etc.) being one of most primary, multi-purposed behaviors that has roles in both helping and influencing human-to-human communication. Several of the specific uses of gaze are motivated by social dynamics and expectations, and therefore is influenced by culture as well.

### 1.1 Gaze in Conversation

Gaze can enhance the conveying and interpretation of emotion [1–4], can be used to signal or gauge attention and interest [4–6], and indicate the roles of participants during multi-party communication [7]. Mutual gaze can be used to indicate both party's attentiveness

to a conversation, and direct gaze the conversation partner's eyes and face can similarly be used by a single party to indicate attentiveness or check the attentiveness of the other party, even if they have momentarily broken mutual gaze [4]. Mutual gaze that is too long or short based on the preferences of the conversation partners can encourage feelings of discomfort and unpleasantness.

Averted gaze especially is commonly used and interpreted in a gestural fashion to signal internal mental states such as thinking. Thinking gestures are a good example of how gaze behavior can be subconscious, but purposeful and sensitive to social motivations - people are more likely to look up while thinking when their conversation partner is facing towards them rather than away [8]. It is suggested that these thinking shifts may also help regulate cognitive load and relieve social pressure, as they become more frequent with increasing emotional content and task difficulty [4, 9]. Gaze shifts during conversations involving questioning are even sensitive to small differences such as question type (spatial vs. verbal), which affects the direction of initial gaze shifts following the question (verbal questions eliciting much more downward gaze) [10]. The temporal nature of the questions can influence this as well, with future-oriented questions pushing gaze to the right, and past oriented-questions encouraging shifts to the left [11].

Gaze behavior has a clear utility in social communication that can seem random or inconsistent at first glance, but actually has specific purposes and subconsciously conveys information to other parties.

## 1.2 Gaze and Culture

Despite the abundance of studies done in the context of examining the social impacts and uses of gaze in communication, the majority of studies in eye behavior are often done under the assumption that these behaviors are universal and driven by innate properties of human behavior. It is well-documented that much of social behavior, especially non-verbal, is often influenced by cultural norms and expectations and recognition that the majority of studies on gaze have been carried out primarily Westerns, and often in a descriptive rather than quantitative manner [12]. However, there are some studies that have demonstrated significant differences between the eye gaze displays and eye contact durations of subjects born in Canada, Trinidad, and Japan [8, 13]. With Canadian and Trinidad subjects looking upwards during thinking (facing their questioner) but the vast majority of Japanese subjects looking down. There are also some behaviors, many based on the role of the subject in the conversation (speaker vs. addressee) that seem to have more universality across cultures, such as the tendency for the addressee to engage in more direct/mutual eye contact in comparison to the speaker [12].

If the motivation behind looking up to signal thinking is primarily social, then it is reasonable to assume that the unique social pressures present in different cultures will result in different behaviors. Looking down often can have negative connotations in Western society since it is commonly associated with shame or embarrassment so its use goes down during face-to-face communication [8], while in Eastern societies like Japan, looking down can be used to indicate humility, which is seen as a more positive trait, which may explain the frequency with which it manifests. Looking upwards or directly too often or too long, consequently, can be interpreted negatively as arrogance,

intimidation or overconfidence in Eastern societies. Perhaps as a result of such cultural pressures, Japanese subjects typically display shorter eye contact durations during normal questioning [13].

Of course, culture does not only influence the display of gaze and eye behavior, but the perception and interpretation of gaze in a way that lines up with the typical displays of different cultures [14]. Consistently, Western populations show higher preferences for longer eye contact, while Eastern populations are more likely to interpret prolonged eye contact as indicating anger, unapproachability and other negative emotions and therefore show lower preference. It is also unlikely these behaviors are the result of some physiological differences between individuals born in different countries as opposed to cultural differences [15].

### 1.3 Gaze in Robotics

In recent years, implementations of gaze behaviors in robots has been increasing and recognition of gaze as an important nonverbal component to include in robot design has been on the rise. Most studies have been finding success in improving human impressions of robot conversation partners, inducing human-likeness, increasing the perception of autonomy and deliberateness in robots, and increasing user motivation and engagement in tasks using robot gaze [16–18] or even measuring human gaze and using that data to improve robot behavior [19]. Similar to human interactions, many studies have found that robots can also use gaze to indicate cooperative behaviors such as turn-taking and role-signaling and can take advantage of shared attention behaviors and signals that humans use regularly in conversation or interaction [20, 21].

A large number of studies of gaze in robots, however, have used head position as proxy for actual gaze indicated by pupils due to the overwhelming majority of easily commercially available social robots not having movable pupils or eyes. Even though such proxied gaze clearly improves human-robot interactions, head direction in multi-party human-to-human interactions has been found to be weakly correlated with actual gaze direction [22]. Historically, robots that display increasingly human-like behavior or behavior that is similar or matches the user's behavior [16] are often evaluated more positively than robots that are less or lack certain human behaviors [18]. Thus, there may be value in using and designing robots and robot behavior that can more faithfully replicate and demonstrate human gaze patterns.

### 1.4 Establishing a Baseline

In order to understand the role that culture has in gaze, we propose a study in which we robustly observe and record gaze behavior, controlling factors such as question and topic type (logical/analytical, personal, temporal orientation). We will be looking at variables such as duration of eye contact and averted gaze, direction of averted gaze, and average rate of shifts in gaze. These metrics and measuring methods were chosen to cover the majority of primary eye behaviors and for their simplicity. The data obtained in this experiment will be further used to aid in constructing general profiles of culturally-influenced eye behavior that can be easily compared, recognized and translated into other

mediums, such as robot behavior, so behaviors that cannot or are difficult to represent in that medium have been excluded.

It is expected that Japanese subjects will look down more often, have more shifts of gaze, and have less eye contact than Westerners. Western subjects would be expected to look upwards more frequently and have more prolonged eye contact. Although there are no Western subjects in this study, from a cultural research standpoint, it is important to observe the influence of time spent living abroad or experience with other languages, since with increased communication and exposure to other cultures, we may be witnessing a slow converging of cultures or simply a higher willingness and awareness of the need to adapt to other culture's behaviors among those with significant experience living or visiting other countries [23]. We would expect that Japanese subjects that are highly exposed to more Western cultures may display behaviors that reflect a middle ground between Western or Eastern, or, if their overseas experience is recent or dominating, a stronger behavioral 'switch flip' into the corresponding culture's behaviors. Analysis is conducted on both an overall Japanese basis as well as on a foreign exposure basis.

## 2 Method

### 2.1 Participants and Equipment

Ten Japanese students participated in the study. Seven participants had no experience being abroad beyond 1 month, and three participants had experience living in a country in either Europe (Italy, Germany), the Americas (Chile) or Oceania (Australia) for more than 2 years. All participants were male in order to reduce influences of gender on communication portions of the experiment. The subjects were recruited through Kyoto University's student part-time jobs recruitment system and had a mean age of 24.5 (SD = 2.99).

Over two different tasks, subjects were recorded with a Ricoh Theta V 360-degree camera for the entire duration. Subjects were also asked to wear an Empatica E4 Connect wristband, for one five-minute conversation session each in the first task, and for the entire duration of the second task. The E4 Connect wristband measures physiological responses such as inter-beat interval, blood volume pulse, heart rate, temperature, acceleration and EDA (skin conductance levels and response). Subjects were asked to keep their arms relatively still during the experiments since wrist and body acceleration was not a point of interest for this study, and was known to interfere with the readings of other physiological signals.

In the second task only, subjects wore a Tobii Pro Glasses 2 eye tracker device which is equipped with two eye cameras, a wide-angle scene camera, microphone and gyro and accelerometer. Nine out of ten subjects successfully performed a calibration with the eye tracker before beginning the experiment (Fig. 1).

### 2.2 Experiment

The experiment contains two tasks, communication and questions, in order to observe how gaze behaviors manifest in different conversational settings. Each experiment ran five subjects simultaneously to allow rotation of conversation partners. One communication session was lost due to recording error, and the eye tracking data of two subjects



**Fig. 1.** Sample image of experiment set-up with subject wearing the E4 Connect wristband and Tobii Pro Glasses 2 eye tracker, with a Ricoh Theta V 360-degree camera taking footage from the center of the table.

were lost due to a recording error, so the preserved 360-degree camera recordings were used primarily during analysis.

Each subject was recorded for approximately 22 min over the course of both tasks, with 4 five-minute communication sessions and 1 two-to-three-minute questioning session, for a total of roughly 210 min of conversation and questioning data (excluding the lost session) across ten different subjects.

**Task 1: Communication.** Task 1 observes simple, unrestricted communication and conversation behaviors when talking about various topics that relate to one's personal present, past and future.

Subjects were instructed to sit in groups of two, sitting across from each other with a small table in between. The subjects had not met before the experiment. Four of the five subjects were given a topic to discuss amongst themselves for 5 min each while one subject rested. The subjects spoke in their native language of Japanese. Topics were rotated as conversation pairs were rotated so that no subject talked about the same topic more than once, and no subject talked to the same person more than once. Among the given topics, two were neutral or present-oriented, two were past-oriented, and one was future oriented. See the Appendix for the topics given.

**Task 2: Thinking Questions.** Task 2 observes the behavior of subjects when challenged with a task that explicitly encourages thinking or recollection of random knowledge, as opposed to that of personal history or preferences.

Subjects were asked, individually, to answer 10 simple logical questions, designed to encourage either analytical thinking or remembering (see appendix for questions). The questions were obtained from previous studies [8, 13] in which cultural differences in



gaze behavior were observed. Similar levels of difficulty in the questions were maintained due to the known influence of difficulty on thinking behavior [9] but some questions were modified to be more independent of cultural or language biases. Given the questions asked are primarily logical, analytical and mathematical questions as opposed to spatial or personal, it is not expected that the type of question will have a significant effect on eye behavior, and the previous studies in which similar questions were used demonstrated no significant changes in behavior due to question type [8, 13]. The full roster of questions can be seen in the Appendix.

Subjects sat across from the experimenter in the same orientation across a table as they had in the communication task. All questions were read aloud in Japanese and printed on cue cards for the experimenter to read from aloud. Subjects did not get to see the questions on paper and were asked to answer verbally.

**Annotations.** Immediately after answering the questions, subjects filled out a brief demographic and cultural background survey detailing their experience with other cultures and countries.

Subjects were called back in the days following the experiment and asked to annotate videos of themselves with the duration and direction of their averted gaze (8 directions, ‘up’, ‘down’, ‘left’, ‘right’, ‘right-up’, ‘right-down’, ‘left-up’, ‘left-down’), and the duration of time they were looking directly at their conversation partner and attempting to establish eye contact or mutual gaze. An experimenter went through each of the annotations to fix minor errors and double-check annotations that subjects had marked as ambiguous or difficult to annotate (by writing two different directions or a question mark), using the angle of the gaze direction and notes from the subject to make the final decision.

### 3 Results and Discussion

The duration of participant gaze, divided into direct gaze, defined here as where the subject has established or is attempting to establish mutual gaze or eye contact by looking at their conversation partner’s eyes and face, and averted gaze, in which the subject is pointedly looking away from their conversation partner, was examined by average percentage durations of the conversations, and average seconds (s) duration.

#### 3.1 Direct vs. Averted Gaze

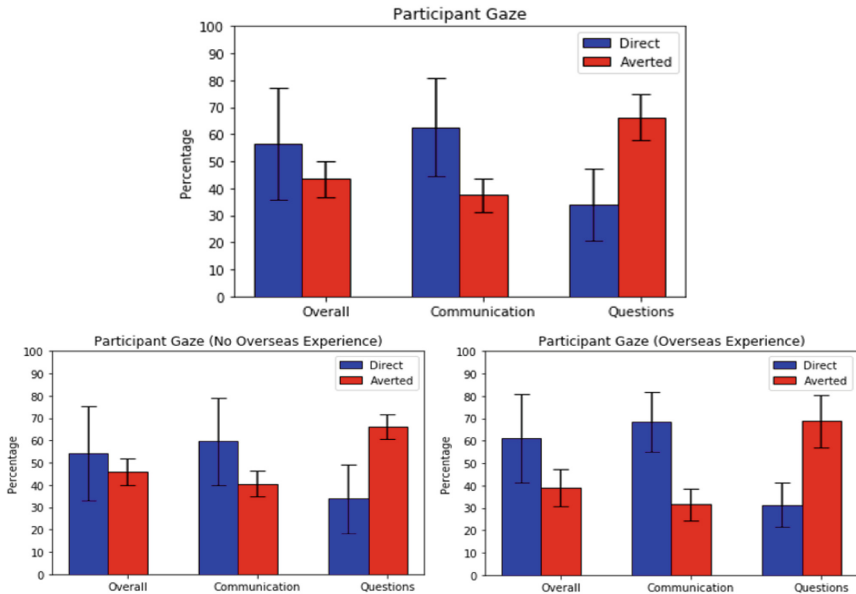
The average percentage durations of gaze for all subjects were broken down into direct gaze and averted gaze, according to task type and subject group. The gaze percentages when split into direct and averted add up to 100%. See Table 1 below for the mean percentage values and Fig. 2 for a visual comparison with error bars.

No significant differences were found between different populations of subjects, but across all subjects, the percentage of direct gaze was significantly higher,  $t(46) = 4.59$ ,  $p < .001$ , during the communications task in comparison to the questions task.

This may be due to the fact that tasks that have increased cognitive load, which has been associated with higher rates of averted gaze [9]. Consequently, direct gaze was

**Table 1.** Average duration percentages of direct vs. averted gaze across tasks (1 = communication, 2 = questions) and subject groups.

| (%)            | All subjects |      |      | No overseas experience |      |      | Overseas experience |      |      |
|----------------|--------------|------|------|------------------------|------|------|---------------------|------|------|
| Task           | All          | 1    | 2    | All                    | 1    | 2    | All                 | 1    | 2    |
| <i>Direct</i>  | 56.5         | 62.5 | 33.8 | 54.1                   | 59.5 | 33.8 | 57.6                | 63.9 | 33.2 |
| <i>Averted</i> | 43.5         | 37.5 | 66.2 | 45.8                   | 40.5 | 66.1 | 42.4                | 36.1 | 66.8 |



**Fig. 2.** Average durations (%) of direct and averted gaze across tasks and subject pools, all subjects (top), no overseas experience (bottom-left), significant overseas experience (bottom-right).

much higher during communication tasks, potentially due to the task being more social in nature and requiring more signaling of attention in comparison to the questions task.

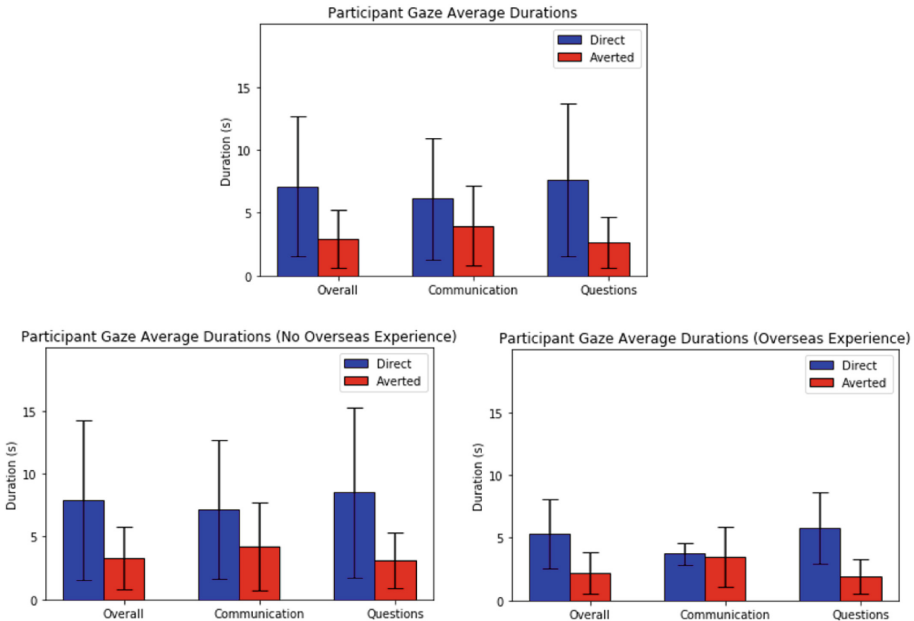
The amounts of direct gaze are also quite high in comparison to past records of Japanese eye contact in conversation, but this could be due to the definition of direct gaze here as attempts to establish eye contact in addition to genuine mutual gaze.

In addition to examining the average percentage of direct vs. average gaze, the average duration in seconds (s) of direct vs. averted gaze was broken down across task type and subject groups (See Table 2 and Fig. 3 for values and visualization, respectively).

In this context, the average duration of gaze is how long a single, unbroken period of gaze with no shifts is. Durations of direct gaze were typically higher across almost all tasks and subjects in comparison to averted gaze, and subjects with overseas experience had generally shorter durations of gaze across both direct and averted gaze. Subjects

**Table 2.** Average durations (s) of direct vs. averted gaze across tasks (1 = communication, 2 = questions) and subject groups.

| (s)            | All subjects |      |      | No overseas experience |      |      | Overseas experience |      |      |
|----------------|--------------|------|------|------------------------|------|------|---------------------|------|------|
| Tasks          | All          | 1    | 2    | All                    | 1    | 2    | All                 | 1    | 2    |
| <i>Direct</i>  | 7.10         | 6.12 | 7.60 | 7.92                   | 7.15 | 8.49 | 5.30                | 3.72 | 5.82 |
| <i>Averted</i> | 2.89         | 3.97 | 2.69 | 3.28                   | 4.22 | 3.10 | 2.18                | 3.47 | 1.93 |



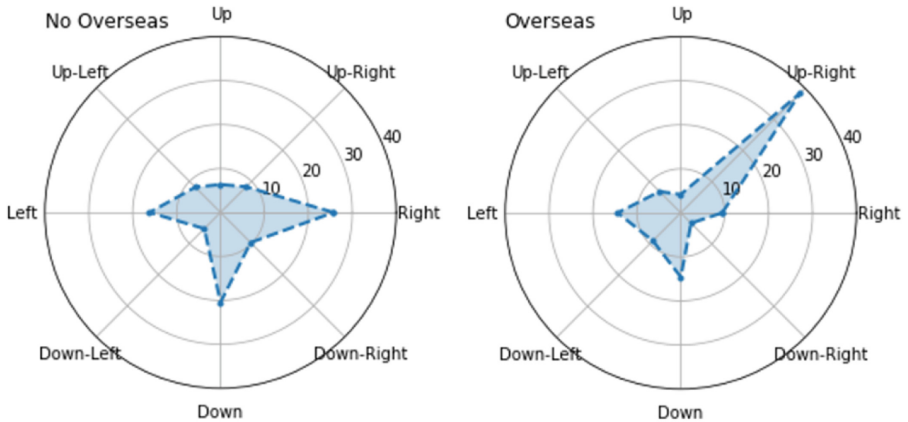
**Fig. 3.** Average durations (s) of direct and averted gaze across tasks and subject pools, all subjects (top), no overseas experience (bottom-left), significant overseas experience (bottom-right).

with overseas experience also had significantly higher average number of gaze shifts at 81 gaze shifts per session to 56 gaze shifts per session for subjects with no overseas experience,  $t(46) = 2.24, p < .05$ . Given that Westerns are typically expected to have longer durations of gaze, particularly during eye contact, and often demonstrate less shifts in gaze, this result cannot be explained by the subject’s exposure to other cultures.

### 3.2 Directions in Averted Gaze

The averted gaze duration percentages of each subject were further broken down into percentages limited to 8 directions (‘up’, ‘down’, ‘right’, ‘left’, ‘right-up’, ‘right-down’, ‘left-up’, ‘left-down’) such that the total percentage of all 8 directions equal 100% per subject, and compared across subject groups and task types.

In Fig. 4, we show a comparison of subjects with and without overseas experience, with subjects with experience showing a strong bias to the right-up direction, and subjects without experience showing a bias to the straight down and right directions. When broken down into the up, right, left, and down directions, the bias upwards in overseas experienced subjects is even more pronounced, with on average more than 50% of averted gaze trending up, right-up, or left-up, while averted gaze in no experience subjects trending down more than 30% and right more than 25% of the time.



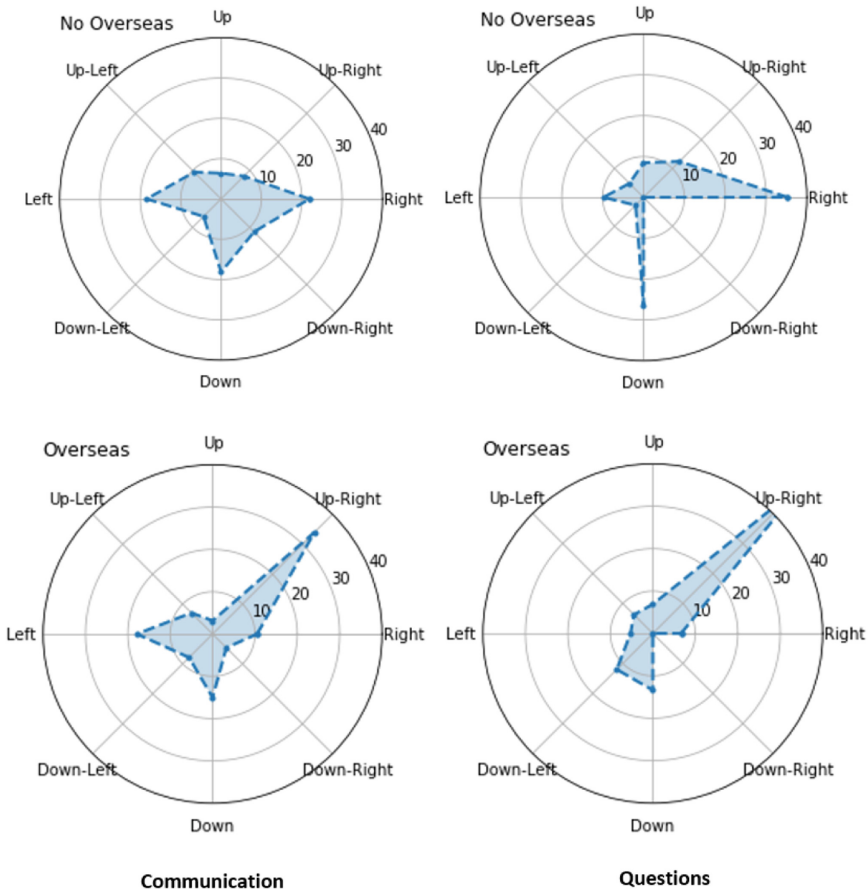
**Fig. 4.** Distribution of averted gaze directions by percentage and by whether subjects have no overseas experience (left) or significant overseas experience (right).

These results are fairly consistent with what is observed in comparisons between the directional gaze patterns in Canadians (Westerners) and Japanese subjects [8, 13], with Canadians tending to look right-up and Japanese subjects looking primarily downwards, though the bias upwards is even stronger in Canadians. This seems to indicate that despite all the subjects being Japanese natives who have spent the majority of their lives living in Japan, those that have been sufficiently or recently exposed to foreign cultures with different gaze behaviors may find their behavior mimicking and shifting towards that of what is typical in the novel foreign countries, and their behavior may arrive at a midpoint between their home and visited countries.

These results contribute to the notion that while in most cases we can proxy a subject's culture based on the country the subject lives in, an individual's cultural behavior is malleable, and the effects of residing in a foreign country and having recent and high amounts of exposure to foreign behaviors may be relatively persistent, so it is there important to take into account a subject's personal history when doing any sort of culture related research.

Analysis was also conducted according to the averted gaze directions across different tasks (Fig. 5). Across both subject groups, the questions task contributes more to the strong biases in the overall gaze direction patterns. Biases can still be observed in communication but they are much less pronounced, and the distribution of gaze directions is much more balanced. This could be attributed to the questions task requiring more active

thinking and cognitive load in comparison to casual conversation, as higher cognitive load has been associated with emphasizing or increasing the amount of averted gaze.



**Fig. 5.** Averted gaze breakdowns, separated by both task type – communication (left) and questions (right), and overseas experience – no experience (top) and with experience (bottom).

## 4 Conclusions and Future Works

This paper details a preliminary study in the scope of a larger project to construct culture-based robot gaze profiles. This experiment was run in order to observe the behavior of Japanese subjects, confirm established ideas about Japanese (and more broadly, Eastern) gaze behaviors, and solidify the tasks and methods to be used in future studies. One of the major aims of this study was also to observe how strong an influence cultural experiences, as opposed to country of residence or ethnicity, has an effect on human behavior.

The results in this experiment confirmed that cultural experience could have some significant effects on gaze behavior, particularly in averted gaze direction and frequency of gaze shifts or changes in gaze direction. Some behaviors were also found to be relatively consistent across different types of Japanese subjects such as the percentage of averted gaze to direct gaze, indicating that while some behaviors can be affected by cultural experience, some patterns are less malleable.

In subsequent studies related to culture and human robot interaction or nonverbal behavior, it is important to take into account the detailed experiences of the individual subjects, and analyze those with significant experiences living abroad from those with little to no experience. It is also important to take into cultural differences in behavior when designing robot nonverbal behavior, and consider carefully who the target audience of such behaviors are – especially if the goal of implementing human-like behaviors in the robots is to increase familiarity and therefore evaluation of the robot. A mismatched cultural behavior profile could potentially result in a reduction in the expected or desired effect, but this required confirmation in a future study.

Additionally, the current study lacks a sufficient number of subjects in order to make conclusive claims about nonverbal behavior. While there is a sizable amount of data per subject, at ten subjects total, increasing the *n* per group to the typical 15 to 20 in this field would allow us to make more confident conclusions. This research would also highly benefit from a direct comparison between Western subjects and Japanese subjects, including Western subjects with significant experience living in Eastern countries or exposure to Eastern culture.

Cultural surveys were conducted to confirm whether subjects maintain cultural behavior that matches their country of birth when subjects have exposure to environments significantly different from their native culture, but the surveys were lacking comprehension. Future investigation on the persistence of foreign cultural influence on behavior should include more detailed surveys that include questions about the recency and the type of environment lived in. There is also a possibility that subjects that tend to seek residency overseas naturally display behaviors or personality traits that are close to their target country, so ideally, a long-term study that observed subjects before and after several years into long-term residency would be the most conclusive. Future iterations would also benefit from implementing a comprehensive personality test prior to the experiment, in order to observe how much and whether personality contributes significantly to non-verbal behavior, and in turn confirm the degree to which culture and personality traits coincide.

Further analysis using current and future experiment data could also reveal more insights about conversational gaze behavior, such as looking more closely at the differences between direct gaze, which is defined here as gaze directed towards the conversation partner's face and eyes, and mutual gaze, in which there is a genuine mutuality in the gaze. Examining the data pairs one by one and comparing interactions between Japanese subjects with and without overseas experiences as opposed to Japanese subjects only with or only without overseas experience would be valuable and more comprehensive. And further examining the roles of participants during conversation and confirming whether previous results finding that levels of eye contact vs. averted gaze remain rather

consistent across cultures would be especially important in constructing gaze profiles [12].

Future work for this research also includes using the data acquired from such human-to-human communication experiments to implement culture-based profiles for human-robot interactions, and test whether users are sensitive and/or biased towards profiles that are familiar and close to their own culture or profiles that are foreign and novel. Trends in research related to robots and culture indicate that the majority of users prefer robots that display behavior that is similar to their own or similar to the established norms in the user's residing country in areas such as proxemics, so we would expect similar results in future experiments. It would be valuable to observe as to how subjects with mixed backgrounds (an Eastern native with experience living in the West, and vice versa) would respond to such cultural-based gaze profiles.

**Acknowledgements.** Special thanks to Anjanie McCarthy for providing the original collection of questions used in previous culture and gaze studies. And a special acknowledgment and thank you to Miyuki Iwamoto for her help in running experiments and providing valuable guidance and advice on equipment. This work was supported by RIKEN-AIP, JST CREST Grant Number JPMJCR17A5 and JSPS KAKENHI 17H01779, Japan.

## Appendix

|   |                      |
|---|----------------------|
| Talking topics                                      |                      |
| Question  | Temporal orientation |
| How you spent your summer vacation                  | Past                 |
| Your favourite holiday of the year                  | Neutral/Present      |
| As a kid, what did you want to be when you grew up? | Past                 |
| Your favourite type of cuisine/food                 | Neutral/Present      |
| The next place you would most like to travel to     | Future               |

### Thinking Questions.

1. What is the name of a flower that is also a female name?
2. If your mother's brother has a child what is your relationship to that child?
3. What is the name of a fruit with red flesh?
4. If the current day is Thursday, what day will it be after 16 days?
5. What is  $9 \times 4 / 2 + 7$ ?
6. What is the 14th letter of the alphabet?
7. It took A 12 h to run to B's house at 8 km per hour. How far is it between A's house and B's house?
8. What is the name of a month that has only 30 days?
9. What is a color that is NOT found in a rainbow?
10. How many words can a person with a typing speed of 65 wpm, type in 8 min?

## References

1. Adams, R.B., Kleck, R.E.: Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion* **5**, 3–11 (2005). <https://doi.org/10.1037/1528-3542.5.1.3>
2. Rychlowska, M., Zinner, L., Musca, S.C., Niedenthal, P.M.: From the eye to the heart. In: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction - Gaze-In 12 (2012). <https://doi.org/10.1145/2401836.2401841>
3. Ganel, T., Goshen-Gottstein, Y., Goodale, M.A.: Interactions between the processing of gaze direction and facial expression. *Vis. Res.* **45**, 1191–1200 (2005). <https://doi.org/10.1016/j.visres.2004.06.025>
4. Kendon, A.: Some functions of gaze-direction in social interaction. *Acta Physiol.* **26**, 22–63 (1967). [https://doi.org/10.1016/0001-6918\(67\)90005-4](https://doi.org/10.1016/0001-6918(67)90005-4)
5. Frischen, A., Bayliss, A.P., Tipper, S.P.: Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychol. Bull.* **133**, 694–724 (2007). <https://doi.org/10.1037/0033-2909.133.4.694>
6. Kleinke, C.L.: Gaze and eye contact: a research review. *Psychol. Bull.* **100**, 78–100 (1986). <https://doi.org/10.1037/0033-2909.100.1.78>
7. Nakano, Y.I., Yoshino, T., Yatsushiro, M., Takase, Y.: Generating robot gaze on the basis of participation roles and dominance estimation in multiparty interaction. *ACM Trans. Interact. Intell. Syst.* **5**, 1–23 (2015). <https://doi.org/10.1145/2743028>
8. McCarthy, A., Lee, K., Itakura, S., Muir, D.W.: Gaze display when thinking depends on culture and context. *J. Cross Cult. Psychol.* **39**, 716–729 (2008). <https://doi.org/10.1177/0022022108323807>
9. Doherty-Sneddon, G., Phelps, F.G.: Gaze aversion: a response to cognitive or social difficulty? *Mem. Cogn.* **33**, 727–733 (2005). <https://doi.org/10.3758/bf03195338>
10. Ehrlichman, H.: Effects of verbal and spatial questions on initial gaze shifts. *Neuropsychologia* **12**, 265–277 (1974). [https://doi.org/10.1016/0028-3932\(74\)90012-8](https://doi.org/10.1016/0028-3932(74)90012-8)
11. Martarelli, C.S., Mast, F.W., Hartmann, M.: Time in the eye of the beholder: gaze position reveals spatial-temporal associations during encoding and memory retrieval of future and past. *Mem. Cogn.* **45**(1), 40–48 (2016). <https://doi.org/10.3758/s13421-016-0639-2>
12. Rossano, F., Brown, P., Levinson, S.C.: Gaze, questioning, and culture. *Conv. Anal.*, 187–249 (2009). <https://doi.org/10.1017/cbo9780511635670.008>
13. McCarthy, A., Lee, K., Itakura, S., Muir, D.W.: Cultural display rules drive eye gaze during thinking. *J. Cross Cult. Psychol.* **37**, 717–722 (2006). <https://doi.org/10.1177/0022022106292079>
14. Senju, A., Verneti, A., Kikuchi, Y., Akechi, H., Hasegawa, T., Johnson, M.H.: Cultural background modulates how we look at other persons' gaze. *Int. J. Behav. Dev.* **37**, 131–136 (2013). <https://doi.org/10.1177/0165025412465360>
15. Akechi, H., Senju, A., Uibo, H., Kikuchi, Y., Hasegawa, T., Hietanen, J.K.: Attention to eye contact in the west and east: autonomic responses and evaluative ratings. *PLoS One* **8** (2013). <https://doi.org/10.1371/journal.pone.0059312>
16. Andrist, S., Mutlu, B., Tapus, A.: Look like me. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI 2015 (2015). <https://doi.org/10.1145/2702123.2702592>
17. Andrist, S., Tan, X.Z., Gleicher, M., Mutlu, B.: Conversational gaze aversion for humanlike robots. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction - HRI 2014 (2014). <https://doi.org/10.1145/2559636.2559666>
18. Mutlu, B., Forlizzi, J., Hodgins, J.: A storytelling robot: modeling and evaluation of humanlike gaze behavior. In: 2006 6th IEEE-RAS International Conference on Humanoid Robots (2006). <https://doi.org/10.1109/ichr.2006.321322>



19. Nakano, Y.I., Ishii, R.: Estimating users engagement from eye-gaze behaviors in human-agent conversations. In: Proceedings of the 15th International Conference on Intelligent User Interfaces - IUI 2010 (2010). <https://doi.org/10.1145/1719970.1719990>
20. Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., Ishiguro, H.: Conversational gaze mechanisms for humanlike robots. *ACM Trans. Interact. Intell. Syst.* **1**, 1–33 (2012). <https://doi.org/10.1145/2070719.2070725>
21. Moon, A., et al.: Meet me where im gazing. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction - HRI 2014 (2014). <https://doi.org/10.1145/2559636.2559656>
22. Vrzakova, H., Bednarik, R., Nakano, Y.I., Nihei, F.: Speakers head and gaze dynamics weakly correlate in group conversation. In: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications - ETRA 2016 (2016). <https://doi.org/10.1145/2857491.2857522>
23. Anawati, D., Craig, A.: Behavioral adaptation within cross-cultural virtual teams. *IEEE Trans. Prof. Commun.* **49**, 44–56 (2006). <https://doi.org/10.1109/tpc.2006.870459>



# Investigation on the Fusion of Multi-modal and Multi-person Features in RNNs for Detecting the Functional Roles of Group Discussion Participants

Hung-Hsuan Huang<sup>1,2(✉)</sup> and Toyoaki Nishida<sup>1,2</sup>

<sup>1</sup> Faculty of Informatics, University of Fukuchiyama, Fukuchiyama, Japan  
hhhuang@acm.org

<sup>2</sup> Center for Advanced Intelligence Project, RIKEN, Kyoto, Japan

**Abstract.** More and more companies are putting emphasis on communication skill in the recruitment of their employees and adopt group discussion as part of their recruitment interview. In our ongoing project, we aim to develop a training system that can provide advices to its users in improving the perception of their communication skill during group discussion. In order to realize this goal, a conceptual unit of communicational behaviors and a template of communication style are required. We propose the use of *functional roles* of the participants in group discussions as this unit. In order to incorporate the use of functional roles for improving the perception of participants' communication skill, the first task is automatic detection of the participants' functional roles in real-time. We previously proposed a SVM based model for this task but the results were only moderate. We expect including temporal characteristics, frame-wise interaction of modalities, and inter-person interaction can improve the classification accuracy and explored the use of RNN based networks to see the effectiveness of these factors.

**Keywords:** Functional roles · Multiparty interaction · Group discussion · Multimodal interaction · Deep learning · Recurrent neural network (RNN) · Long short-term memory (LSTM) · Gated recurrent unit (GRU)

## 1 Introduction

While companies are running projects, the communication skill of individual member largely affects the relationship with other members and thus has great influence on team performance. According to the investigation conducted by Japan Business Federation (Keidanren), communication skill has been the most important factor in recruiting new graduates for more than 15 years [14]. There is a growing number of companies that adopted group discussion in the recruitment of their employers (38% of the major companies [13] in Japan). During a

group discussion interview, job applicants have to collaborate with each other to deliberate productive results on an assigned topic. During the discussion, their communication skill and personality are observed by the investigators of the companies. Therefore, the perception of higher communication skills is expected to lead the applicant's success in job hunting.

In our ongoing project, we aim to develop a training system that can provide advices to its users in improving the perception of their communication skill in group discussion. It has been proofed that the participants' communication skill of group discussion sessions can be estimated by low-level verbal and nonverbal features at high accuracy [20]. However, the results of this work is based on the data of the whole experiment session which lasts for 15 min. In this setting, the evaluation results for individual participants cannot be obtained until the end of the session. Therefore, if this result was used to build a training system, the evaluation can only be provided to the participant off-line. The users need to recall what happened in the whole session and could be hard to identify where and how to improve their behaviors. Furthermore, it is also difficult to generate practical and comprehensive advices directly from low-level signals. For an on-line or even a real-time system, that is, a system which provides advices to its user while the group discussion is in progress is more desired. The user then can immediately notice what was wrong after the user did or said something and is easier to improve that behavior.

In order to develop such an on-line training system for group discussion, a conceptual unit of a batch of behaviors is required. It should be able to be treated as a template of communication style, appropriate for generating advices that is comprehensive enough for the user to capture the idea and improve their behaviors right away. This unit also need to have relatively small size so that the system can provide advices at fine granularity. We propose the use of *functional roles* for such purpose. We define a functional role of a participant of group discussion as: the role played by a participant at certain moment and displays the participant's contribution to the flow of the discussion. We believe that they can be used to analyze the dynamics of the interaction among the participants. By monitoring the dynamics of functional role transitions, e.g. the system can encourage some participant to show more opinions, if the system found that he/she is a passive participant. The system can suppress certain subject if it found that he/she took too much time for insisting on his/her own opinion. The first step for building such a system is automatic detection of the participants' functional roles during runtime. We previously proposed a support vector machine (SVM) based model for this purpose. This model classifies a participant's role from six roles after each utterance with hand-crafted features extracted from himself/herself [12]. This model achieved overall F-measure value, 0.32 with both verbal and non-verbal features or 0.28 with only non-verbal features in leave-one-subject-out cross validation.

On the other hand, the observation of someone's role in a group is supposed not to only depend on his/her own behavior but also how this person is interacting with other participants. It could be useful to utilize the character-

istics extracted from the interaction among the participants. Since such group dynamics can be so sophisticated that cannot be described sufficient by heuristics, instead of hand-crafted features based on heuristics which are required in SVM, we expect neural network models to be able to extract the characteristics of interaction automatically and improve the accuracy of the automatic detection. In the multimodal learning task involving multiparty interaction, the data streams of multiple modalities from multiple people can potentially contribute to the classification. During the design of neural network models, the order of the following processes can be considered:

- Fusion of the data streams from each participant
- Fusion of the data streams of modalities from one participant
- Extraction of the temporal characteristics of one modality

Different orders treat the raw data streams unequally, and therefore emphasize different parts of multiparty interaction, e.g. the overall multimodal performance of one person or the interaction of one modality among all of the participants. This then lead to different overall performance of the networks. In this work, we investigated the performance of three recurrent neural network (RNN) architectures which feature different orders in integrating these processes and compared them with baseline models: previously proposed SVM model and classic multi-layer perceptron (MLP) network where the inter-person features and frame-level modality interaction features are not available.

This paper is organized as the follows: Sect. 2 introduces related works. After the introduction of the dataset used and the definition of functional roles in Sect. 3, Sect. 4 describes the models compared in this paper including the baseline models, SVM and MLP as well as three RNN based models. Section 5 presents and analyzes the experiment results using the models described in previous session. Finally, Sect. 6 concludes this paper.

## 2 Related Works

Researchers in organizational psychology have studied the communication skill or the individual personality in group meetings for decades, uncovering statistical relationships between nonverbal behaviors, personality, hireability, and professional performance. It has been reported in social science that the non-verbal behaviors like gaze or gestures during the interaction with the others could have large influence on the flow of conversation [2, 6, 9, 15, 17]. In the context of group meeting, based on the nonverbal features, including features like speaking turn, voice prosody, visual activity, and visual focus of attention feature and so on. Aran and Gatica-Perez [1] presented an analysis on the participants' personality prediction in small groups. Similar to our goal, Schiavo et al. [22] presented a system that monitors the group members' non-verbal behaviors and acts for an automatic facilitator. It supports the flow of communication in a group conversation activity.

Furthermore, job interviews also have been studied in the research field of multimodal interaction, too. Raducanu et al. [21] made use of a TV show, “The Apprentice”, which features a competition for a real, highly paid corporate jobs. The study was carried out using non-verbal audio cues to predict the person with highest status and to predict the candidates going to be fired. Muralidhar et al. [18] implemented a behavioral training framework for students with the goal of improving the perception their hospitality perceived by others. They also evaluated the relationship between automatically extracted nonverbal cues and various social signals in a correlation analysis.

These studies show that prediction models could be achieved by using verbal and nonverbal multimodal features like speech turn, voice prosody, visual activity, visual focus of attention and so on. In our study, we aim to develop a training system that can provide advices to its users in improving the perception of their communication skill during group discussion. From previous studies, we expect that multimodal features can also be used to detect the functional roles defined by us.

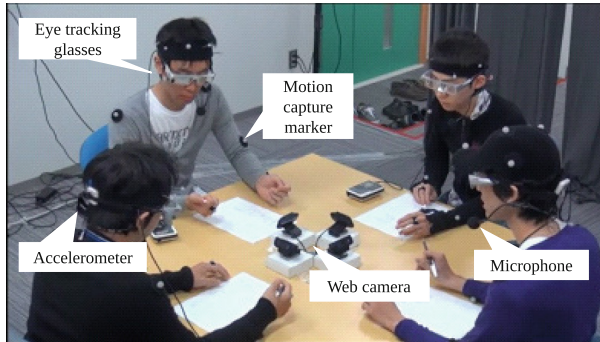
Zancanaro et al. [24] also proposed an automatic detection model using low-level video/audio features for functional roles. Their support vector machine (SVM) model classify two categories of functional roles, task area and socio-emotional area, each one has five classes. The performance was from 0.52 to 0.55 in the sense of F-score. Different to their general-purpose definition of functional roles, our definition has the specific purpose, the improvement on perception of communication skill in mind. In order to be able to feedback on the performance of the participants timely, we analyzed the relationships with the perception of communication skill and conversational situations in shorter interval which was not addressed by previous studies.

### 3 Data Corpus and the Definition of Functional Roles

In order to analyze the relationships between the functional roles and the perception of communication skill, it is necessary to conduct participant experiments to record group discussion sessions as the corpus. Also, the participants’ communication skill needed to be assessed by the experts who are familiar with this.

#### 3.1 Data Recording Experiment

This work is based on the MATRICS data corpus [19]. It is a multimodal corpus in which groups of four people discussed three different topics in two typical styles, make a choice from a list or make up something from the scratch. These topics were chosen in considering the ones which should be easier for Japanese college students to discuss: the selection of entertainers to be invited to university festival from a list, booth planning for school festival from the scratch, and the conduction of a travel planning for a foreigner friend who is going to visit Japan. 40 college students were recruited for the recording experiment. All of them



**Fig. 1.** Setup of the recording environment

were native Japanese speakers. They were divided into 10 groups, each one had four people, each group made three discussion sessions on the topics described above, each session lasts from 15 to 20 min. In order to prevent gender bias, all members in a group were in the same gender, or in equal number of two genders. In order to simulate the situation in the group discussion sessions of recruitment, the combination of the participants was considered so that they did not know each other before the experiment. Two video cameras as well as a number of sensors were used to record all the discussion sessions. Each participant wears a headset microphone, eye tracking glasses, and an accelerometer on their head. Everyone of them had a dedicated Web camera to capture his or her face in large size. Motion capture and Microsoft Kinect sensors were used to record the upper body movements of the participants as well. The setup of the recording experiment is shown in Fig. 1.

### 3.2 Functional Roles

In terms of functional roles, previous studies [4, 10] indicated that the participants perform one of three types of roles: group-task role, maintenance role, and individual role in group discussion. By referring previous works and the observation on our own data corpus, we defined our own set of functional roles. It includes six types of functional roles: *follower*, *gatekeeper*, *information giver*, *opinion provider*, *passive participant*, and *summarizer*. Table 1 lists the descriptions of all these six types.

Three coders watched the videos and annotated the roles of participants according to the definition above. Each coder was assigned with the data of four groups and used the software tool, Elan [16] for the annotation task. Among them, one was randomly selected for the measurement of inter-coder reliability of the coding, and every coder annotated that group. The pair-wise Kappa Coefficients among the three coders were 0.41, 0.51, and 0.63.

Three coders watched the videos and annotated the roles of participants according to the definition above. Each coder was assigned with the data of four

groups and used the software tool, Elan [16] for the annotation task. Among them, one was randomly selected for the measurement of inter-coder reliability of the coding, and every coder annotated that group. The pair-wise Kappa Coefficients among the three coders were 0.41, 0.51, and 0.63. Among the three coders, two of them had more consistent annotation but the other one was less consistent to them. After removing the sessions with incomplete data caused by recording problems, we therefore separate the dataset to two subsets, one contains the annotation of the two more reliable coders (9 sessions, 5 individual groups, and 20 individual subjects) and another one contains the annotations from all three coders (13 sessions, 8 individual groups, and 32 individual subjects) out of the all 120 available data due to the absence of face direction information (described hereafter).

## 4 Automatic Classification Models

As the usage scenario of the planned training system explained in Sect. 1, the system needs to always track the functional roles of all participants. That means, an automatic detection model using easily observable features (possible for the machines to process) and generates results in fine grain is required for this task. This section proposes such a classification model using low-level verbal and non-verbal features. The first issue of the detection model is, when should it generate the classification results, i.e. the *detection points*. In the previous study, Zancanaro et al. proposed a detection model using fix-length windows [24]. They explored the window sizes from 330 ms to 14 s and found that the performance of the model increases while the window size increases.

**Table 1.** Definitions of functional roles

| Role                | Definition   |
|---------------------|--|
| Follower            | Go along with the activity of the group, praise, agree with, and accept the contributions of others. Often look at the person who is speaking and nod or say back-channeling words |
| Gatekeeper          | Facilitate the flow of the discussion. Encourage the participants who are not so willing to join the discussion. Often look at other participants                                  |
| Information giver   | Provide objective information which is supplementary to the discussion   |
| Opinion provider    | State the participant's own opinion which might be subjective and try to convince others to agree with it  |
| Passive participant | Does not join the discussion actively. Almost does not provide the participant's own opinion in the topic being discussed. Often stay silent and look downward to the table        |
| Summarizer          | Summarize or conclude the discussion in the current topic  |

**Table 2.** Distribution of the detection points of the functional roles

| Dataset                    | 9 sessions |            | 13 sessions |            |
|----------------------------|------------|------------|-------------|------------|
| Role                       | Instances  | Percentage | Instances   | Percentage |
| <i>Follower</i>            | 3,202      | 47.8%      | 5,219       | 50.1%      |
| <i>Gatekeeper</i>          | 462        | 6.9%       | 539         | 5.2%       |
| <i>Information giver</i>   | 200        | 3.0%       | 288         | 2.8%       |
| <i>Opinion provider</i>    | 1,169      | 17.4%      | 1,958       | 18.8%      |
| <i>Passive participant</i> | 1,407      | 21.0%      | 2,081       | 20.0%      |
| <i>Summarizer</i>          | 265        | 4.0%       | 341         | 3.3%       |
| <b>Total</b>               | 6,705      | 100%       | 10,426      | 100%       |

In our work, however, we have a determined target application, integrating the classification model into the training system. The question then becomes, when should the training system provides advices to its users. Since it is often considered impolite to interrupt people while they are speaking, we assumed the appropriate timings should be the end of certain participant's utterance. Table 2 shows the distribution of functional roles regarding to the candidates of detection points.

#### 4.1 Baseline Models

In order to clarify the performance improvements from introducing frame-wise and inter-person features. Two baseline models based support vector machine (SVM) and classic multi-layer perceptron (MLP) on are also investigated. For these two models, frame-wise information cannot be automatically extracted from the raw data, but a vector of feature values has to be deliberated for one data instance based on various heuristics, i.e. feature engineering. We selected the following 15 features for the preliminary classification model. All of the feature values are extracted from one participant and are used to classify that participant's functional roles immediately after each of his/her utterances.

- Prosodic features. Phonetic analysis tool, Praat<sup>1</sup> was used to compute the following prosodic features of the current utterance. Since prosody characteristics are person dependent, these values are normalized by each participant.
  - Maximum pitch
  - Minimum pitch
  - Difference between maximum pitch and minimum pitch
  - Average of pitch
  - Maximum intensity
  - Minimum intensity
  - Difference between maximum intensity and minimum intensity

<sup>1</sup> <http://www.fon.hum.uva.nl/praat/>.



- Average of intensity
- Speech turn features.
  - Average utterance length up to now
  - Standard deviation of utterance length up to now
  - Ratio of speaking period up to now
  - Ratio of silent period up to now
- Face direction features. Because the participants wore eye tracking glasses during the experiment, the face directions of them was difficult for automatic face recognition tools. Therefore the face directions were manually labeled and the following features were used.
  - Time ratio when the participant was looking other participants in last segment (either speaking or not)
  - Time ratio when the participant was looking at the table in last segment (either speaking or not)

Due to the fact that the dataset was imbalanced, Follower and Opinion Provider have larger amount of instances than the other classes. We oversampled smaller classes with SMOTE [5] algorithm and under-sampled the larger classes while keeping the total weight (amount) of the dataset both in training and testing phases.

The SVM classification model was then built with the features above. During the training phase, RBF kernel was adopted while the cost parameter  $C$  was explored from 1 to 10 at the step size 1, kernel parameter  $\gamma$  was explored from  $10^{-3}$  to  $10^3$  at the step size  $10^1$ , feature normalization and standardization trials were conducted. The MLP based model shared the same dataset with SVM model. This network has simple structure where each modality (P, S, and F) are fed into the network in the cluster of neurons with equal number (32), concatenated, then fed to two more fully connected layers (128 and 64 neurons), and finally to the output layer. Dropouts are conducted between these layers in the rate, 0.3 to improve the generalization performance of the network. ReLU [3] is used as the activation function of hidden layers and Softmax [8] is used as the activation function of the output layer.

## 4.2 RNN Based Recurrent Neural Network Models

Unlike the based models where the characteristics of a single participant and the interaction among the participants need to be extracted according to heuristics and projected to a vector, RNN allows the characteristics in temporal direction to be derived from the raw data. The dataset prepared for RNNs can be raw values in frame level. The video taken by the two video cameras were used for the extraction of multimodal features. Since the objective is the generation of listening behaviors while the speaker is talking, it is necessary to identify the time periods when the speaker is speaking. The behaviors of the speaker in those periods are then extracted as explanatory variables, and the behaviors of the

listener in corresponding periods are extracted as response variables. Speakers' speech activities were automatically identified by the annotation software, ELAN [16] with additional manual corrections. Since the voice from other subjects may be recorded in one subject's headset, the speaker of each utterance is manually labeled.

All participants' facial expressions are extracted with an open source tool, OpenFace<sup>2</sup>. OpenFace estimates head postures (only rotations are used in this work, three values), gaze directions (eight values), and facial action units (AU, 17 out of 46 in the original definition) in accordance with Ekman's Facial Action Coding System (FACS) [7]. This resulted into 28 parameters from OpenFace in total. The posture information were extracted by using the open source tool, OpenPose<sup>3</sup>. Since OpenPose only generates two-dimensional coordinates of the joints of human bodies, posture information (leaning in two axes, forward/backward and left/right) are approximated with the assumption that the widths of the participants' shoulder are the minimum values when they are sitting straight up (i.e. they only lean forward but backward). Prosodic features of speakers' voice were extracted by using the open source tool, OpenSMILE<sup>4</sup>. The 16 low-level descriptors (LLD) of the Interspeech 2009 Emotion feature set [23] were extracted at 100 Hz. The features include root-mean-square of signal frame energy, zero-crossing rate of time signal, voicing probability, F0, and MFCC (12 dimensions). This resulted into 16 parameters from OpenSmile in total. The features like voice property are person dependent, relative changes are considered better describing the internal state of the speaking person rather than their absolute values. All of the feature values are standardized to fulfill the condition where mean is 0 and standard deviation is 1.0, and are then normalized to be within the range between 0.0 and 1.0.

We then investigated the performance of neural networks in three architectures where each one emphasize on one aspect of modality fusion. Emphasizing here means the execution of the process in the first stage of the processing flow. These networks share similar settings as the MLP base model described in last section in the aspects of the layers organization in later stages of processing flow, activation functions, and dropout ratio. Both video and audio input data are fed to the networks in 10-s fixed window at 30 fps.

**Network A:** This network emphasizes the interaction of the multimodal behaviors of one participant and then extract the temporal characteristics of it via LSTM (long short-term memory) [11] units. Finally the extracted characteristics of all participants are fused and the outputs are generated (Fig. 2).

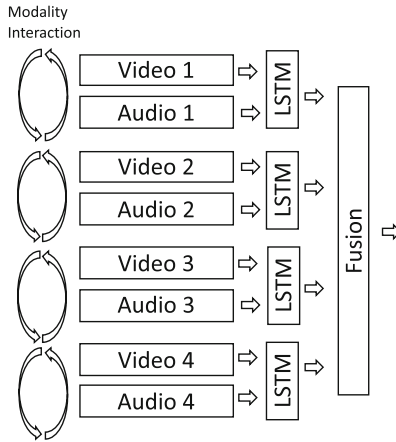
**Network B:** This network emphasizes the interaction among the participants in individual modalities. It extracts the characteristics of the interaction among the participants in each modality separately (video and audio) and then extracts the temporal characteristics of them via LSTM units. Finally all the modalities are fused and then the final results are generated (Fig. 3).

<sup>2</sup> <https://github.com/TadasBaltrusaitis/OpenFace>.

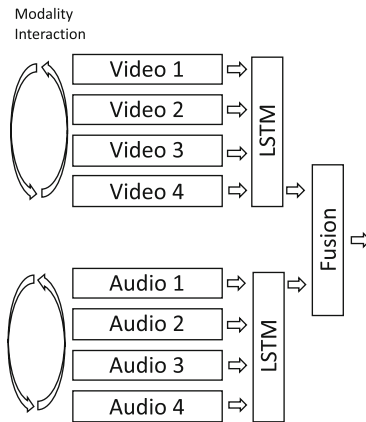
<sup>3</sup> <https://github.com/CMU-Perceptual-Computing-Lab/openpose>.

<sup>4</sup> <https://www.audeering.com/what-we-do/opensmile/>.

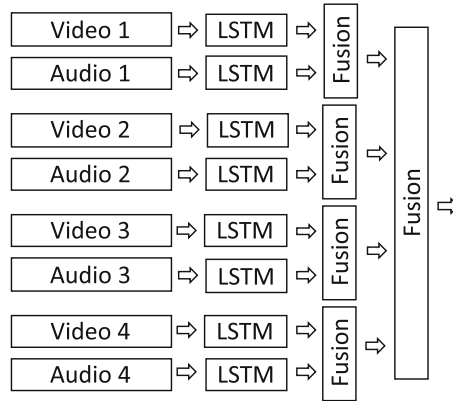
**Network C:** Instead of emphasizing inter-modality effects (Network A), this network extracts the temporal characteristics of each modality of one participants individually and fuse them later. Finally the streams from all participants are fused and the results are generated (Fig. 4).



**Fig. 2.** Conceptual diagram of neural network A. Only the order of feature extraction and fusion are shown, and layer details are omitted



**Fig. 3.** Conceptual diagram of neural network B. Only the order of feature extraction and fusion are shown, and layer details are omitted



**Fig. 4.** Conceptual diagram of neural network C. Only the order of feature extraction and fusion are shown, and layer details are omitted

## 5 Experiment Results

All models above are evaluated in leave-one-subject-out (LOSO) cross validation: the data of one subject is left as the test data while the data of the other  $N - 1$  subjects ( $N =$  the number of all subjects in the dataset) is used as the training set. The process is repeated for  $N$  times until the data of all subjects have been used as the test set. Then the overall results are computed. During the learning process, neural network models are trained for 300 epochs and batch size is 16. These values are found to be generating best results from preliminary experimental trials. We compared the variations (the inputs from one or all four participants) of these three networks on the 9-session and 13-session datasets as well as the base line models (SVM and MLP). The experiment results are sorted in Table 3 for 9-session data set and Table 4 for 13-session data set, respectively. Table 5 shows the confusion matrix of the best mode, LSTM network in architecture C with the features from all participants. The followings were found in the experiments results:

- The quantity of the data is more important than its quality (annotation reliability), 13-session dataset generally has better results.
- LSTM based neural networks performed better than simple MLP network (F-measure: 0.42), which is again better than SVM (F-measure: 0.28).
- The interaction of all four participants does contribute to the classification, the networks with four-person inputs outperform one-person networks.
- Network C (F-measure: 0.69) always performs better than network A and B (F-measure: 0.61). Network A and B have similar performance if the data from all four participants are available, but network A performs better than network B if the data from only one participant is available.
- For LSTM networks, Summarizer class always perform the best while Follower class always performs the worst. This may caused by how the features

distinguish the class from others and also from the number of data instances. Generally the classes with more instances have lower accuracy and vice versa.

To summarize, both frame-wise features and inter-person interaction features are shown to contribute to the classification task than their counterparts which do not include these features. LSTM mechanism does extract the temporal characteristics from the data in frame level in single modality and works better than hand-crafted features. But the frame-level interaction between modalities seems not useful. The inter-person interaction is useful when the features are extracted in later stage of the process, that is, after the features of each person has been explored. These results imply that the judgement of the participants' functional role more depends on overall impression rather than micro-level inter-modality or inter-person interactions.

**Table 3.** F1 values of the evaluated models with the 9-session dataset. LSTM\_X denotes the neural network based on LSTM in architecture X

| Role                       | Features from single participant only |      |        |        |        | Features from all participants |        |        |
|----------------------------|---------------------------------------|------|--------|--------|--------|--------------------------------|--------|--------|
|                            | SVM                                   | MLP  | LSTM_A | LSTM_B | LSTM_C | LSTM_A                         | LSTM_B | LSTM_C |
| <i>Follower</i>            | 0.27                                  | 0.24 | 0.43   | 0.31   | 0.44   | 0.44                           | 0.41   | 0.47   |
| <i>Gatekeeper</i>          | 0.22                                  | 0.47 | 0.55   | 0.50   | 0.56   | 0.56                           | 0.56   | 0.62   |
| <i>Information giver</i>   | 0.21                                  | 0.52 | 0.60   | 0.42   | 0.50   | 0.64                           | 0.50   | 0.52   |
| <i>Opinion provider</i>    | 0.28                                  | 0.34 | 0.50   | 0.56   | 0.62   | 0.45                           | 0.67   | 0.73   |
| <i>Passive participant</i> | 0.34                                  | 0.48 | 0.52   | 0.47   | 0.56   | 0.47                           | 0.58   | 0.64   |
| <i>Summarizer</i>          | 0.21                                  | 0.42 | 0.67   | 0.51   | 0.69   | 0.68                           | 0.61   | 0.71   |
| <b>Macro avg.</b>          | 0.24                                  | 0.41 | 0.55   | 0.46   | 0.56   | 0.54                           | 0.55   | 0.62   |

**Table 4.** F1 values of the evaluated models with the 13-session dataset. LSTM\_X denotes the neural network based on LSTM in architecture X

| Role                       | Features from single participant only |      |        |        |        | Features from all participants |        |        |
|----------------------------|---------------------------------------|------|--------|--------|--------|--------------------------------|--------|--------|
|                            | SVM                                   | MLP  | LSTM_A | LSTM_B | LSTM_C | LSTM_A                         | LSTM_B | LSTM_C |
| <i>Follower</i>            | 0.29                                  | 0.25 | 0.47   | 0.34   | 0.51   | 0.53                           | 0.43   | 0.52   |
| <i>Gatekeeper</i>          | 0.21                                  | 0.48 | 0.59   | 0.51   | 0.60   | 0.64                           | 0.60   | 0.67   |
| <i>Information giver</i>   | 0.31                                  | 0.55 | 0.70   | 0.45   | 0.52   | 0.73                           | 0.53   | 0.61   |
| <i>Opinion provider</i>    | 0.31                                  | 0.31 | 0.54   | 0.63   | 0.72   | 0.52                           | 0.75   | 0.79   |
| <i>Passive participant</i> | 0.34                                  | 0.48 | 0.58   | 0.55   | 0.64   | 0.53                           | 0.63   | 0.71   |
| <i>Summarizer</i>          | 0.22                                  | 0.44 | 0.71   | 0.64   | 0.77   | 0.76                           | 0.70   | 0.84   |
| <b>Macro avg.</b>          | 0.28                                  | 0.42 | 0.60   | 0.52   | 0.63   | 0.61                           | 0.61   | 0.69   |

**Table 5.** Confusion matrix of the best mode, LSTM network in architecture C with the features from all participants. Numbers depicts the ratio regarding to the number of instances. Rows are actual instances and columns are predicted classes

|                            | F           | G           | I           | O           | P           | S           |
|----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| <i>Follower</i>            | <b>0.45</b> | 0.29        | 0.15        | 0.03        | 0.05        | 0.02        |
| <i>Gatekeeper</i>          | 0.14        | <b>0.76</b> | 0.05        | 0.01        | 0.02        | 0.01        |
| <i>Information giver</i>   | 0.09        | 0.07        | <b>0.68</b> | 0.05        | 0.07        | 0.03        |
| <i>Opinion provider</i>    | 0.01        | 0.04        | 0.15        | <b>0.74</b> | 0.03        | 0.03        |
| <i>Passive participant</i> | 0.04        | 0.07        | 0.17        | 0.02        | <b>0.66</b> | 0.04        |
| <i>Summarizer</i>          | 0.01        | 0.03        | 0.07        | 0.02        | 0.05        | <b>0.82</b> |

## 6 Conclusions and Future Directions

In the task to classify the functional roles of group discussion participants, we explored the use of LSTM based networks to see the effectiveness of data interaction among modalities and participants. The investigation compares two based models (SVM and MLP) with three different organizations of LSTM based networks. The LSTM variations include different order in utilizing the multimodal data streams from one participant and the trunks from all participants. From the results, LSTM shows its effectiveness in extracting temporal information in single modality than hand-crafted features (SVM and MLP). Also, data fusion in later stage of the process performs better than early ones.

The findings of this study may help to give the insights in analyzing group dynamics and also provide the hints in developing support system. In this particular paper, we only focused on non-verbal features, we would like to explore the effects of semantic features in the future. Furthermore, we divided the features to only two classes (video and audio) in this work, it would be interesting to see the effects of more detailed features like facial expression and gaze.

**Acknowledgements.** This research is partially supported by KAKENHI: Grant-in-Aid for Scientific Research (A), Grant No. 19H01120.

## References

1. Aran, O., Gatica-Perez, D.: One of a kind: inferring personality impressions in meetings. In: Proceedings of 15th ACM International Conference on Multimodal Interaction (ICMI 2013), Sydney, Australia, December 2013
2. Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press, Cambridge (1976)
3. Bengio, Y., Glorot, X., Bordes, A.: Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 15, pp. 315–323 (2011)

4. Benne, K.D., Sheats, P.: Functional roles of group members. *J. Soc. Issues* **4**(2), 41–49 (1948)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
6. Clark, H.H., Carlson, T.B.: Hearers and speech acts. *Language* **58**(2), 332–373 (1982)
7. Ekman, P., Friesen, W.V., Hager, J.C.: Facial Action Coding System (FACS). Website (2002). <http://www.face-and-emotion.com/dataface/facs/description.jsp>
8. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press, Cambridge (2016)
9. Hall, J.A., Coats, E.J., LeBeau, L.S.: Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. *Psychol. Bull.* **131**(6), 898–924 (2005)
10. Hare, P.: Types of roles in small groups, a bit history and a current perspective. *Small Group Res.* **25**(3), 433–448 (1994)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
12. Huang, H.H., Zhang, Q., Okada, S., Kuwabara, K., Nishida, T.: Adopting functional roles for improving participants' communication skill in group discussion conversation. In: *Workshop on Group Interaction Frontiers in Technology (GIFT 2018)*, 20th ACM International Conference on Multimodal Interaction (ICMI 2018), Boulder, USA, October 2018
13. Institute, H.R.: Report of the 2017 investigation on the trend in the recruitment of new graduates (2017). (in Japanese)
14. Keidanren Japan Business Federation: Reports of the 2017 investigation on the recruitment of new graduates, November 2017. (in Japanese)
15. Kendon, A.: Some functions of gaze direction in social interaction. *Acta Psychol.* **26**, 22–63 (1967)
16. Lausberg, H., Sloetjes, H.: Coding gestural behavior with the NEUROGES-ELAN system. *Behav. Res. Methods* **41**(3), 841–849 (2009)
17. McNeill, D.: *Hand and Mind*. The University of Chicago Press, Chicago (1992)
18. Muralidhar, S., Nguyen, L.S., Frauendorfer, D., Odobez, J.M., Mast, M.S., Gatica-Perez, D.: Training on the job: behavioral analysis of job interviews in hospitality. In: *18th ACM International Conference on Multimodal Interaction (ICMI 2016)*, Tokyo, Japan, pp. 84–91, November 2016
19. Nihei, F., Nakano, Y.I., Hayashi, Y., Huang, H.H., Okada, S.: Predicting influential statements in group discussions using speech and head motion information. In: *16th International Conference on Multimodal Interaction (ICMI 2014)*, Istanbul, pp. 136–143, November 2014
20. Okada, S., Nakano, Y., Hayashi, Y., Takase, Y., Nitta, K.: Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In: *18th ACM International Conference on Multimodal Interaction (ICMI 2016)*, Tokyo, pp. 169–176, November 2016
21. Raducanu, B., Vitria, J., Gatica-Perez, D.: You are fired! Nonverbal role analysis in competitive meetings. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, Taipei, Taiwan, April 2009
22. Schiavo, G., Cappelletti, A., Mencarini, E., Stock, O., Zancanaro, M.: Overt or subtle? Supporting group conversations with automatically targeted directives. In: *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI 2014)*, pp. 225–234 (2014)

23. Schuller, B., Steidl, S., Batliner, A.: The interspeech 2009 emotion challenge. In: 10th Annual Conference of the International Speech Communication Association (Interspeech 2009), Brighton, United Kingdom, September 2009
24. Zancanaro, M., Lepri, B., Pianesi, F.: Automatic detection of group functional roles in face to face interactions. In: Proceedings of the 8th International Conference on Multimodal Interfaces (ICMI 2006), pp. 28–34 (2006)





# Exploring Gaze Behaviour and Perceived Personality Traits

Koki Ijuin<sup>(✉)</sup> and Kristiina Jokinen

AI Research Center, National Institute of Advanced Industrial Science and Technology (AIST),  
2-3-26 Aomi, Koto-Ku, Tokyo, Japan  
{koki-ijuin, kristiina.jokinen}@aist.go.jp

**Abstract.** The paper discusses correlation between interlocutors' eye-gaze behavior and their perceived personality traits in human-human and human-robot interactions. Given that personality is related to the person's typical manners and styles of behaving, it can be assumed that such underlying characteristics are reflected in the person's gaze patterns as well. Starting from the comparison of human-human and human-robot interaction, the participant's gaze frequency and length in regard to the human vs. robot partner's face and body are related to the participant's perceived personality traits. A positive correlation is found concerning the differences in gaze patterns and the extrovert personality trait. This seems highly reasonable, considering the basic function of gaze as a means to collect situational information and the extrovert communication style as actively looking for new information.

**Keywords:** Eye-gaze activity · Personality traits · Human-human interaction · Human-robot interaction

## 1 Introduction

In face-to-face situations, humans are sensitive to the other person's gaze: gazing is used to construct shared knowledge, communicate experiences, and create social linkages [1–3]. Also turn-taking is commonly coordinated by gaze [4, 5]. In virtual human and human-robot interactions gaze has been widely studied (see e.g. Broz et al. [6] for an overview), and our earlier work suggests that gaze plays an important role in grounding and it supports the view that robot agents are regarded as communicative agents rather than simply interactive tools [7, 8].

Starting from the comparison of gaze patterns in human-human and human-robot interactions, the paper explores correlation between interlocutors' eye-gaze behavior and their perceived personality traits. Gaze indicates where the person's visual focus is directed to and consequently, which actions and objects in the interaction context are considered important and relevant for further processing and for building common ground between the interlocutors. Since different participants have different overt reactions in communicative situations, different personality traits have been proposed to describe, study and classify typical human behavior and communication styles. It is

further assumed that the differences in gaze-patterns relate to the person's interaction behavior, and moreover, that the differences may be pinned down to the characteristic personality traits. Given that the main function of gaze is to provide information from the environment to the participants, it is natural to assume that different eye-gaze patterns are linked to the different ways to collect information and consequently, related to behavior patterns which count as typical personalities. For instance, the person's openness and curiosity towards the interaction partner can correlate with their high and intensive eye-gaze activity focused on the partner so as to observe and absorb as much information from the environment and the partner as possible.

The goal of our research is to study human-robot interaction and to improve the robot's engagement and interaction capabilities by building models for natural interaction. In this paper we focus on gaze patterns and personality. The paper first describes the "Big-5" personality traits and gives a theoretical starting point for the work in Sect. 2, and presents the AICO-corpus, used as the basis for our experimental studies, is described in Sect. 3. The preliminary analyses are given in Sect. 4 and discussion provided in Sect. 5.

## 2 The Big-5 Model of Personality Traits

The common personality traits described in the Five-Factor Model of Personality, the so-called Big-5 Model [9], are broad dimensions of human personality. The Big-5 personality traits are: extroversion (being assertive, energetic, friendly), agreeableness (being cooperative and trustworthy), conscientiousness (being self-disciplined, organised and reliable), neuroticism (having a tendency to negative emotions, being anxious), and openness to experience (being adventurous and easily embracing new ideas). Although they have been subject to much discussion and critique, they seem to convey some general features of the human behaviour which also seem to hold across cultures [9]. In this paper we do not go into discussion of the validity of the Big-5 personality traits but use them as a descriptive tool to analyse and understand differences in the participants' gaze behaviour with respect to different communicative situations.

In this paper, we are interested in the differences in human gaze patterns when the person interacts with a fellow human and with a robot agent. Assumption is that the person's underlying personality is the same across the different communicative situations, so the differences in the gaze patterns can be correlated with the type of dialogue partner and the person's typical manner of getting visual information about their partner in order to establish conversational common ground. The work builds on studies such as Aran et al. [10], Jayagopi et al. [11] and Okada [12] who studied first impression of the personality using multimodal cues. We extend their work with the focus on adapting real-time robot behaviour with respect to the personality of the person interacting with the robot [13, 14].

Nonverbal behaviours such as head pose, gaze, facial expression and body language are fundamental cues used to predict the personality of people interacting with each other [15, 16]. Through the observation of facial postures and eye contact durations, as well as through the frequency of the head movements, hand gestures, their amplitudes, and the shifts in the body postures, it is possible to infer certain traits of others' personalities [17, 18]. Computer vision and deep learning have been started to be extensively used for

video analysis and classification [19, 20] and, within the ChaLearn Looking at People Apparent Personality Analysis competition [21], have been proven effective also in the field of personality recognition, take advantage of the presence of audio and visual information.

### 3 AICO Corpus

We use the AICO corpus [22] to investigate whether it is possible to correctly infer personality traits from human-human interactions and what are the differences compared to inferring the same person’s personality in a human-robot interaction setting. The corpus consists of 30 participants (20 Japanese, 10 English), each having both HHI (human-human) and HRI (human-robot) interaction. The participants (10 female) were students and researchers, age 20–60, with experience on IT, but no experience on robots. The corpus was collected in Japanese and English depending on the participant’s preferred language. Before the experiments, the participants signed a consent form and filled in a pre-experiment questionnaire of their background and expectations. After each interaction (HRI and HHI), they filled in another seven-point scale questionnaire focusing on their experience in the interaction.

The instructions were the same for both HRI and HHI conditions (see the setup in Fig. 1). Of the participants, 14 had instruction dialogues concerning best practices in care-giving situation, and 16 had chat dialogues mostly focusing on music, films and every-day life. The corpus is described in Jokinen [22].

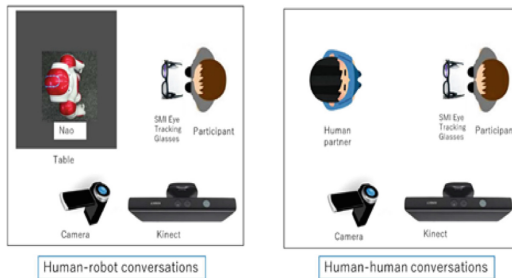
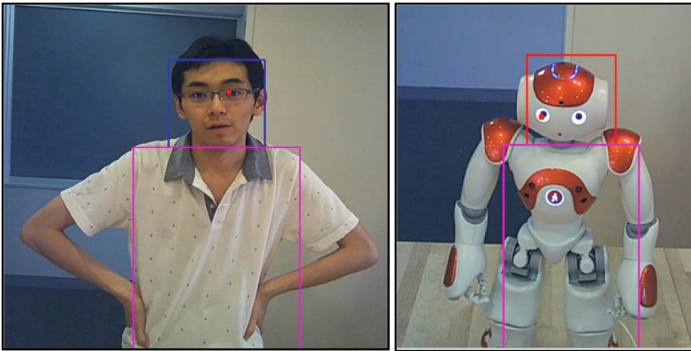


Fig. 1. Experimental setup (from [8]).

Annotation for duration of utterances were done with automatic silence segmentation of ELAN. The eye gaze activities were automatically annotated into two groups; Gaze Face and Gaze Body [8]. Figure 2 shows the snapshots from participant’s eye gaze tracker, and the automatic face detection for annotation. The gesture annotation followed the MUMIN annotation scheme and is described in Mori et al. [23].

For the personality annotation, each personality traits was regarded as a scale which span between two extremes, e.g. extroversion was judged on the positive end of the scale as “extroverted, enthusiastic” and on the negative end of the scale as “reserved, quiet”. The related questions were formulated to address both ends of the scale, and the other end of each scale functioned as a negative control question, with the assumption that

consistent annotators would rank these questions with a near-opposite score related to the other end. Its value was changed to the corresponding positive one in the analysis.



**Fig. 2.** Snapshot from eye-tracker. Note that the red points show the gaze point of the participant, and rectangles are drawn by the automatic annotation system we created. (Color figure online)

The perceived personality questionnaire was composed of 10 questions. The questions were implemented in the ELAN annotation tools as 10 tiers representing the Short Big-5 questionnaire [24]. The evaluators answered the questions for each sub-dialogue highlighted in the tier “Personality of”, corresponding to the left and right participant. The evaluator looked at the human participant’s behavior in the video related to the highlighted section, and marked the perceived personality using a 7-point Likert scale where 1 corresponds to “strongly disagree” and 7 to “strongly agree”. The questions were of the format:

*I believe participant .... to be:*

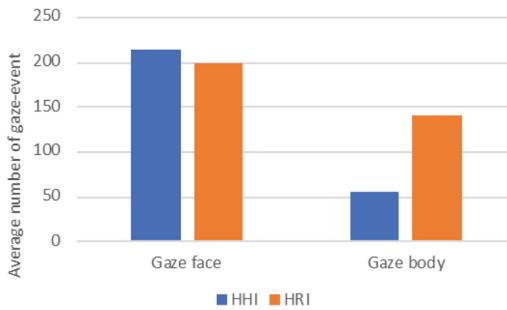
1. *Extraverted, enthusiastic*
2. *Critical, quarrelsome*
3. *Dependable, Self-disciplined*
4. *Anxious, easily upset*
5. *Open to new experiences, Complex*
6. *Reserved, quiet*
7. *Sympathetic, warm*
8. *Disorganized, careless*
9. *Calm, emotionally stable*
10. *Conventional, uncreative*

The questions 1 and 6, 2 and 7, 3 and 8, 4 and 9 and 5 and 10 are paired to form the personality scale. The instructions for the evaluators pointed out that they needed to answer the questions by taking into consideration the overall behaviour of both participants throughout the video.

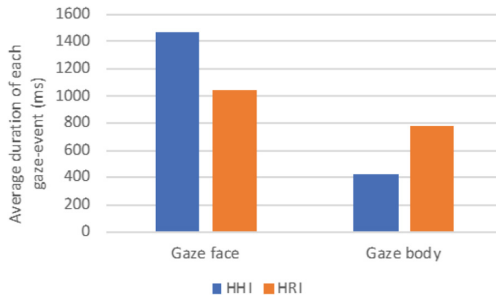
### 4 Preliminary Analyses

We have finished eye gaze annotation for 22 HHI and HRI conversations, so the preliminary analyses were conducted with those data. First, we calculated the number, total and average duration of gaze-event toward human and robot partner. Figure 3 and 4 show the average number and duration of gaze-event. We also calculated the average ratios of gaze-event duration during the conversations for HHI and HRI. The results are shown in Fig. 5. The paired-t tests were conducted in order to verify the difference of those ratios between HHI and HRI. The results show that there is a significant difference with gaze face event( $t_{(21)} = 2.14, p < .05$ ) and gaze body event( $t_{(21)} = -4.07, p < .01$ ), shown in Table 1.

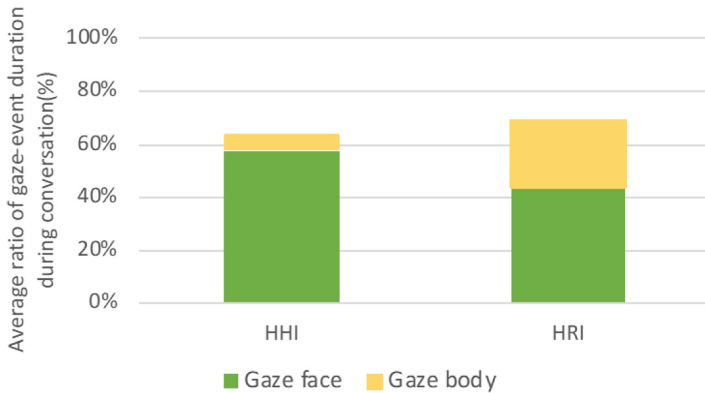
The results show that the participants gaze more at the human partner’s face in HHI than robot partner’s face in HRI, and they gaze more at the partner’s body in HRI than in HHI.



**Fig. 3.** Average Number of gaze-event, both gaze face and gaze body, toward partner for HHI and HRI.



**Fig. 4.** Average duration of gaze-event, both gaze face and gaze body, toward partner for HHI and HRI.



**Fig. 5.** Average ratios of gaze-event duration during conversations for HHI and HRI. Note that the rest of the ratio is that the participants do not gaze at the partner or their gaze could not be tracked by eye-tracker.

**Table 1.** Results of paired-*t* test for average ratios of gaze-event duration during conversations for HHI and HRI.

|           |            | <i>df</i> | <i>t</i> value | <i>p</i> value      |
|-----------|------------|-----------|----------------|---------------------|
| Gaze face | HHI vs HRI | 21        | 2.14           | 0.04 ( $p < .05$ )  |
| Gaze body | HHI vs HRI | 21        | -4.07          | 0.001 ( $p < .01$ ) |

In order to explain the difference why the participants gaze more at body in HRI than in HHI, a preliminary correlation analysis between gaze-events and perceived personality of the participant was conducted. As mentioned, the personality evaluation was done for 10 participants. We conducted Spearman's rank correlation test for those 10 participants data, and the results with significant difference are listed in Table 2. The results show that there are positive correlations between "emotion stability" and number of all gaze-event and gaze face event in HHI, and "openness to experience" and average duration of gaze body event in HRI. Negative correlation between "openness to experience" and number of gaze body event in HHI was also shown.

**Table 2.** Results of correlation tests between gaze-event and personality. Note that n/s stands for no significance.

|   | Emotion stability     | Openness to experience |
|---|-----------------------|------------------------|
| Number of all gaze-event in HHI ( $N = 10$ )            | $\rho = .69, p < .05$ | n/s                    |
| Number of gaze face event in HHI ( $N = 10$ )           | $\rho = .73, p < .05$ | n/s                    |
| Number of gaze body event in HHI ( $N = 10$ )           | n/s                   | $\rho = -.72, p < .05$ |
| Average duration of gaze body event in HRI ( $N = 10$ ) | n/s                   | $\rho = .69, p < .05$  |

## 5 Discussion

We conducted preliminary analyses of eye gaze activities during whole conversation of HHI and HRI. The results showed that:

1. the participants gaze more at the partner's body in HRI than in HHI,
2. the participants gaze more at the partner's face in HHI than in HRI,
3. positive correlation between frequency of gazing toward partner's face in HHI and emotional stability of the participant,
4. positive correlation between average duration of gazing toward partner's body in HRI and openness to experience of the participant.

Results 1 and 2 suggest that the participants gaze more at the partner's face in HHI than in HRI. The degree of change in robot's face is only blinking by changing colour of the eye, which does not provide any information. This might be the reason why the participants gaze less at the robot's face compared to human face. The result 2 also shows that the participants gaze longer at the robot's body than at the human partner's body. In HRI condition, the robot moves its arms and head randomly during its utterances in the conversation. This might be the reason that the participants gaze at the robot's body in order to understand the meaning of the gesture of the robot. Although, we need further analyses in order to confirm why this difference occurs.

To confirm why the eye gaze activities in HHI and HRI differs, we conducted the preliminary analyses considering participant's personality. The result 3 shows that the participants gaze at the human partner's face more often if their emotion was stable, although this tendency was not shown in HRI. This might be the reason that the social politeness affects in the HHI, whereas they do not need to show that in HRI.

The Result 4 shows that the more participants are open to experience, the more they gaze longer at the robot's body. This might reinforce the previous interpretation that the participants gaze at the robot's body in order to understand its random movement. The participants who is open mind to experience, which the interaction with the robot is not common, might have interests to the robot's gesture which seems it tries to tell some information.

These results suggest that there might be the difference of eye gaze activities between interaction with human and robot. However, some participants tried to understand the

robot's gesture which they might think that there is some reason or meaning like human do, where there is no meaning in this condition. This suggests that the gesture of the robot can compensate the participant's understanding if the robot acts properly to the context of the conversations. The results also suggest that the personality of the participants affects their eye gaze activities so that establishing the smooth interaction with the robot needs to consider the personality, at least from these preliminary analyses. These results even revealed with the perceived personality, which was evaluated for each conversation, so that the further analyses considering the changes of perceived personality and the context of the conversations is needed in order to verify and predict how the participants use their behavior during the conversations in those dynamic conditions.

## 6 Summary and Future Plan

We conducted the preliminary analyses of eye gaze activities in human-human and human-robot interaction. The results show that the eye gaze activities are different between interaction with human and robot, and the results also suggest that the general perceived personality of the participants by the evaluators affect to the fixation location of the gaze in both human-human interaction and human-robot interaction. These results suggest that changing the robot's behaviour for participant's personality might be the key to establish smooth interaction.

As we mentioned in this paper, this was the preliminary analyses. The further quantitative analyses of gesture, eye gaze activities and personality should be done in order to reinforce those possibilities. We are now annotating the detail gesture of the participants and human partners, and evaluating the participants' personality for each utterances in order to improve the accuracy of the analyses.

**Acknowledgement.** This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

## References

1. Kendon, A.: Some functions of gaze-direction in social interaction. *Acta Physiol.* **26**(1), 22–63 (1967)
2. Argyle, M., Cook, M.: *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge (1976)
3. Clark, H., Wilkes-Gibbs, D.: Referring as a collaborative process. *Cognition* **22**, 1–39 (1986)
4. Jokinen, K., Furukawa, H., Nishida, M., Yamamoto, S.: Gaze and turn-taking behavior in casual conversational interactions. *ACM Trans. Interact. Intell. Syst. (TiiS)* **3**(2), 12:1–12:30 (2013). Special Section on Eye-gaze and Conversational Engagement
5. Ijuin, K., Umata, I., Kato, T., Yamamoto, S.: Difference in eye gaze for floor apportionment in native-and second-language conversations. *J. Nonverbal Behav.* **42**(1), 113–128 (2018)
6. Broz, F., Lehmann, H., Mutlu, B., Nakano, Y.: *Gaze in Human-Robot Communication*. John Benjamins Publishing Company, Amsterdam (2015)
7. Jokinen, K.: Conversational gaze modelling in first encounter robot dialogues. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Languages Resources Association (ELRA) (2018)



8. Ijuin, K., Jokinen, K., Kato, T., Yamamoto, S.: Eye-gaze in social robot interactions – grounding of information and eye-gaze patterns. In: JSAI (2019)
9. McCrae, R.R.: Cross-cultural research on the five-factor model of personality. *Online Readings Psychol. Cult.* **4**(4), 1–12 (2002). <https://doi.org/10.9707/2307-0919.1038>
10. Aran, O., Gatica-Perez, D.: One of a kind: inferring personality impressions in meetings. In: *Proceedings of the ACM International Conference of Multimodal Interaction, ICMI 2013*, pp. 11–18 (2013)
11. Jayagopi, D.B., et al.: The vernissage corpus: a conversational human-robot-interaction dataset. In: *Proceedings of the 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 149–150. IEEE (2013)
12. Okada, S., Aran, O., Gatica-Perez, D.: Personality trait classification via co-occurrent multi-party multimodal event discovery. In: *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI 2015)*, Seattle, WA, USA, pp. 15–22 (2015)
13. Aly, A., Tapus, A.: A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. In: *Proceedings of the 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 325–332, March 2013
14. Jokinen, K., Wilcock, G.: Multimodal open-domain conversations with the Nao Robot. In: *Natural Interaction with Robots, Knowbots and Smartphones - Putting Spoken Dialog Systems into Practice*, pp. 213–224. Springer, New York (2014)
15. Koppensteiner, M.: Motion cues that make an impression: predicting perceived personality by minimal motion information. *J. Exp. Soc. Psychol.* **49**(6), 1137–1143 (2013). <http://www.sciencedirect.com/science/article/pii/S0022103113001467>
16. Subramanian, R., Yan, Y., Staiano, J., Lanz, O., Sebe, N.: On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In: *Proceedings of the 15th ACM on International Conference on Multimodal Interaction Association for Computing Machinery*, pp. 3–10 (2013). <https://doi.org/10.1145/2522848.2522862>
17. Larsen, R.J., Shackelford, T.K.: Gaze avoidance: personality and social judgments of people who avoid direct face-to-face contact. *Pers. Individ. Differ.* **21**(6), 907–917 (1996). <http://www.sciencedirect.com/science/article/pii/S0191886996001481>
18. Riggio, R.E., Friedman, H.S.: Impression formation: the role of expressive behavior. *J. Pers. Soc. Psychol.* **50**(2), 421–427 (1986)
19. Vinciarelli, A., Mohammadi, G.: A survey of personality computing. *IEEE Trans. Affect. Comput.* **5**(3), 273–291 (2014)
20. Mehta, Y., Majumder, N., Gelbukh, A., Cambria, E.: Recent trends in deep learning based personality detection. *Artif. Intell. Rev.* **53**(4), 2313–2339 (2019). <https://doi.org/10.1007/s10462-019-09770-z>
21. Ponce-López, V., et al.: ChaLearn LAP 2016: first round challenge on first impressions - dataset and results. In: Hua, G., Jégou, H. (eds.) *ECCV 2016. LNCS*, vol. 9915, pp. 400–418. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-49409-8\\_32](https://doi.org/10.1007/978-3-319-49409-8_32)
22. Jokinen, K.: The AICO multimodal corpus. In: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC-2020)*. European Languages Resources Association (ELRA) (2020)
23. Mori, T., Jokinen, K., Den, Y.: Analysis of body behaviours in human-human and human-robot interactions. In: *Proceedings of the LREC Workshop on People in Language, Vision and the Mind* (2020)
24. Rammstedt, B., John, O.P.: Measuring personality in one minute or less: a 10-item short version of the big five inventory in English and German. *J. Res. Pers.* **41**(1), 203–212 (2007). <http://www.sciencedirect.com/science/article/pii/S0092656606000195>



# Users of Fitbit Facebook Groups: A Gender- and Generation-Determined Investigation of Their Motivation and Need

Aylin Ilhan<sup>(✉)</sup> 

Department of Information Science, Heinrich Heine University, Düsseldorf, Germany  
aylin.ilhan@hhu.de

**Abstract.** This investigation focused on gender- and generation-determined differences regarding the need and use of Fitbit Facebook groups and the motivation to join these groups. Therefore, we applied the Uses and Gratifications Theory (U&GT) and Self-Determination Theory (SDT). This investigation aims to better understand the needs of activity tracking technology users who joined these groups. For this aim, we used an online survey. All in all, 268 participants are analyzed in this investigation. Results reveal that there are only a few gender- and generation-determined differences. This investigation draws on previous studies and allows to expand further research and to stress factors that needed to be considered.

**Keywords:** Fitbit Facebook groups · Uses and Gratifications Theory · Self-Determination Theory · Gender · Generations

## 1 Introduction

The importance of healthy living is a lifetime challenge. Being physically active can secure well-being, improve quality of life, and reduce the sedentary lifestyle. According to [1], “the failure to enjoy adequate levels of physical activity increases the risk of cancer, heart disease, stroke, and diabetes by 20–30% and shortens lifespan by 3–5 years.” For a few years, companies such as Fitbit, Samsung, Garmin, and Huawei regularly present new models of activity tracking technologies. These technologies enable users to easily monitor, analyze, and to use health-metrics such as steps, heart rate, burned calories, and sleep quality. According to [2], 30% of US citizens already use wristbands to track activities. These wristbands try to support users to be physically more active through setting step-goals, receiving reminders, or gamification elements (i.e., step challenges, achievements). Not only the promises activity tracking technologies reveal, but the increasing adaption and interest of these technologies might show the capability of these technologies. Studies already investigated the acceptance and usefulness of activity tracking technologies (i.e., [3, 5–7]). The activity tracking technology users surveyed by [7] agree on the usefulness and the impact of activity trackers.

Besides the possibility to track health and fitness data, the corresponding mobile applications of the activity tracking technologies such as Fitbit include the option of

forming and maintaining a community. There, users can share information, post pictures, achievements, or anything else on different topics. The offering of these communities within the application can be supportive of regulating one's objective. Within communities, users can encourage each other or share information and answer questions.

Spaces to seek, produce, and share information online as well as to discuss topics are widespread. From online forums to social networking sites (SNSs), there is a great variety. Facebook is one of the most popular SNSs with about 2.45 billion monthly active users [8] and is today a digital space to connect, share information, and to pass the time. Facebook enables to connect with Facebook users who are sharing diverse common interests (e.g., political, housekeeping (e.g., cooking), health, and fitness). Facebook offers more than 10 million Facebook groups [9]. Therefore, it comes as no real surprise that the initial search for Fitbit groups yields many Facebook groups (both private and public). About nine thousand Facebook users joined the private Facebook group *Fitbit Charge 2 Group*<sup>1</sup>, about three thousand the private Facebook group *Fitbit UK*<sup>2</sup>, and about one thousand the private *Facebook group Fitbit Charge 3 & Ionic Group*<sup>3</sup>. This raises the question why users of activity trackers are joining those online communities, since activity tracking technologies have all the functionalities to be aware of one's health and fitness and improvement of being physically more active? First insights already showed that the primary motivation to use fitness and health-related Facebook groups is to seek for information [4].

Indeed, Facebook can be a source for getting health-related information. The study of [10] showed that users of a Facebook group *Talking about Vaccines* are seeking information on this topic as well. At least the quality and usefulness of the received information are still individually evaluated and perceived. Regarding the thematic priority, Facebook groups are of interest to users with a common interest and need to be connected [11]. Even if [11] concentrated on Facebook groups thematizing schizophrenia, findings such as creating awareness, supporting users with this disorder could be general characteristics of health- and fitness-related Facebook groups. Similar characteristics could be stressed out in the study of [12]. Users of the investigated Facebook group shared clinical information, were seeking for guidance and feedback, and emotional support [10]. A broad thematic view showed that within Facebook groups, specific health information and experiences are shared and sought. Even though such Facebook groups seem to occupy an important role, the study of [4] highlights that users who are using fitness- and health-related Facebook groups would use their activity trackers also without using the groups. The use of this kind of Facebook groups is crystallized as a supportive opportunity to meet needs but not necessarily to continue the use of activity trackers [4].

To the best of our knowledge, there are several studies on Facebook groups regarding health-related topics (i.e., [10–12]), but the research on activity tracker-related groups [4] is still limited. Therefore, to expand the research and to better understand the benefits of these groups as perceived by the users, the following investigation analyses motivations to join Fitbit-related Facebook groups by applying two theories that have the potential to complement each other. Firstly, we applied the Uses and Gratifications Theory (U&GT).

<sup>1</sup> <https://www.facebook.com/groups/1322932601106678/>.

<sup>2</sup> <https://www.facebook.com/groups/663824983756452/>.

<sup>3</sup> <https://www.facebook.com/groups/fitbitstar/>.

The theory claims that people are using a specific media source out of specific needs (e.g., seek for information). The U&GT is an adequate approach to better understand the need for why users choose a particular media, but it is not telling which motivation underlies to join these groups. Therefore, secondly, we made use of the Self-Determination Theory (SDT) to understand not only the need of users but also their motivation to join these Fitbit-related Facebook groups as well. The core of the SDT is defined by different motivational driving forces (extrinsic and intrinsic), which leads people to do activities or decisions.

This investigation is a follow-up study and makes use of the collected data by [4]. In association with activity tracking technologies, this study concentrates on Fitbit-related Facebook groups. This investigation should create an added value in many ways. Firstly, it will offer insights if there are gender- and generation-determined differences regarding the need to use Fitbit-related Facebook groups. For this purpose, we apply the U&GT. Secondly, to complete the reasons why users are using these groups, we applied the SDT to allow insights if there are gender- and generation-determined differences as well.

First of all, the theoretical background will show related literature, as well as the applied theories U&GT and SDT followed by the research questions. Subsequently, the used methodology (online survey) and the preparation and applied statistical methods to answer the research questions are presented. The results will be reasonably processed to answer the research questions adequately. In the end, the results will be discussed in order to develop implications.

## 2 Theoretical Background

In this section, we introduce the core of the used theories for this investigation and related literature regarding gender- and generation-determined differences.

The **Uses and Gratifications Theory** enables an audience-centered investigation. This approach leads to understanding why users decided to use specific media concerning the needs they desire to satisfy. U&GT traces back to the time of traditional media channels, where the chosen media enables the seeking of different gratifications [13, 14]. Part of the U&GT is not only users' seeking behavior for gratifications but also the obtaining of them [15–17]. However, sought and obtained gratifications do not need to accord with each other [17].

Today, the approach is used in different contexts, such as activity tracking technologies in general [18], activity tracking technologies with regard to social media or mobile applications [4, 19], and diverse social media platforms [20–27]. There are different types of gratifications detected and investigated. Four common gratifications, based on [28] are *information*, *self-presentation*, *socialization*, and *entertainment*. According to these four gratifications, the study by [29] confirmed that they were sought while participating in Facebook groups as well. According to [4], activity tracking technology users mainly look for the possibility to receive information within fitness- and health-related Facebook groups. Nevertheless, apart from these four gratifications, [25] identified ten motives “social interaction, information seeking, pass time, entertainment, [...]” as well. According to [24], the surveyed participants use Facebook, among others, for socializing and communicating with their friends and obtaining information about social events. In

the end, gratifications needed to be adapted and modified related to their context and the investigated aim. Now, as we already know that there are several studies on the U&GT and social media, what about gender-determined differences? Is there any evidence that the sought and obtained gratifications differentiate regarding gender, especially in the context of physical activity and activity tracking technologies?

According to [30], gender-specific differences are apparent. While women of their study survey within the Facebook groups.

end more to use SNSs regarding socializing aspects and getting social information, men tend to use SNSs to seek general information [30]. By applying the U&GT, [19] investigated the gender-determined information sharing behavior of physical activity within mobile applications, and Facebook. Furthermore, they found out that men tend to share their results with others (achieved with the Runtastic mobile application) more often than women in Facebook groups. Interestingly, from the surveyed Runtastic users, 84.7% joined Facebook groups related to fitness-related Facebook groups (e.g., Runtastic) [19].

Apart from gender-determined differences, we are focusing on generation-determined differences as well. According to [31], different generations use social media platforms differently. Besides gender-specific differences, [19] investigated generation-determined differences as well. Older participants are willingly sharing more results in the fitness app Runtastic than younger participants [19]. For getting a comprehensive insight, it is necessary to expand the gender- and generation-determined investigations in this research area. While it is confirmed that some frequent gratifications appear (i.e., information, social aspects), it still remains uncertain if they might be affected by aspects such as gender and age. Especially since the “Facebook usage is significantly different between gender, with 63% of men using the site compared to 75% of women [...]” and “it wide disparity among age groups” [9]. Therefore, based on the U&GT and the rarely available investigations, we formulate the following research questions (RQ):

**RQ1:** To what extent do male and female activity tracker users differ regarding their sought and obtained gratifications within the Fitbit Facebook groups?

**RQ2:** To what extent do the generations differ regarding the sought and obtained gratification within the Fitbit Facebook groups?

The **Self-Determination Theory** states that motivation is not simply dichotomous (intrinsic or extrinsic) [32, 33]. Extrinsic motivation is defined based on the source and to what extent it is self-determined. Activities and decisions are distinguishable regarding their external and internal nature. The more external circumstances influence actions, the less self-determined people are [34]. According to [34, 35], extrinsic motivation can be divided into four subcategories regarding their regulation nature *external regulation*, *introjected regulation*, *identified regulation*, and *integrated regulation*. The external regulation is the lowest of all self-determined motivational actions. Someone is extrinsically motivated if s/he is doing something only to avoid punishments or to get rewards. These actions are leading back to external regulations. More self-determined actions within the extrinsic motivation are characterized through the identified and integrated regulation. Even though people identify external values with their values and recognize them as harmonious, it is not the activity itself that is perceived enjoyable and leads to specific

actions. This leads to intrinsic motivation. People are intrinsically motivated when they perceive the activities themselves enjoyable and exciting. Here, the activity itself is at the forefront. Actions based on intrinsic motivation are the strongest self-determined ones [34, 35]. In the end, people are also able to do something without any reason or intention. According to [34, 35], there is the talk of *amotivation*. Here, the regulation is called impersonal and is motivated by having no intention at all, incompetence, or lack of control [34, 35].

There are few studies which already investigated users' driving motivational force and willingness to share information about their physical activity on Facebook [36] or to join health- and fitness-related Facebook groups [4]. According to [36], the surveyed users of the mobile application Strava are rather intrinsically motivated to share their physical activity status on Facebook. Further, surveyed users by [4] joined mostly out of internal reasons, here because they were intrinsically motivated. There are also existing a few studies combining physical activity intervention with the use of Facebook [37]. [37] stresses that more research is needed to understand the effect of Facebook regarding physical activity intervention. They also mention "that the additional use of Facebook may not have increased the level of physical activity participation significantly" [37]. Another study [38] found out that the use of Facebook can have a positive impact on exercise motivation, even if using Facebook is related to external and introjected regulation. Apart from concentrating on the SDT and Facebook, the study of [18] investigated activity tracking technology users' motivation based on the SDT. The participants were both intrinsically and extrinsically (external and integrated regulation) motivated to use the activity tracking technologies [18]. [39] stresses out that it is also crucial to investigate the older populations regarding the SDT and their motivational regulations. Drawn on the studies that already did the first step into understanding the SDT and the use of Facebook, we would like to continue this research field by answering the following two research questions:

**RQ3:** To what extent does users' motivation to join Fitbit-related Facebook groups differ based on gender?

**RQ4:** To what extent does the generations' motivation to join Fitbit-related Facebook groups differ?

### 3 Methods

#### 3.1 Online Survey

The online survey was distributed from January 2018 to February 2018 on different health- and fitness-related Facebook groups. The online survey was created with a free online tool<sup>4</sup>. This investigation is using the dataset by [4] collected in 2018. As Fitbit is vigorously investigated comparing to the other activity tracking technology brands, we decided to restrict the sample. Investigations around Fitbit are mostly focusing on interventions, the accuracy, the feasibility, and the acceptance of them. Therefore, we concentrated only on Fitbit-related Facebook groups, where the survey back then was

<sup>4</sup> <https://www.esurveycreator.com>.

distributed. The administrators or moderators were asked through a direct Facebook message, if it is allowed to share the survey within the Facebook groups. All in all, the sample of 268 participants is connected to one of eight Fitbit-related Facebook groups selected for this investigation. The survey can be divided into three parts. Firstly (1), the survey contains questions regarding demographics (i.e., gender and age) and general questions to verify that the participants were a member of the surveyed Facebook group. Secondly (2), eight items were assigned to the U&GT to investigate the sought and obtained gratifications. The gratifications investigated for both sought and obtained are *information, self-presentation, socialization, and entertainment* [28]. Sought gratifications were formulated with “I use this Facebook group, because I’m looking for the possibility ...” while obtained gratifications were introduced with the sentence “The use of the Facebook group actually enables me...” Thirdly (3), the last section represents the statements (all in all seven statements) based on the SDT. All statements, according to Theory Sect. (2) and (3), are equipped with a 7-point Likert scale from 1–‘Strongly Disagree’ to 7–‘Strongly Agree.’ The survey was available both in German and English and was pretested by five persons.

### 3.2 Data Preparation and Statistical Analysis

The data was prepared and analyzed with the Syntax of IBM SPSS Statistics 26. The answer possibility *I don’t know* was coded as missing value; otherwise, statistical calculations are getting falsified. The data compiled through the Likert scale were handled as ordinal-scaled as some of the variables were not normally distributed. This was tested with the Shapiro-Wilk test. Because we handled our data as ordinal scaled, we used two nonparametric tests to investigate differences regarding the distributions.

To investigate if there are gender-determined differences (RQ1 & RQ3), we used the Mann-Whitney U test by using the new dialog fields instead of the legacy dialogs. The Mann-Whitney U test is a nonparametric rank-based test based on two groups on an independent variable, here gender (female, male) and a dependent ordinal-scaled variable such as statements with a Likert scale [40]. To investigate generation-determined differences (RQ2 & RQ4), we used the Kruskal-Wallis H test (new dialog fields within SPSS), because we have more than two groups of generations [41]. To determine generation-determined differences, the grouping of participants’ age, based on [31] (*Silver Surfers (older than 59 years old)*, *Generation X (between 40 and 59 years old)*, *Generation Y (between 24 and 39 years old)*, *Generation Z (younger than 24 years old)*), was needed.

## 4 Results

### 4.1 Gender-Determined Differences Regarding Their Sought and Obtained Gratifications Within the Fitbit Facebook Groups (RQ1)

Table 1 shows that there are no gender-determined differences regarding the distribution of the sought gratifications. Female and male participants are looking mainly for the possibility to receive information (Median equals 6 (*Agree*)) and are not looking for self-presentation (Median equals 2 (*Disagree*)). Further, regarding the sought gratification

*socialization*, the interquartile range (IQR) for the female participants is higher than for the male participants. This indicates that even if the median equals 4 (*Neither Agree nor Disagree*), female participants' answers are strongly scattered around this value.

**Table 1.** Gender-determined differences regarding sought gratifications, N = 267. Abbrev. Mean Rank (MR), Median (Mdn), Interquartile Range (IQR), Mann-Whitney U Test (U Test),  $p < .05^*$ ,  $p < .01^{**}$ ,  $p < .001^{***}$ .

| Sought gratifications | Gender | N   | Descriptive statistics |     |     | Rank-based nonparametric test            |
|-----------------------|--------|-----|------------------------|-----|-----|--|
|                       |        |     | MR                     | Mdn | IQR | U test                                   |
| Information           | Male   | 25  | 130.96                 | 6   | 2   | U = 2949.000, $z = -.112$ ,<br>$p > .05$ |
|                       | Female | 239 | 132.66                 | 6   | 2   |  |
| Self-presentation     | Male   | 25  | 132.32                 | 2   | 3.5 | U = 2983.000, $z = .131$ ,<br>$p > .05$  |
|                       | Female | 235 | 130.31                 | 2   | 3   |  |
| Socialization         | Male   | 25  | 142.06                 | 4   | 2   | U = 3226.500, $z = .703$ ,<br>$p > .05$  |
|                       | Female | 238 | 130.94                 | 4   | 4   |  |
| Entertainment         | Male   | 25  | 142.82                 | 4   | 2   | U = 3245.500, $z = .797$ ,<br>$p > .05$  |
|                       | Female | 237 | 130.31                 | 4   | 2   |  |

According to Appendix 1, the aggregated values of participants who disagree that they are looking for the possibility to socialize (Likert values (1)–(3)) are about 45% out of 238 female participants. The median for female and male participants regarding sought gratification *entertainment* equals 4. Nevertheless, if we add the values from (1) to (3) (see Appendix 1) for the sought gratification *entertainment*, the results show that female and male participants agree on any level that they are seeking for the possibility to be entertained within the Fitbit groups.

According to Table 2, this investigation shows no significant gender-determined differences regarding the distribution of the obtained gratifications, even if there are few differences recognizable. Female participants (81.6% out of 239 female participants) and male participants (72% out of 25 male participants) agree that using the Fitbit Facebook group enables them to receive information (see Appendix 2). As for male participants, the median for obtained gratification *self-presentation* equals 3 (*Somewhat Disagree*), and for female participants, the median equals 4 (*Neither Agree nor Disagree*).

According to the general distribution and the aggregating of values (see Appendix 2), female participants equally tend to disagree more (46.7% out of 201 female participants) and to agree a little less (42.3% out of 201 female participants). The answers of male participants are distributed differently since there are 56% out of 25 male participants fully disagreeing (aggregating all levels of disagreement).



**Table 2.** Gender-determined differences regarding obtained gratifications, N = 267. Abbrev. Mean Rank (MR), Median (Mdn), Interquartile Range (IQR), Mann-Whitney U Test (U Test),  $p < .05^*$ ,  $p < .01^{**}$ ,  $p < .001^{***}$ .

| Obtained gratifications |        |     | Descriptive statistics |     |     | Rank-based nonparametric test     |
|-------------------------|--------|-----|------------------------|-----|-----|-----------------------------------|
|                         | Gender | N   | MR                     | Mdn | IQR | U test                            |
| Information             | Male   | 25  | 121.20                 | 6   | 3   | U = 2705.000, z = -.810, p > .05  |
|                         | Female | 239 | 133.68                 | 6   | 2   |                                   |
| Self-presentation       | Male   | 25  | 99.90                  | 3   | 2.5 | U = 2172.500, z = -1.116, p > .05 |
|                         | Female | 201 | 115.19                 | 4   | 3.5 |                                   |
| Socialization           | Male   | 25  | 117.86                 | 5   | 2.5 | U = 2621.500, z = -.128, p > .05  |
|                         | Female | 213 | 119.69                 | 4   | 3   |                                   |
| Entertainment           | Male   | 25  | 113.02                 | 4   | 2   | U = 2500.500, z = -.989, p > .05  |
|                         | Female | 227 | 127.98                 | 5   | 3   |                                   |

**4.2 Generation-Determined Differences Regarding Their Sought and Obtained Gratifications Within the Fitbit Facebook Groups (RQ2)**

According to Table 3, the Kruskal-Wallis H test reports that agreeing on seeking for information (reason to join Fitbit Facebook groups) significantly differed between generations,  $H(3) = 8.555$ ,  $p = .036$ . Considering the sought gratifications of *self-presentation*, *socialization*, and *entertainment*, there are no significant generation-determined differences detected. Even though Silver Surfers tend to agree (Median equals 5 (*Somewhat Agree*)) more than the other generations that they are using the Fitbit Facebook groups because they are looking for the possibility to socialize (i.e., being motivated for challenges, and emotional reinforcement). As the participants already use the Fitbit Facebook groups, the question arises whether there are existing generation-determined differences regarding the obtained gratifications. According to Table 4, the Kruskal-Wallis H test reports that agreeing on obtaining information (during the use of Fitbit Facebook groups) significantly differed between generations,  $H(3) = 9.390$ ,  $p = .025$ .

**Table 3.** Generation-determined differences regarding sought gratifications, N = 268. Abbrev. Generation (Gen.), Silver Surfers (SS), Generation X (GX), Generation Y (GY), Generation Z (GZ), Mean Rank (MR), Median (Mdn), Interquartile Range (IQR), Kruskal-Wallis H Test (H Test),  $p < .05^*$ ,  $p < .01^{**}$ ,  $p < .001^{***}$ .

| Sought gratifications | Descriptive statistics |     |        |     |      | Rank-based nonparametric test |
|-----------------------|------------------------|-----|--------|-----|------|-------------------------------|
|                       | Gen.                   | N   | MR     | Mdn | IQR  |                               |
| Information           | SS                     | 12  | 158.17 | 7   | 1.75 | H(3) = 8.555,<br>p = .036*    |
|                       | GX                     | 105 | 143.78 | 7   | 2    |                               |
|                       | GY                     | 130 | 125.04 | 6   | 2    |                               |
|                       | GZ                     | 17  | 101.74 | 6   | 1.5  |                               |
| Self-presentation     | SS                     | 12  | 163.21 | 4   | 5.75 | H(3) = 3.008,<br>p > .05      |
|                       | GX                     | 101 | 125.48 | 2   | 3    |                               |
|                       | GY                     | 131 | 132.86 | 2   | 3    |                               |
|                       | GZ                     | 17  | 126.76 | 2   | 3    |                               |
| Socialization         | SS                     | 12  | 152.08 | 5   | 4.75 | H(3) = 3.616,<br>p > .05      |
|                       | GX                     | 104 | 140.24 | 4   | 4    |                               |
|                       | GY                     | 131 | 126.95 | 4   | 3    |                               |
|                       | GZ                     | 17  | 114.09 | 3   | 3    |                               |
| Entertainment         | SS                     | 12  | 135.17 | 4   | 3.75 | H(3) = .681,<br>p > .05       |
|                       | GX                     | 103 | 135.08 | 4   | 3    |                               |
|                       | GY                     | 131 | 130.89 | 4   | 2    |                               |
|                       | GZ                     | 17  | 119.71 | 4   | 3.5  |                               |

### 4.3 Gender-Determined Motivation to Join Fitbit-Related Facebook Groups (RQ3)

The need why users decided to use Fitbit Facebook groups enables the first insight, but what about the motivation, which leads to the decision to join these groups. All in all, Table 5 shows that the Mann-Whitney U test revealed one significant difference between female (Mean Rank = 131.16) and male participants (Mean Rank = 145.90) regarding the external regulations. Even though the median is equal for both, the mean rank for the male participants is higher. This shows that male participants tend to disagree a little bit less than female participants. According to Table 5, both female and male participants mainly agree that they joined the Fitbit Facebook group out of intrinsic motivation. These Fitbit Facebook groups are for both male and female participants not only as pastime, as both are mainly disagreeing that they joined this group because they

**Table 4.** Generation-determined differences regarding sought gratifications, N = 268. Abbrev. Generation (Gen.), Silver Surfers (SS), Generation X (GX), Generation Y (GY), Generation Z (GZ), Mean Rank (MR), Median (Mdn), Interquartile Range (IQR), Kruskal-Wallis H Test (H Test),  $p < .05^*$ ,  $p < .01^{**}$ ,  $p < .001^{***}$ .

| Obtained gratifications |      |     | Descriptive statistics |     |      | Rank-based nonparametric test |
|-------------------------|------|-----|------------------------|-----|------|-------------------------------|
|                         | Gen. | N   | MR                     | Mdn | IQR  | H test                        |
| Information             | SS   | 12  | 155.42                 | 6.5 | 2    | H(3) = 9.390,<br>p = .025*    |
|                         | GX   | 105 | 144.83                 | 6   | 2    |                               |
|                         | GY   | 131 | 119.19                 | 5   | 3    |                               |
|                         | GZ   | 17  | 150.50                 | 6   | 2    |                               |
| Self-presentation       | SS   | 11  | 120.14                 | 5   | 5    | H(3) = 1.314,<br>p > .05      |
|                         | GX   | 88  | 116.43                 | 4   | 4    |                               |
|                         | GY   | 113 | 109.77                 | 3   | 3    |                               |
|                         | GZ   | 15  | 127.10                 | 5   | 4    |                               |
| Socialization           | SS   | 11  | 126.32                 | 4   | 3    | H(3) = .676,<br>p > .05       |
|                         | GX   | 91  | 123.85                 | 5   | 3    |                               |
|                         | GY   | 121 | 116.67                 | 4   | 3    |                               |
|                         | GZ   | 16  | 118.94                 | 5   | 3.75 |                               |
| Entertainment           | SS   | 12  | 123.79                 | 4.5 | 2.5  | H(3) = 1.503,<br>p > .05      |
|                         | GX   | 98  | 133.97                 | 5   | 3    |                               |
|                         | GY   | 127 | 122.55                 | 4   | 3    |                               |
|                         | GZ   | 16  | 122.03                 | 4.5 | 3    |                               |

were bored. Regarding the determined extrinsic regulations, the more self-determined the regulations are (identified and integrated regulation), the more the median value increases.

#### 4.4 Generation-Determined Motivation to Join Fitbit-Related Facebook Groups (RQ4)

When analyzing the generation-determined differences, the Kruskal-Wallis H test reveals only generation-determined differences for the amotivation *Boredom*. According to Table 6, the Kruskal-Wallis H test reports that the amotivation (Boredom) (reason to join Fitbit Facebook groups) significantly differed between generations,  $H(3) = 10.542$ ,  $p = .014$ . The mean rank for Generation Y (139.10) is higher than for Generation Silver Surfer

**Table 5.** Gender-determined differences regarding the motivation, N = 267. Abbrev. Mean Rank (MR), Median (Mdn), Interquartile Range (IQR), Mann-Whitney U Test (U Test),  $p < .05^*$ ,  $p < .01^{**}$ ,  $p < .001^{***}$ .

| Self-determination                  |                |        | Descriptive statistic |        |     |      | Rank-based nonparametric test            |
|-------------------------------------|----------------|--------|-----------------------|--------|-----|------|--|
| Motivation                          | Regulation     | Gender | N                     | MR     | Mdn | IQR  | U Test                                   |
| Extrinsic                           | External       | Male   | 24                    | 145.90 | 1   | 0    | U = 3201.500,<br>z = 2.180,<br>p = .029* |
|                                     |                | Female | 240                   | 131.16 | 1   | 0    |  |
|                                     | Introjected    | Male   | 24                    | 147.00 | 1   | 0    | U = 3228.000,<br>z = 1.921,<br>p > .05   |
|                                     |                | Female | 240                   | 131.05 | 1   | 0    |  |
|                                     | Identified     | Male   | 24                    | 118.90 | 3   | 3.75 | U = 2553.500,<br>z = -.643,<br>p > .05   |
|                                     |                | Female | 231                   | 128.95 | 4   | 3    |  |
|                                     | Integrated     | Male   | 23                    | 108.39 | 4   | 4    | U = 2217.00,<br>z = -.476,<br>p > .05    |
|                                     |                | Female | 205                   | 115.19 | 4   | 3    |  |
| Intrinsic                           | Intrinsic      | Male   | 24                    | 113.08 | 5   | 3.75 | U = 2414.000,<br>z = -1.306,<br>p > .05  |
|                                     |                | Female | 239                   | 133.90 | 5   | 3    |  |
| Amotivation (“Boredom”)             | Non-regulation | Male   | 24                    | 126.88 | 1   | 2    | U = 2745.000,<br>z = .069,<br>p > .05    |
|                                     |                | Female | 227                   | 125.91 | 1   | 2    |  |
| Amotivation (“Just for heck of it”) | Non-regulation | Male   | 24                    | 127.71 | 4   | 3.5  | U = 2765.000,<br>z = -.161,<br>p > .05   |
|                                     |                | Female | 235                   | 130.23 | 3   | 4    |  |

(90.67). Even if both generations still tend to disagree that they joined the Fitbit Facebook groups out of boredom, Generation Y tends a little bit less to disagree overall. Likewise, this is similar to Generation X (Mean Rank = 129.36) and Generation Y (Mean Rank = 139.10). Nevertheless, all four generations mainly agree that they joined the Fitbit Facebook group because it was a self-determined decision and intrinsically regulated (Table 6).

## 5 Discussion

The investigation’s aim was the identification of gender- and generation-determined differences regarding the motivation and need to join and use Fitbit Facebook groups. In doing so, the investigation revealed, except for a few significant differences, that both female and male participants and the four investigated generations have similar needs

and motivational reasons. For this study, we used a dataset (based on an online survey) that was already collected in 2018.

**Table 6.** Generation-determined differences regarding the motivation, N = 268. Abbrev. Generation (Gen.), Silver Surfers (SS), Generation X (GX), Generation Y (GY), Generation Z (GZ), Mean Rank (MR), Median (Mdn), Interquartile Range (IQR), Kruskal-Wallis H Test (H Test),  $p < .05^*$ ,  $p < .01^{**}$ ,  $p < .001^{***}$ .

| Self-determination                     |                 |      |        | Descriptive statistics |      |                          | Rank-based nonparametric test |
|--|-----------------|------|--------|------------------------|------|--------------------------|-------------------------------|
| Motivation                             | Regulation      | Gen. | N      | MR                     | Mdn  | IQR                      | H test                        |
| Extrinsic                              | External        | SS   | 12     | 135.75                 | 1    | 0                        | H(3) = 1.290,<br>p > .05      |
|  |                 | GX   | 103    | 132.83                 | 1    | 0                        |                               |
|  |                 | GY   | 133    | 133.91                 | 1    | 0                        |                               |
|  |                 | GZ   | 17     | 125.00                 | 1    | 0                        |                               |
|  | Introjected     | SS   | 12     | 120.50                 | 1    | 0                        | H(3) = 1.499,<br>p > .05      |
|  |                 | GX   | 103    | 132.39                 | 1    | 0                        |                               |
|  |                 | GY   | 133    | 134.21                 | 1    | 0                        |                               |
|  |                 | GZ   | 17     | 136.03                 | 1    | 0                        |                               |
|  | Identified      | SS   | 12     | 125.33                 | 4    | 1.75                     | H(3) = 1.560,<br>p > .05      |
|  |                 | GX   | 101    | 122.21                 | 3    | 5                        |                               |
|  |                 | GY   | 126    | 132.21                 | 4    | 3                        |                               |
|  |                 | GZ   | 17     | 140.59                 | 4    | 3.5                      |                               |
| Integrated                             | SS              | 12   | 124.21 | 4                      | 2.75 | H(3) = 1.214,<br>p > .05 |                               |
|  | GX              | 93   | 109.49 | 3                      | 4    |                          |                               |
|  | GY              | 109  | 118.09 | 4                      | 3    |                          |                               |
|  | GZ              | 15   | 119.30 | 4                      | 2    |                          |                               |
| Intrinsic                              | Intrinsic       | SS   | 12     | 173.63                 | 6    | 2                        | H(3) = 5.059,<br>p > .05      |
|  |                 | GX   | 103    | 129.36                 | 5    | 3                        |                               |
|  |                 | GY   | 132    | 133.77                 | 5    | 3                        |                               |
|  |                 | GZ   | 17     | 112.65                 | 5    | 2                        |                               |
| Amotivation<br>("Boredom")             | Non-regulation  | SS   | 12     | 90.67                  | 1    | 0                        | H(3) = 10.542,<br>p = .014**  |
|  |                 | GX   | 100    | 117.07                 | 1    | 1                        |                               |
|  |                 | GY   | 123    | 139.10                 | 2    | 3                        |                               |
|  |                 | GZ   | 17     | 116.12                 | 1    | 1.5                      |                               |
| Amotivation ("Just for<br>heck of it") | Non- regulation | SS   | 12     | 120.83                 | 3.5  | 4                        | H(3) = .359,<br>p > .05       |
|  |                 | GX   | 102    | 129.95                 | 4    | 5                        |                               |
|  |                 | GY   | 128    | 131.67                 | 3    | 4                        |                               |
|  |                 | GZ   | 17     | 124.21                 | 2    | 5                        |                               |

Answering **RQ1** shows that both female and male participants are mainly looking for the possibility to seek for information. Here, there are no gender-determined differences. All in all, Facebook is recognized as a digital space for exchanging information on different topics. Individuals who joined Facebook and Facebook groups bring along different experiences, and varying amounts of knowledge, which might be enriching. For a better understanding of which kind of information female and male participants are looking for, further studies are needed. According to [24], participants obtained information about social events. Here, we do not know exactly what kind of information participants sought and obtained. For this approach, a content analysis of postings could be useful, as the content itself can be characterized. As this study investigated the gratification sought *information* very broadly, a subdivision in different types of information might show gender-determined differences and overall varying distribution regarding the agreement as well. An assumption for non-gender-determined differences could be the fact that the use of Fitbit Facebook groups for information is predefined by the groups itself. Some Facebook groups have descriptions, where the main aim might be the exchange of information. Therefore, users who joined this group might share a common need and interest.

Interestingly, according to [19], where men tend to share their results with others more than women in Facebook groups could not be confirmed by this study. On the contrary, both female and male participants mainly disagree (Median equals 2 (*Disagree*)) that they are looking for the possibility to show their success, aims, and obtained achievements. Here the question arises if there are other factors that could affect the need for self-presentation. For example, how long do the participants have their activity tracking device? If they had the wearable for a long-time, the need to share the results might decrease. Further, it also can depend on the reason, why users bought an activity tracking device. In the beginning, users might be enthusiastic and excited, for example, collecting and sharing badges. Further, the behavioral stage might play an important role, as well. For instance, if they are at the beginning of changing their behavior, to be physically more active, sharing their success might be more important than later. According to the obtained gratifications, there are no significant gender-determined differences.

The results for **RQ2** show that there are existing generation-determined differences regarding the sought gratification *information*. Interestingly, even if the sample size for the Silver Surfers is small, the mean rank for sought information is the highest. This could indicate that especially the older participants have an increased need to receive information in a straightforward setting. All they need is a Facebook account and to join a Facebook group. Further, according to [9], the elderly population is more and more joining Facebook. Also, if the generation-determined differences regarding socialization are not statically significant, it is still evident that the Silver Surfers tend to somewhat agree that they are seeking for the possibility to get social support such as an invitation for challenges or emotional reinforcement. As the sample size of the Silver Surfers is small, in-depth interviews might be reasonable to understand better why the Silver Surfer somewhat agree that they are looking for this possibility and also if it is easier to get it through the Facebook group. We assume that for the generation Silver Surfers,

it might be more challenging to connect with other Fitbit users as compared to younger participants who might be surrounded by more users in their everyday life. Therefore, those Fitbit Facebook groups might be a chance for the Silver Surfers to get social contacts supporting them in being physically more active and motivated. The generalization of this assumption needs further studies. Besides the sought gratifications, are there generation-determined differences regarding the obtained gratifications? There is one significant generation-determined difference in obtaining information. Interestingly, while Silver Surfers mainly somewhat agree that they are using the Fitbit Facebook group to socialize, they also primarily somewhat agree that the Fitbit groups indeed enable it.

RQ1 and RQ2 are focusing on the sought and obtained gratifications, the results of **RQ3** show to what extent the decision to join a Facebook was self-determined, and if there are gender-determined differences. Male and female participants significantly differ regarding external regulation. All in all, female and male participants have mainly joined the Fitbit Facebook group based on intrinsic regulation. Based on the results, we assume to have some bearing on the sought gratification information and the motivational driving forces. We know that the female and male participants sought mainly for information. This indicates they have a specific need and did not join the group because they were bored. Further, it could be assumed that the participants joined other Facebook groups as well and experienced information exchange. This could be a reason why they prefer to join Fitbit Facebook groups to receive information as well.

Last but not least, answering **RQ4** focused on generation-determined differences regarding the SDT. In contrast to the one gender-determined significant difference, here the Kruskal-Wallis H test revealed generation-differences based on the amotivation (Boredom). But, overall, participants did not join Fitbit Facebook groups out of boredom.

## 6 Conclusion

Based on answering the research questions (RQ1–RQ4), what are the take-home messages?

First of all, it should be mentioned that this study has some limitations. As this investigation focused on gender- and generation-determined differences, the sample size for male participants and as well as for Generation Z and Silver Surfer, is rather small. According to [9], the elderly population is joining Facebook more and more. Therefore, there is a further study needed. As it is a non-probabilistic distributed survey, the controlling of getting specific participants from each group is not possible, as we do not know how the distribution of the groups looks like. Further, as this study focused only on two aspects (gender and age), other factors might influence the results as well. How long do the participants have and use their wearable? Might this factor influence the need to self-present and socialize? Here it would be adequate to have a sample with newbies as well as users like in this one, who have their wearable up to 1–3 years or more.





## Appendix 2

| Obtained Gratifications | N      | Likert Scale<br>(1-Strongly Disagree – 7-Strongly Agree) |               |               |               |               |               |               |               |
|-------------------------|--------|--|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                         |        | 1  | 2             | 3             | 4             | 5             | 6             | 7             |               |
| Information             | Male   | 25   | 1<br>(4.0%)   | 2<br>(8.0%)   | 2<br>(8.0%)   | 2<br>(8.0%)   | 4<br>(16.0%)  | 6<br>(24.0%)  | 8<br>(32.0%)  |
|                         | Female | 239  | 4<br>(1.7%)   | 6<br>(2.5%)   | 12<br>(5.0%)  | 22<br>(9.2%)  | 61<br>(25.5%) | 41<br>(17.2%) | 93<br>(38.9%) |
|                         |        |  | 7<br>9.2%     | 3<br>3.8%     | 5<br>5.9%     | 2<br>2.3%     | 2<br>2.3%     | 81.6%         |               |
| Self-Presentation       | Male   | 25   | 4<br>(16.0%)  | 7<br>(28.0%)  | 3<br>(12.0%)  | 5<br>(20.0%)  | 2<br>(8.0%)   | 2<br>(8.0%)   | 2<br>(8.0%)   |
|                         | Female | 201  | 41<br>(20.4%) | 25<br>(12.4%) | 28<br>(13.9%) | 22<br>(10.9%) | 35<br>(17.4%) | 15<br>(7.5%)  | 35<br>(17.4%) |
|                         |        |  | 46.7%         | 28<br>32.8%   | 22<br>21.1%   | 10.9%         | 42.3%         |               |               |
| Socialization           | Male   | 25   | 3<br>(12.0%)  | 3<br>(12.0%)  | 2<br>(8.0%)   | 4<br>(16.0%)  | 8<br>(32.0%)  | 0<br>(0.0%)   | 5<br>(20.0%)  |
|                         | Female | 213  | 28<br>(13.1%) | 17<br>(8.0%)  | 25<br>(11.7%) | 45<br>(21.1%) | 25<br>(11.7%) | 33<br>(15.5%) | 40<br>(18.8%) |
|                         |        |  | 32%           | 16.0%         | 52%           | 46%           |               |               |               |
| Entertainment           | Male   | 25   | 2<br>(8.0%)   | 1<br>(4.0%)   | 4<br>(16.0%)  | 7<br>(28.0%)  | 7<br>(28.0%)  | 2<br>(8.0%)   | 2<br>(8.0%)   |
|                         | Female | 227  | 14<br>(6.2%)  | 20<br>(8.8%)  | 26<br>(11.5%) | 50<br>(22.0%) | 45<br>(19.8%) | 30<br>(13.2%) | 42<br>(18.5%) |
|                         |        |  | 26.5%         | 22.0%         | 51.5%         |               |               |               |               |

## Appendix 3

| Sought Gratifications | N         | Likert Scale<br>(1-Strongly Disagree – 7-Strongly Agree) |               |               |               |               |               |               |               |
|-----------------------|-----------|--|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                       |           | 1  | 2             | 3             | 4             | 5             | 6             | 7             |               |
| Information           | Silver S. | 12   | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)     | 2<br>(16.7%)  | 1<br>(8.3%)   | 1<br>(8.3%)   | 8<br>(66.7%)  |
|                       | Gen. X    | 105  | 2<br>(1.9%)   | 4<br>(3.8%)   | 2<br>(1.9%)   | 6<br>(5.7%)   | 15<br>(14.3%) | 21<br>(20.0%) | 55<br>(52.4%) |
|                       | Gen. Y    | 130  | 2<br>(1.5%)   | 0<br>(0%)     | 12<br>(9.2%)  | 11<br>(8.5%)  | 28<br>(21.5%) | 24<br>(18.5%) | 53<br>(40.8%) |
|                       | Gen. Z    | 17   | 0<br>(0%)     | 0<br>(0%)     | 1<br>(5.9%)   | 3<br>(17.6%)  | 4<br>(23.5%)  | 6<br>(35.3%)  | 3<br>(17.6%)  |
|                       |           |  | 5.9%          | 17.6%         | 76.4%         |               |               |               |               |
| Self-Presentation     | Silver S. | 12   | 3<br>(25.0%)  | 2<br>(16.7%)  | 0<br>(0%)     | 2<br>(16.7%)  | 0<br>(0%)     | 1<br>(8.3%)   | 4<br>(33.3%)  |
|                       | Gen. X    | 101  | 39<br>(38.6%) | 14<br>(13.9%) | 15<br>(14.9%) | 14<br>(13.9%) | 4<br>(4.0%)   | 5<br>(5.0%)   | 10<br>(9.9%)  |
|                       | Gen. Y    | 131  | 40<br>(30.5%) | 26<br>(19.8%) | 19<br>(14.5%) | 15<br>(11.5%) | 13<br>(9.9%)  | 10<br>(7.6%)  | 8<br>(6.1%)   |
|                       | Gen. Z    | 17   | 6<br>(35.3%)  | 3<br>(17.6%)  | 2<br>(11.8%)  | 3<br>(17.6%)  | 1<br>(5.9%)   | 1<br>(5.9%)   | 1<br>(5.9%)   |
|                       |           |  | 64.7%         | 13.9%         | 17.6%         | 17.7%         |               |               |               |
| Socialization         | Silver S. | 12   | 3<br>(25.0%)  | 0<br>(0%)     | 0<br>(0%)     | 1<br>(8.3%)   | 5<br>(41.7%)  | 0<br>(0%)     | 3<br>(25.0%)  |
|                       | Gen. X    | 104  | 19<br>(18.3%) | 8<br>(7.7%)   | 15<br>(14.4%) | 13<br>(12.5%) | 16<br>(15.4%) | 15<br>(14.4%) | 18<br>(17.3%) |
|                       | Gen. Y    | 131  | 21<br>(16.0%) | 20<br>(15.3%) | 18<br>(13.7%) | 28<br>(21.4%) | 13<br>(9.9%)  | 17<br>(13.0%) | 14<br>(10.7%) |
|                       | Gen. Z    | 17   | 3<br>(17.6%)  | 4<br>(23.5%)  | 4<br>(23.5%)  | 0<br>(0%)     | 3<br>(17.6%)  | 1<br>(5.9%)   | 2<br>(11.8%)  |
|                       |           |  | 64.6%         | 0%            | 35.3%         |               |               |               |               |
| Entertainment         | Silver S. | 12   | 2<br>(16.7%)  | 1<br>(8.3%)   | 1<br>(8.3%)   | 3<br>(25.0%)  | 0<br>(0%)     | 3<br>(25.0%)  | 2<br>(16.7%)  |
|                       | Gen. X    | 103  | 11<br>(10.7%) | 14<br>(13.6%) | 17<br>(16.5%) | 17<br>(16.5%) | 22<br>(21.4%) | 10<br>(9.7%)  | 18<br>(17.5%) |
|                       | Gen. Y    | 131  | 6<br>(4.6%)   | 13<br>(9.9%)  | 18<br>(13.7%) | 42<br>(32.1%) | 25<br>(19.1%) | 18<br>(13.7%) | 9<br>(6.9%)   |
|                       | Gen. Z    | 17   | 1<br>(5.9%)   | 5<br>(29.4%)  | 2<br>(11.8%)  | 1<br>(5.9%)   | 4<br>(23.5%)  | 3<br>(17.6%)  | 1<br>(5.9%)   |
|                       |           |  | 47.1%         | 5.9%          | 47%           |               |               |               |               |

# Appendix 4

| Obtained Gratifications |           | Likert Scale<br>(1-Strongly Disagree – 7-Strongly Agree) |               |               |               |               |               |               |               |
|-------------------------|-----------|--|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                         |           | N  | 1             | 2             | 3             | 4             | 5             | 6             | 7             |
| Information             | Silver S. | 12   | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)     | 1<br>(8.3%)   | 3<br>(25.0%)  | 2<br>(16.7%)  | 6<br>(50%)    |
|                         |           |  | <b>91.7%</b>  |               |               |               |               |               |               |
|                         | Gen. X    | 105  | 2<br>(1.9%)   | 5<br>(4.8%)   | 5<br>(4.8%)   | 4<br>(3.8%)   | 24<br>(22.9%) | 15<br>(14.3%) | 50<br>(47.6%) |
|                         |           |  | <b>84.8%</b>  |               |               |               |               |               |               |
|                         | Gen. Y    | 131  | 3<br>(2.3%)   | 3<br>(2.3%)   | 9<br>(6.9%)   | 19<br>(14.5%) | 34<br>(26.0%) | 24<br>(18.3%) | 39<br>(29.8%) |
|                         |           |  | <b>74.1%</b>  |               |               |               |               |               |               |
| Gen. Z                  | 17        | 0<br>(0%)  | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)     | 5<br>(29.4%)  | 6<br>(35.3%)  | 6<br>(35.3%)  |               |
|                         |           | <b>100%</b>  |               |               |               |               |               |               |               |
| Self-Presentation       | Silver S. | 11   | 3<br>(27.3%)  | 1<br>(9.1%)   | 0<br>(0%)     | 0<br>(0%)     | 4<br>(36.4%)  | 2<br>(18.2%)  | 1<br>(9.1%)   |
|                         |           |  | <b>63.7%</b>  |               |               |               |               |               |               |
|                         | Gen. X    | 88   | 19<br>(21.6%) | 11<br>(12.5%) | 11<br>(12.5%) | 10<br>(11.4%) | 13<br>(14.8%) | 5<br>(5.7%)   | 19<br>(21.6%) |
|                         |           |  | <b>46.6%</b>  |               |               |               |               |               |               |
|                         | Gen. Y    | 113  | 21<br>(18.6%) | 12<br>(10.6%) | 19<br>(16.8%) | 15<br>(13.3%) | 18<br>(15.9%) | 6<br>(5.3%)   | 16<br>(14.2%) |
|                         |           |  | <b>51.3%</b>  |               |               |               |               |               |               |
| Gen. Z                  | 15        | 2<br>(13.3%)   | 2<br>(13.3%)  | 1<br>(6.7%)   | 2<br>(13.3%)  | 3<br>(20.0%)  | 4<br>(26.7%)  | 1<br>(6.7%)   |               |
|                         |           | <b>33.3%</b>   |               |               |               |               |               |               |               |
| Socialization           | Silver S. | 11   | 1<br>(9.1%)   | 0<br>(0%)     | 2<br>(18.2%)  | 3<br>(27.3%)  | 1<br>(9.1%)   | 2<br>(18.2%)  | 2<br>(18.2%)  |
|                         |           |  | <b>27.3%</b>  |               |               |               |               |               |               |
|                         | Gen. X    | 91   | 13<br>(14.3%) | 7<br>(7.7%)   | 10<br>(11.0%) | 15<br>(16.5%) | 14<br>(15.4%) | 10<br>(11.0%) | 22<br>(24.2%) |
|                         |           |  | <b>33%</b>    |               |               |               |               |               |               |
|                         | Gen. Y    | 121  | 14<br>(11.6%) | 12<br>(9.9%)  | 12<br>(9.9%)  | 31<br>(25.6%) | 16<br>(13.2%) | 18<br>(14.9%) | 18<br>(14.9%) |
|                         |           |  | <b>31.4%</b>  |               |               |               |               |               |               |
| Gen. Z                  | 16        | 3<br>(18.8%)   | 3<br>(18.8%)  | 0<br>(0%)     | 3<br>(18.8%)  | 3<br>(18.8%)  | 3<br>(18.8%)  | 3<br>(18.8%)  |               |
|                         |           | <b>43.9%</b>   |               |               |               |               |               |               |               |
| Entertainment           | Silver S. | 12   | 1<br>(8.3%)   | 1<br>(8.3%)   | 1<br>(8.3%)   | 3<br>(25.0%)  | 3<br>(25.0%)  | 2<br>(16.7%)  | 2<br>(16.7%)  |
|                         |           |  | <b>24.9%</b>  |               |               |               |               |               |               |
|                         | Gen. X    | 98   | 11<br>(11.2%) | 6<br>(6.1%)   | 8<br>(8.2%)   | 17<br>(17.3%) | 20<br>(20.4%) | 14<br>(14.3%) | 22<br>(22.4%) |
|                         |           |  | <b>25.5%</b>  |               |               |               |               |               |               |
|                         | Gen. Y    | 127  | 4<br>(3.1%)   | 12<br>(9.4%)  | 17<br>(13.4%) | 35<br>(27.6%) | 27<br>(21.3%) | 14<br>(11.0%) | 18<br>(14.2%) |
|                         |           |  | <b>25.9%</b>  |               |               |               |               |               |               |
| Gen. Z                  | 16        | 0<br>(0%)  | 2<br>(12.5%)  | 4<br>(25.0%)  | 2<br>(12.5%)  | 3<br>(18.8%)  | 3<br>(18.8%)  | 2<br>(12.5%)  |               |
|                         |           | <b>37.5%</b>   |               |               |               |               |               |               |               |

# Appendix 5

| Self-Determination                  |            | Likert Scale<br>(1-Strongly Disagree – 7-Strongly Agree) |                |               |               |               |               |               |               |
|-------------------------------------|------------|--|----------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Motivation                          | Regulation | N  | 1              | 2             | 3             | 4             | 5             | 6             | 7             |
| External                            | Male       | 24   | 20<br>(83.3%)  | 3<br>(12.5%)  | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)     | 1<br>(4.2%)   |
|                                     |            |  | <b>95.8%</b>   |               |               |               |               |               |               |
|                                     | Female     | 240  | 228<br>(95.0%) | 1<br>(0.4%)   | 0<br>(0%)     | 1<br>(0.4%)   | 2<br>(0.8%)   | 0<br>(0%)     | 8<br>(3.3%)   |
|                                     |            |  | <b>95.4%</b>   |               |               |               |               |               |               |
| Introjected                         | Male       | 24   | 19<br>(79.2%)  | 3<br>(12.5%)  | 0<br>(0%)     | 1<br>(4.2%)   | 0<br>(0%)     | 0<br>(0%)     | 1<br>(4.2%)   |
|                                     |            |  | <b>91.7%</b>   |               |               |               |               |               |               |
|                                     | Female     | 240  | 220<br>(91.7%) | 8<br>(3.3%)   | 2<br>(0.8%)   | 1<br>(0.4%)   | 0<br>(0%)     | 0<br>(0%)     | 9<br>(3.8%)   |
|                                     |            |  | <b>95.8%</b>   |               |               |               |               |               |               |
| Extrinsic                           | Male       | 24   | 5<br>(20.8%)   | 3<br>(12.5%)  | 5<br>(20.8%)  | 4<br>(16.7%)  | 4<br>(16.7%)  | 1<br>(4.2%)   | 1<br>(4.2%)   |
|                                     |            |  | <b>54.1%</b>   |               |               |               |               |               |               |
|                                     | Female     | 231  | 45<br>(19.5%)  | 34<br>(14.7%) | 20<br>(8.7%)  | 46<br>(19.9%) | 29<br>(12.6%) | 30<br>(13.0%) | 27<br>(11.7%) |
|                                     |            |  | <b>42.9%</b>   |               |               |               |               |               |               |
| Integrated                          | Male       | 23   | 7<br>(30.4%)   | 3<br>(13.0%)  | 1<br>(4.3%)   | 5<br>(21.7%)  | 4<br>(17.4%)  | 3<br>(13.0%)  | 0<br>(0%)     |
|                                     |            |  | <b>47.7%</b>   |               |               |               |               |               |               |
|                                     | Female     | 205  | 47<br>(22.9%)  | 24<br>(11.7%) | 28<br>(13.7%) | 46<br>(22.4%) | 28<br>(13.7%) | 17<br>(8.3%)  | 15<br>(7.3%)  |
|                                     |            |  | <b>48.3%</b>   |               |               |               |               |               |               |
| Intrinsic                           | Male       | 24   | 0<br>(0%)      | 4<br>(16.7%)  | 4<br>(16.7%)  | 3<br>(12.5%)  | 3<br>(12.5%)  | 3<br>(12.5%)  | 6<br>(25.0%)  |
|                                     |            |  | <b>33.4%</b>   |               |               |               |               |               |               |
|                                     | Female     | 239  | 7<br>(2.9%)    | 8<br>(3.3%)   | 16<br>(6.7%)  | 37<br>(15.5%) | 55<br>(23.0%) | 53<br>(22.2%) | 63<br>(26.4%) |
|                                     |            |  | <b>12.9%</b>   |               |               |               |               |               |               |
| Amotivation ("Boredom")             | Male       | 24   | 14<br>(58.3%)  | 2<br>(8.3%)   | 3<br>(12.5%)  | 4<br>(16.7%)  | 2<br>(8.3%)   | 2<br>(8.3%)   | 1<br>(4.2%)   |
|                                     |            |  | <b>79.1%</b>   |               |               |               |               |               |               |
|                                     | Female     | 227  | 127<br>(55.9%) | 39<br>(17.2%) | 13<br>(5.7%)  | 20<br>(8.8%)  | 17<br>(7.5%)  | 4<br>(1.8%)   | 7<br>(3.1%)   |
|                                     |            |  | <b>78.8%</b>   |               |               |               |               |               |               |
| Amotivation ("Just for heck of it") | Male       | 24   | 6<br>(25.0%)   | 5<br>(20.8%)  | 0<br>(0%)     | 3<br>(12.5%)  | 3<br>(12.5%)  | 0<br>(0%)     | 3<br>(12.5%)  |
|                                     |            |  | <b>45.8%</b>   |               |               |               |               |               |               |
|                                     | Female     | 235  | 74<br>(31.5%)  | 28<br>(11.9%) | 21<br>(8.9%)  | 27<br>(11.5%) | 27<br>(11.5%) | 26<br>(11.1%) | 32<br>(13.6%) |
|                                     |            |  | <b>52.3%</b>   |               |               |               |               |               |               |

# Appendix 6

| Self-Determination |            |     | Likert Scale<br>(1-Strongly Disagree – 7-Strongly Agree) |               |               |               |                |               |               |             |
|--------------------|------------|-----|--|---------------|---------------|---------------|----------------|---------------|---------------|-------------|
| Motivation         | Regulation | N   | 1  | 2             | 3             | 4             | 5              | 6             | 7             |             |
| External           | Silver S.  | 12  | 11<br>(91.7%)  | 0<br>(0%)     | 0<br>(0%)     | 1<br>(8.3%)   | 0<br>(0%)      | 0<br>(0%)     | 0<br>(0%)     |             |
|                    |            |     | <b>91.7%</b>   |               |               | 8.3%          |                | 0%            |               |             |
|                    |            |     | 1  | 0             | 0             | 0             | 0              | 0             | 5             |             |
|                    | Gen. X     | 103 | 97<br>(94.2%)  | 1<br>(1.0%)   | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)      | 0<br>(0%)     | 0<br>(0%)     | 5<br>(4.9%) |
|                    |            |     | <b>95.2%</b>   |               |               | 0%            |                | 4.9%          |               |             |
|                    |            |     | 124  | 3             | 0             | 0             | 2              | 0             | 4             |             |
|                    | Gen. Y     | 133 | 124<br>(93.2%)   | 3<br>(2.3%)   | 0<br>(0%)     | 0<br>(0%)     | 1.5%<br>(1.5%) | 0<br>(0%)     | 0<br>(0%)     | 4<br>(3.0%) |
|                    |            |     | <b>95.5%</b>   |               |               | 0%            |                | 4.5%          |               |             |
|                    |            |     | 17   | 0             | 0             | 0             | 0              | 0             | 0             |             |
|                    | Gen. Z     | 17  | 17<br>(100%)   | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)      | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)   |
|                    |            |     | <b>100%</b>  |               |               | 0%            |                |               |               |             |
|                    |            |     | 12   | 0             | 0             | 0             | 0              | 0             | 0             |             |
| Introjected        | Silver S.  | 12  | 12<br>(100%)   | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)      | 0<br>(0%)     | 0<br>(0%)     |             |
|                    |            |     | <b>100%</b>  |               |               | 0%            |                |               |               |             |
|                    |            |     | 94   | 2             | 0             | 1             | 0              | 0             | 6             |             |
|                    | Gen. X     | 103 | 94<br>(91.3%)  | 2<br>(1.9%)   | 0<br>(0%)     | 1<br>(1.0%)   | 0<br>(0%)      | 0<br>(0%)     | 0<br>(0%)     | 6<br>(5.8%) |
|                    |            |     | <b>93.2%</b>   |               |               | 1.0%          |                | 5.8%          |               |             |
|                    |            |     | 119  | 9             | 0             | 1             | 0              | 0             | 4             |             |
|                    | Gen. Y     | 133 | 119<br>(89.5%)   | 9<br>(6.8%)   | 0<br>(0%)     | 1<br>(0.8%)   | 0<br>(0%)      | 0<br>(0%)     | 0<br>(0%)     | 4<br>(3.0%) |
|                    |            |     | <b>96.3%</b>   |               |               | 0.8%          |                | 3.0%          |               |             |
|                    |            |     | 15   | 0             | 2             | 0             | 0              | 0             | 0             |             |
|                    | Gen. Z     | 17  | 15<br>(88.2%)  | 0<br>(0%)     | 2<br>(11.8%)  | 0<br>(0%)     | 0<br>(0%)      | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)   |
|                    |            |     | <b>100%</b>  |               | 0%            |               |                |               |               |             |
|                    |            |     | 1  | 2             | 2             | 5             | 0              | 1             | 1             |             |
| Extrinsic          | Silver S.  | 12  | 1<br>(8.3%)  | 2<br>(16.7%)  | 2<br>(16.7%)  | 5<br>(41.7%)  | 0<br>(0.0%)    | 1<br>(8.3%)   | 1<br>(8.3%)   |             |
|                    |            |     | <b>41.7%</b>   |               |               | 16.8%         |                |               |               |             |
|                    |            |     | 26   | 14            | 11            | 15            | 8              | 13            | 14            |             |
|                    | Gen. X     | 101 | 26<br>(25.7%)  | 14<br>(13.9%) | 11<br>(10.9%) | 15<br>(14.9%) | 8<br>(7.9%)    | 13<br>(12.9%) | 14<br>(13.9%) |             |
|                    |            |     | <b>50.5%</b>   |               |               | 14.9%         |                | 34.7%         |               |             |
|                    |            |     | 22   | 18            | 10            | 26            | 20             | 17            | 13            |             |
|                    | Gen. Y     | 126 | 22<br>(17.5%)  | 18<br>(14.3%) | 10<br>(7.9%)  | 26<br>(20.6%) | 20<br>(15.9%)  | 17<br>(13.5%) | 13<br>(10.3%) |             |
|                    |            |     | <b>39.7%</b>   |               |               | 20.6%         |                | 39.7%         |               |             |
|                    |            |     | 1  | 3             | 2             | 4             | 2              | 5             | 0             |             |
|                    | Gen. Z     | 17  | 1<br>(5.9%)  | 3<br>(17.6%)  | 2<br>(11.8%)  | 4<br>(23.5%)  | 2<br>(11.8%)   | 5<br>(29.4%)  | 0<br>(0%)     |             |
|                    |            |     | <b>100%</b>  |               |               | 23.5%         |                | 41.2%         |               |             |
|                    |            |     | 2  | 1             | 2             | 3             | 2              | 2             | 0             |             |
| Integrated         | Silver S.  | 12  | 2<br>(16.7%)   | 1<br>(8.3%)   | 2<br>(16.7%)  | 3<br>(25.0%)  | 2<br>(16.7%)   | 2<br>(16.7%)  | 0<br>(0%)     |             |
|                    |            |     | <b>41.7%</b>   |               |               | 25.0%         |                | 33.4%         |               |             |
|                    |            |     | 26   | 12            | 12            | 17            | 10             | 9             | 7             |             |
|                    | Gen. X     | 93  | 26<br>(28.0%)  | 12<br>(12.9%) | 12<br>(12.9%) | 17<br>(18.3%) | 10<br>(10.8%)  | 9<br>(9.7%)   | 7<br>(7.5%)   |             |
|                    |            |     | <b>53.8%</b>   |               |               | 18.3%         |                | 28%           |               |             |
|                    |            |     | 23   | 13            | 14            | 25            | 19             | 7             | 8             |             |
|                    | Gen. Y     | 109 | 23<br>(21.1%)  | 13<br>(11.9%) | 14<br>(12.8%) | 25<br>(22.9%) | 19<br>(17.4%)  | 7<br>(6.4%)   | 8<br>(7.3%)   |             |
|                    |            |     | <b>45.8%</b>   |               |               | 22.9%         |                | 31.1%         |               |             |
|                    |            |     | 3  | 1             | 1             | 7             | 1              | 2             | 0             |             |
|                    | Gen. Z     | 15  | 3<br>(20.0%)   | 1<br>(6.7%)   | 1<br>(6.7%)   | 7<br>(46.7%)  | 1<br>(6.7%)    | 2<br>(13.3%)  | 0<br>(0%)     |             |
|                    |            |     | <b>33.4%</b>   |               |               | 46.7%         |                | 20%           |               |             |
|                    |            |     | 0  | 0             | 0             | 0             | 4              | 3             | 5             |             |
| Intrinsic          | Silver S.  | 12  | 0<br>(0%)  | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)     | 4<br>(33.3%)   | 3<br>(25.0%)  | 5<br>(41.7%)  |             |
|                    |            |     | <b>0%</b>  |               |               |               | 100%           |               |               |             |
|                    |            |     | 6  | 6             | 8             | 13            | 20             | 23            | 27            |             |
|                    | Gen. X     | 103 | 6<br>(5.8%)  | 6<br>(5.8%)   | 8<br>(7.8%)   | 13<br>(12.6%) | 20<br>(19.4%)  | 23<br>(22.3%) | 27<br>(26.2%) |             |
|                    |            |     | <b>19.4%</b>   |               |               | 12.6%         |                | 67.9%         |               |             |
|                    |            |     | 1  | 5             | 11            | 21            | 31             | 29            | 34            |             |
|                    | Gen. Y     | 132 | 1<br>(0.8%)  | 5<br>(3.8%)   | 11<br>(8.3%)  | 21<br>(15.9%) | 31<br>(23.5%)  | 29<br>(22.0%) | 34<br>(25.8%) |             |
|                    |            |     | <b>12.9%</b>   |               |               | 15.9%         |                | 71.3%         |               |             |
|                    |            |     | 0  | 1             | 1             | 6             | 3              | 3             | 3             |             |
|                    | Gen. Z     | 17  | 0<br>(0%)  | 1<br>(5.9%)   | 1<br>(5.9%)   | 6<br>(35.3%)  | 3<br>(17.6%)   | 3<br>(17.6%)  | 3<br>(17.6%)  |             |
|                    |            |     | <b>11.8%</b>   |               |               | 35.3%         |                | 52.8%         |               |             |

| Self-Determination                       |                |           | Likert Scale<br>(1-Strongly Disagree – 7-Strongly Agree) |               |               |               |               |               |               |             |
|--|----------------|-----------|--|---------------|---------------|---------------|---------------|---------------|---------------|-------------|
| Motivation                               | Regulation     | N         | 1  | 2             | 3             | 4             | 5             | 6             | 7             |             |
| Amotivation<br>("Boredom")               | Non-Regulation | Silver S. | 12   | 10<br>(83.3%) | 1<br>(8.3%)   | 0<br>(0%)     | 1<br>(8.3%)   | 0<br>(0%)     | 0<br>(0%)     | 0<br>(0%)   |
|  |                |           |  | <b>91.6%</b>  |               |               | 8.3%          |               | 0%            |             |
|  |                |           |  | 64            | 14            | 6             | 3             | 8             | 2             | 3           |
|  | Gen. X         | 100       | 64<br>(64.0%)  | 14<br>(14.0%) | 6<br>(6.0%)   | 3<br>(3.0%)   | 8<br>(8.0%)   | 2<br>(2.0%)   | 3<br>(3.0%)   |             |
|  |                |           | <b>84%</b>   |               |               | 3.0%          |               | 13%           |               |             |
|  |                |           | 57   | 24            | 9             | 16            | 9             | 3             | 5             |             |
|  | Gen. Y         | 123       | 57<br>(46.3%)  | 24<br>(19.5%) | 9<br>(7.3%)   | 16<br>(13.0%) | 9<br>(7.3%)   | 3<br>(2.4%)   | 5<br>(4.1%)   |             |
|  |                |           | <b>73.1%</b>   |               |               | 13%           |               | 13.8%         |               |             |
|  |                |           | 11   | 2             | 1             | 1             | 2             | 0             | 0             |             |
|  | Gen. Z         | 17        | 11<br>(64.7%)  | 2<br>(11.8%)  | 1<br>(5.9%)   | 1<br>(5.9%)   | 1<br>(5.9%)   | 2<br>(11.8%)  | 0<br>(0%)     | 0<br>(0%)   |
|  |                |           | <b>82.4%</b>   |               |               | 5.9%          |               | 11.8%         |               |             |
|  |                |           | 5  | 0             | 1             | 2             | 2             | 1             | 1             |             |
| Amotivatın<br>("Just for heck<br>of it") | Non-Regulation | Silver S. | 12   | 5<br>(41.7%)  | 0<br>(0%)     | 1<br>(8.3%)   | 2<br>(16.7%)  | 2<br>(16.7%)  | 1<br>(8.3%)   | 1<br>(8.3%) |
|  |                |           |  | <b>50%</b>    |               |               | 16.7%         |               | 33.3%         |             |
|  |                |           |  | 35            | 10            | 5             | 15            | 11            | 11            | 15          |
|  | Gen. X         | 102       | 35<br>(34.3%)  | 10<br>(9.8%)  | 5<br>(4.9%)   | 15<br>(14.7%) | 11<br>(10.8%) | 11<br>(10.8%) | 15<br>(14.7%) |             |
|  |                |           | <b>49%</b>   |               |               | 14.7%         |               | 36.3%         |               |             |
|  |                |           | 34   | 20            | 14            | 16            | 16            | 11            | 17            |             |
|  | Gen. Y         | 128       | 34<br>(26.6%)  | 20<br>(15.6%) | 14<br>(10.9%) | 16<br>(12.5%) | 16<br>(12.5%) | 11<br>(8.6%)  | 17<br>(13.3%) |             |
|  |                |           | <b>53.1%</b>   |               |               | 12.5%         |               | 34.4%         |               |             |
|  |                |           | 6  | 3             | 1             | 1             | 1             | 3             | 2             |             |
|  | Gen. Z         | 17        | 6<br>(35.3%)   | 3<br>(17.6%)  | 1<br>(5.9%)   | 1<br>(5.9%)   | 1<br>(5.9%)   | 3<br>(17.6%)  | 2<br>(11.8%)  |             |
|  |                |           | <b>58.8%</b>   |               |               | 5.9%          |               | 35.3%         |               |             |

## References

1. World Health Organization. <https://www.who.int/dietphysicalactivity/pa/en/>. Accessed 22 Feb 2020
2. Liu, S.: Fitness activity tracker – Statistics Facts (2019). <https://www.statista.com/topics/4393/fitness-and-activity-tracker>. Accessed 22 Feb 2020
3. Fritz, T., Huang, E.M., Murphy, G.C., Zimmermann, T.: Persuasive technology in the real world: a study of long-term use of activity sensing devices for fitness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 487–496. ACM, New York (2014). <https://doi.org/10.1145/2556288.2557383>
4. Ilhan, A.: Motivations to join fitness communities on facebook: which gratifications are sought and obtained? In: Meiselwitz, G. (ed.) SCSM 2018. LNCS, vol. 10914, pp. 50–67. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91485-5\\_4](https://doi.org/10.1007/978-3-319-91485-5_4)
5. Nelson, E.C., Verhagen, T., Noordzij, M.L.: Health empowerment through activity trackers: an empirical smart wristband study. *Comput. Hum. Behav.* **62**, 364–374 (2016). <https://doi.org/10.1016/j.chb.2016.03.065>
6. Shin, G., et al.: Wearable activity trackers, accuracy, adoption, acceptance and health impact: a systematic literature review. *J. Biomed. Inform.* **93**, 103153 (2019). <https://doi.org/10.1016/j.jbi.2019.103153>
7. Ilhan, A., Henkel, M.: 10,000 steps a day for health? user-based evaluation of wearable activity trackers. In: Proceedings of the 51st Hawaii International Conference on System Sciences, pp. 3376–3385. ScholarSpace, Honolulu (2018). <http://hdl.handle.net/10125/50316>
8. Clement, J.: Number of monthly active Facebook users worldwide as of 4th quarter 2019. <https://www.statista.com/statistics/264810/number-of-monthly-activefacebook-users-worldwide/>. Accessed 24 Feb 2020
9. Newberry, C.: 33Facebook Stats That Matter to Marketers in 2020. <https://blog.hootsuite.com/facebook-statistics/>. Accessed 24 Feb 2020
10. Sharon, A.J., Yom-Tov, E., Baram-Tsabari, A.: Vaccine information seeking on social QA services. *Vaccine*, **38**(12), 2691–2699 (2020). <https://doi.org/10.1016/j.vaccine.2020.02.010>
11. Athanasopoulou, C., Sakellari, E.: Facebook and health information: content analysis of groups related to schizophrenia. *Stud. Health Technol. Inf.* **213**, 255–258 (2015). <https://doi.org/10.3233/978-1-61499-538-8-255>
12. Greene, J.A., Choudhry, N.K., Kilabuk, E., Shrank, W.H.: Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. *J. Gen. Intern. Med.* **26**(3), 287–292 (2011). <https://doi.org/10.1007/s11606-010-1526-3>
13. Katz, E., Blumler, J.G., Gurevitch, M.: Uses and gratifications research. *Public Opin. Quart.* **37**(4), 509–523 (1973–1974)
14. Katz, E., Blumler, J.G., Gurevitch, M.: Utilization of mass communication by individual. In: Blumler, J.G., Katz, E. (eds.) *The Uses of Mass Communications: Current Perspective on Gratifications Research*, pp. 19–32. Sage, Beverly Hills CA (1974)
15. Greenberg, B.S.: Gratifications of television viewing and their correlates for British children. In: Blumler, J.G., Katz, E. (eds.) *The Uses of Mass Communications: Current Perspectives on Gratifications Research*, pp. 195–233. Sage, Beverly Hills (1974)
16. Katz, E., Haas, H., Gurevitch, M.: On the use of the mass media for important things. *Am. Sociol. Rev.* **38**, 164–181 (1973). <https://doi.org/10.2307/2094393>
17. Palmgreen, P., Wenner, L.A., Rayburn, J.D.: Relations between gratifications sought and obtained: a study of television news. *Commun. Res.* **7**, 161–192 (1980). <https://doi.org/10.1177/009365028000700202>
18. Schaffarczyk, L., Ilhan, A.: Healthier life and more fun? users’ motivations to apply activity tracking technology and the impact of gamification. In: Meiselwitz, G. (ed.) *HCI 2019*. LNCS, vol. 11579, pp. 124–136. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-21905-5\\_10](https://doi.org/10.1007/978-3-030-21905-5_10)

19. Klenk, S., Reifegerste, D., Renatus, R.: Gender differences in gratifications from fitness app use and implications for health interventions. *Mob. Media Commun.* **5**, 178–193 (2017). <https://doi.org/10.1177/2050157917691557>
20. Joinson, A.N.: ‘Looking at’, ‘looking up’ or ‘keeping up with’ people? motives and uses of facebook. In: CHI 2008, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1027–1036. ACM, New York (2008). <https://doi.org/10.1145/1357054.1357213>
21. Raacke, J., Raacke-Bonds, J.: MySpace and Facebook: applying the uses and gratifications theory to exploring friend-networking sites. *CyberPsychol. Behav.* **11**(2), 169–174 (2008). <https://doi.org/10.1089/cpb.2007.0056>
22. Scheibe, K., Meschede, C., Göretz, J., Stock, W.G.: Giving and taking gratifications in a gamified social live streaming service. In: Proceedings of the 5th European Conference on Social Media. ECSM, pp. 264–273. Academic Conferences and Publishing Limited, Reading (2018)
23. Scheibe, K., Zimmer, F., Stock, W.G.: Social media usage of asylum seekers in germany. In: Proceedings of the 6th European Conference on Social Media. ECSM, pp. 263–272. Academic Conferences and Publishing International, Reading, UK (2019)
24. Tanta, I., Mihovilović, M., Sablić, Z.: Uses and gratification theory – why adolescents use facebook? *Medijska istraž. ivanija* **20**(2), 85–111 (2014)
25. Whiting, A., Williams, D.: Why people use social media: a uses and gratifications approach. *Qual. Market Res.* **16**(4), 362–369 (2013). <https://doi.org/10.1108/QMR-06-2013-0041>
26. Zimmer, F., Scheibe, K., Stock, W.G.: A model for information behavior research on social live streaming services (SLSSs). In: Meiselwitz, G. (ed.) SCISM 2018. LNCS, vol. 10914, pp. 429–448. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91485-5\\_33](https://doi.org/10.1007/978-3-319-91485-5_33)
27. Zimmer, F., Scheibe, K.: What drives streamers? users’ characteristics and motivations on social live streaming services. In: Proceedings of the 52nd Hawaii International Conference on System Sciences, pp. 2538–2547. ScholarSpace, Honolulu (2019) <http://hdl.handle.net/10125/59692>
28. McQuail, D.: Mass communication theory. Sage, London (1983)
29. Park, N., Kee, K.F., Valenzuela, S.: Being immersed in social networking environment: facebook groups, uses and gratifications, and social Outcomes. *CyberPsychol. Behav.* **12**(6), 729–733 (2009). <https://doi.org/10.1089/cpb.2009.0003>
30. Krasnova, H., Veltri, N.F., Eling, N., Buxmann, P.: Why men and women continue to use social networking sites: the role of gender differences. *J. Strateg. Inf. Syst.* **26**(4), 261–284 (2017). <https://doi.org/10.1016/j.jsis.2017.01.004>
31. Fietkiewicz, K.J., Lins, E., Baran, K.S., Stock, W.G.: Inter-Generational comparison of social media use: investigating the online behavior of different generational cohorts. In: Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS), pp. 3829–3838. IEEE Computer Society, Washington (2016). <https://doi.org/10.1109/HICSS.2016.477>
32. Deci, E.L., Ryan, R.M.: Intrinsic Motivation and Self-Determination in Human Behavior. Plenum Press, New York (1985)
33. Deci, E.L., Ryan, R.M.: The “what” and “why” of goal pursuits: human needs and the selfdetermination of behavior. *Psychol. Inq.* **11**, 227–268 (2000). [https://doi.org/10.1207/S15327965PLI1104\\_01](https://doi.org/10.1207/S15327965PLI1104_01)
34. Ryan, R.M., Deci, E.L.: Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* **25**, 54–67 (2000). <https://doi.org/10.1006/ceps.1999.1020>
35. Ryan, R.M., Williams, G.C., Patrick, H., Deci, E.L.: Self-determination theory and physical activity: the dynamics of motivation in development and wellness. *Hell. J. Psychol.* **6**, 107–124 (2009). <https://doi.org/10.1080/17509840701827437>

36. Stragier, J., Evens, T., Mechant, P.: Broadcast yourself: an exploratory study of sharing physical activity on social networking sites. *Media Int. Aust.* **155**, 120–129 (2015). <https://doi.org/10.1177/1329878X1515500114>
37. Wang, C.K.J., Leng, H.K., Kee, Y.H.: Use of facebook in physical activity intervention programme. A test of self-determination theory. *Int. J. Sport Psychol.* **46**(3), 210–224 (2015). <http://hdl.handle.net/10497/17252>
38. Divine, A., Watson, P.M., Baker, S., Hall, C.R.: Facebook, relatedness and exercise motivation in university students: a mixed methods investigation. *Comput. Hum. Behav.* **91**, 138–150 (2019). <https://doi.org/10.1016/j.chb.2018.09.037>
39. Ferguson, R., Gutberg, J., Schattke, K., Paulin, M., Jost, N.: Self-determination theory, social media and charitable causes: an in-depth analysis of autonomous motivation. *Eur. J. Soc. Psychol.* **45**(3), 298–307 (2015). <https://doi.org/10.1002/ejsp.2038>
40. Laerd Statistics: Mann-Whitney U Test SPSS Statistics. <https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php>. Accessed 24 Feb 2020
41. Laerd Statistics: Kruskal-Wallis H Test Stata. <https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php>. Accessed 24 Feb 2020



# Intelligent Auto Technologies Are Here, and Drivers Are Losing Control

Brian M. Jones<sup>(✉)</sup>

Tennessee Technological University, Cookeville, TN 38501, USA  
bjones@tntech.edu

**Abstract.** Vehicles are being embedded with advanced technological capabilities and we are headed quickly towards completely self-driving cars. These same technologies will also enhance the convenience and safety features in all types of vehicles. The era of intelligent cars is here. Most vehicles manufactured and sold in 2019, that are not economy versions, include some form of adaptive cruise control, lane keeping assist, collision mitigation, video monitoring of adjacent lane obstructions, heads up display (HUD), and parking assist up to and including self-parking capabilities. With the plethora of new technologies being incorporated into our vehicles that are still controlled by us, how are we managing? According to anecdotal evidence gathered from several Honda dealerships in middle Tennessee many drivers aren't satisfied with some of these capabilities. This paper attempts to explain the technologies available and in use, and the driver's thoughts about partial assist technologies that are currently in place. While these features are designed to enhance safety and decrease traffic accidents, are they actually increasing driver safety or simply facilitating driver distraction? NTSB standards and recommendations will be summarized, and a questionnaire developed that helps us examine driver reaction to the technologies embedded in 2019 and 2020 Honda vehicles.

**Keywords:** Vehicle-to-vehicle communication · Intelligent vehicles · Autonomous vehicles · Assistive driver technology · Vehicle CAN bus

## 1 Introduction

The rite of passage from vehicle occupant to vehicle driver has been a tradition for teenagers across the United States for years. Most teens dreamed of driving themselves around early in High School as they viewed it as gaining their freedom and becoming independent. Many teens were also into cars and the performance aspect of them as they dreamed of owning a Mustang, Corvette, Z-28, Porsche, Ferrari, or Lamborghini one day. And while many drivers will not own a supercar, the cars they will own continue to evolve. Over the years cars have increased their capabilities and conveniences and now include options that most never dreamed possible.

Current vehicles include standard items such as seatbelts, power steering, power disc brakes, automatic transmissions, cruise control, and variable speed windshield wipers

that were at one time high cost options available only in premium branded cars. Today these “options” are standard on base line vehicles by most car manufacturers and currently available options have evolved with the advent of computer-based systems embedded in most new cars. The Controller Area Network (CAN bus) system is a vehicle standard designed to allow microcontrollers and embedded devices to communicate with each other in applications without a host computer (CiA 2020).

The longevity of CAN bus (Guardigli 2012) is still to be determined but the long view guarantees some type of computer-based network embedded in vehicles for the foreseeable future. Vehicles are gaining additional “intelligent” features and the need for computer control intensifies with these added capabilities and their reliance on networked systems.

This paper proposes a research agenda to investigate the value placed on assistive technology by drivers. The general aim of this research will be to determine what intelligent capabilities are valued by drivers and what capabilities drivers would rather not have. A survey was developed to determine the value drivers perceive in these technologies and if the drivers deem them too intrusive. The following research questions are driving the proposed research:

*Q1: What intelligent driver assist system is valued by drivers?*

*Q2: How intuitive is the control or operation of the assistive technology currently offered?*

*Q3: Do drivers want more assistive technology or less embedded into their cars?*

*Q4: Do drivers trust the technology to allow completely self-driving/autonomous cars?*

## **2 Vehicle Automation Levels**

The National Highway Transportation Safety Administration (NHTSA) categorizes automobile automation into 6 distinct categories from 0 up to 5. At the zero level the driver of a vehicle is responsible for all activities behind the wheel. There are no automatic controls for steering, lighting, braking, etc. At level 1 automation begins to show itself with limited auto steering or braking and acceleration inputs. Level 2 steps automation up to include level 1 capabilities plus the ability to steer and brake simultaneously in certain situations. Level 3 adds the ability for the car to perform all driving functions in specific situations with monitoring required, but at all other times, the driver is in complete control of driving operations. Level 4 adds the ability of full automation in certain situations without driver monitoring, but in all other situations the driver must still monitor and be ready to assume control. Level 5 is defined as complete driverless automation. The “driver” and everyone in the car is a passenger and no monitoring is required for any and all driving tasks (NHTSA 2020a, b, c, d).

## **3 Vehicle Cybersecurity**

Car manufacturers in coordination with the NHTSA uses a multi-layered approach to cybersecurity. Effort focuses on entry points; attention to both the vehicle’s wired and wireless access points and systems. Any system connected to a vehicle’s network could be vulnerable to a cyberattack and therefore must be protected from unwanted intrusion



or access. A layered approach (multiple targets that can be updated and changed when necessary) to vehicle cybersecurity decreases the chance of a successful cyber-attack. This approach also mitigates the negative consequences of successful vehicle system access. NHTSA recommends a comprehensive and systematic approach to developing layered cybersecurity capabilities for vehicles:

1. A risk-based prioritized identification and protection process for safety-critical vehicle control systems;
2. Timely detection and rapid response to potential vehicle cybersecurity incidents on America's roads;
3. Architectures, methods, and measures that design-in cyber resiliency and facilitate rapid recovery from incidents when they occur; and
4. Methods for effective intelligence and information sharing across the industry to facilitate quick adoption of industry-wide lessons learned. NHTSA encouraged the formation of Auto-ISAC, an industry environment emphasizing cybersecurity awareness and collaboration across the automotive industry (Vehicle Cybersecurity 2020).

## **4 Available Automation**

Many studies have tried to pit automation against man by comparing what machines can do best against what humans can do best (Fitts 1951). Modern vehicle manufacturers are trying to see what machine (automation) can do to help the human driver be even better. So while automation is also being looked at to allow full automation in vehicles it is also being developed to enhance the skill and judgement of a human driver. The following is a list of several innovations currently being developed and installed in vehicles.

### **4.1 Adaptive Cruise Control**

An adaptive cruise control system (ACC) automatically varies the speed of the vehicle based on the presence and location of vehicles ahead. The system accepts and maintains the set speed until the presence of a vehicle obstructing its path is registered. The system then adjusts its speed to maintain a safe distance between the primary vehicle and any vehicle in its path. Once traffic is clear of the sensor range, the initially set speed is resumed.

### **4.2 Forward Collision Warning (and Associated Systems)**

A forward collision warning (FCW) system alerts the driver of a potential frontal vehicle crash. This system is an intelligent safety technology that monitors the vehicle's speed and the speed of the vehicle in front of it, as well as the distance between them. If the system detects that the vehicles are getting too close, it alerts the driver that the vehicles are getting to close. The system assumes based on the closure speed that a crash could be imminent and alerts the driver that action is needed to avoid a crash. This type system is a warning only system and requires driver input to avoid the crash.

A new generation of system called a Collision Mitigation Systems (CMS) was designed to prevent forward crashes from happening by allowing automatic vehicle reaction. Once the forward collision sensors detect that a crash is imminent, the system activates the automatic emergency braking (AEB) systems that automatically applies the brakes to bring the vehicle to a complete stop (or at least enough brake pressure to avoid the collision). Most of these systems first alert the driver of the impending collision so that they can take corrective action. If the system deems it appropriate, they supplement the brake pressure applied by the driver to prevent the crash. If the driver takes no corrective action, the system may automatically apply braking to prevent or reduce the severity of the crash.

Many CMS include capabilities of dynamic brake support (DBS) that aids and supports the driver's application of braking pressure and crash imminent braking (CIB) that applies up to maximum brake pressure. Both support capabilities for the CMS are designed to help prevent accidents thereby reducing injury and preventing vehicle deaths.

### **4.3 Park Assist Systems**

Park Assist systems use similar technology found in collision mitigation systems with a computer that does all of the steering calculations required to move the car from adjacent to a parking space to fully parked within the space. Most of these capabilities are being enhanced with the goal of complete driverless parking but currently require braking input from a driver and some systems require manually changing gears as well.

### **4.4 Lane Keep Assist Systems (and Associated Systems)**

Lane keep assist systems (LKS) system is a safety technology enabled by the computer-based system of modern vehicles. It is intended to prevent drivers from accidentally drifting or merging out of the intended lane of traffic.

LKS systems rely on the information relayed to it by sensors included in a lane departure warning system (LDW). This system includes sensors that determine whether a car is about to move out of its intended lane of travel. When the system determines an accidental lane departure is eminent, the LKS activates by issuing a warning or even by steering, braking, or accelerating one or more of the wheels, or a combination of all, that moves the vehicle back into its intended lane of travel.

LDW systems use cameras to monitor roadway lane markings and to detect when a vehicle is moving out of its traffic lane. When lane departure is detected and that the vehicle is out of its lane, an audio, visual, or both, alerts the driver and the LKS of the unintentional lane departure so that the driver (or LKS) can steer the vehicle back into its appropriate lane.

## **5 Key Benefits of Automation**

While all drivers can benefit from assistive technology, different aged drivers value different vehicle embedded intelligent technology (Jenness et al. 2008).

## 5.1 Safety

The safety benefits of automated vehicles are the driving force of recent enhancements. Automation in modern vehicles has the potential to save lives and reduce injuries. This is known because the statistics show that approximately 94% of serious crashes are due to human error. Automation in vehicles has the potential to eliminate human error from the crash equation, which will help protect everyone on the road including pedestrians and bicyclists. When you consider that approximately 37,133 people died in motor vehicle-related crashes in the U.S. in 2017, 36,560 in 2018, and 35,756 in 2019, the full benefit of lifesaving driver assistive technologies is understood (NHTSA 2020a, b, c, d).

Driver studies have shown that driver's preferences for the presentation of information about a situation are correlated with their rating of urgency. Based on the findings in Lerner 2014, and the general characteristics of each variable studied, three general categories of urgency seemed to emerge: High threat, caution, and no urgency (Lerner et al. 2014). The participants of this study rated situations focusing on convenience and sustainability as least urgent and safety related situations as most urgent. This type of study has helped shape the focus on safety information with automobile manufacturers for "intelligent" features in current vehicles.

## 5.2 Economic

A 2010 NHTSA study showed motor vehicle crashes cost \$242 billion in economic activity, including \$57.6 billion in lost worker productivity and \$594 billion due to loss of life and decreased quality of life. If vehicle automation can eliminate a significant number of motor vehicle crashes, we would also eliminate these enormous personal and societal costs.

## 5.3 Efficiency and Convenience

Roads filled with automated vehicles with vehicle to vehicle communication (V2V) technology could also cooperate to smooth traffic flow and reduce traffic congestion. NHTSB estimates show that Americans spent an estimated 6.9 billion hours in traffic delays in 2014, cutting into time at work or with family and significantly increasing fuel costs and vehicle emission. A recent study stated that automated vehicles could free up as much as 50 min each day that had previously been dedicated to driving.

## 5.4 Mobility

While its full societal benefits are difficult to project, the transformative potential of automated vehicles and their driver assistance features can also be understood by reviewing U.S. demographics and the communities these technologies could help to support.

For example, automated vehicles may also provide new mobility options to millions more Americans. Today there are 61 million Americans with some form of disability (CDC 2018). Many of the "intelligent" options included in modern cars can make driving easier by handling some of the surround monitoring thereby easing the cognitive load on the driver.

In many places across the country employment or independent living rests on the ability to drive. Automated vehicles could extend that kind of freedom to millions more.

## 6 The Future of Vehicle Automation

One of the key pieces to a continuing evolution for intelligent cars is the full implementation of Vehicle-to-vehicle (V2V) communication. V2V communication allows vehicles to share their speed, location, and heading information wirelessly. The wireless information is broadcast and received omnidirectionally up to ten times per second. This allows each vehicle to build a situational awareness (anything in its area) that encompasses a 360-degree pattern area surrounding the vehicle.

Vehicles equipped with the appropriate system and software will use the information from surrounding vehicles to determine potential issues. These issues can include stopping, yielding, turn avoidance, etc. when appropriate. Thereby eliminating potential accidents and/or reducing slowdowns before they happen. The system can accomplish this goal by employing visual, tactile, and audible alerts to warn the affected drivers. These alerts allow drivers the ability to take action to avoid crashes or to understand the automatic actions the vehicle is taking in appropriately equipped cars.

V2V messages have a minimum range of 300 m and this range can be extended in certain environments. The systems can simultaneously identify and monitor dangers obscured by surrounding traffic, terrain, and weather. The V2V technology extends the currently available crash avoidance systems that rely on radars and cameras to identify vehicle threats. The intent of this evolution in intelligent vehicle design is to allow drivers to avoid crashes and even traffic slowdowns and obstacles (V2V NHTSA 2020).

The vehicle information broadcast does not identify the driver or the vehicle to any surrounding vehicles. The system also incorporates privacy and security controls to deter vehicle tracking and unauthorized access to the system.

Six and a half million police-reported crashes were recorded in 2017, resulting in more than thirty-seven thousand fatalities and two and a half million injuries in the United States. The hope is that connected intelligent vehicle will provide drivers with the information and capabilities necessary to anticipate potential issues and significantly reduce the number of lives lost to vehicle crashes each year.

The following is a list of future technologies and capabilities currently under development by numerous automotive manufacturers.

1. Smart fuel savings technologies (cylinder shut down when light engine load, engine off at complete stops, etc.)
2. Self-healing paint
3. Smart batteries that detect requirements and adjust alternator inputs
4. Real time vehicle tracking (similar to a “jail broke” Lojack system)
5. Multilanguage support
6. Multi-measurement system support (KM vs Miles)
7. Light weight body panels that store energy for both hybrid and electric car propulsion systems
8. Automobile linking technologies to enhance safety and shorten travel times
9. Built in cellular/Internet capabilities (LTE, 4G, etc.)
10. Semi-autonomous and autonomous driving.

## Appendix 1

### Levels of Automation:

- Level 0: The human driver does all the driving.
- Level 1: An advanced driver assistance system (ADAS) on the vehicle can sometimes assist the human driver with either steering or braking/accelerating, but not both simultaneously.
- Level 2: An advanced driver assistance system (ADAS) on the vehicle can itself actually control both steering and braking/accelerating simultaneously under some circumstances. The human driver must continue to pay full attention (“monitor the driving environment”) at all times and perform the rest of the driving task.
- Level 3: An Automated Driving System (ADS) on the vehicle can itself perform all aspects of the driving task under some circumstances. In those circumstances, the human driver must be ready to take back control at any time when the ADS requests the human driver to do so. In all other circumstances, the human driver performs the driving task.
- Level 4: An Automated Driving System (ADS) on the vehicle can itself perform all driving tasks and monitor the driving environment – essentially, do all the driving – in certain circumstances. The human need not pay attention in those circumstances.
- Level 5: An Automated Driving System (ADS) on the vehicle can do all the driving in all circumstances. The human occupants are just passengers and need never be involved in driving.

From Jan 2020 NHTSB: <https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety>.

## Appendix 2

### Vehicle Automation Survey

**Year:** < >    **Make:** Honda    **Model:** < >

**Age:**                      **Gender:**                      **Education Level:**

**Assistive Automation Present:**     Adaptive Cruise Control     Lane Keeping Assist

Automated Interior Lighting     Automated Exterior Lighting     Park Assist or Self Park

Voice Controlled GPS and/or Telephone     Vehicle Surround/Rear Viewing

Automated Windshield Wipers     Windshield Heads Up Display

**What assist system is a must have for all future vehicles:** < >

**What assist technology have you disabled (if any):** < >

**Why did you disable this capability:**

**Overall do you want more automation, the same as currently available, or less:**

< >

**Are the driver assist technologies intuitive to operate (activate, disable, enable):**

< >

**Does the automation enhance your driving experience:** < >

**Does the automation distract you from looking outside:** < >

**Does the automation input control when not needed (e.g. override):** < >

**Do you feel safer in an intelligence equipped car than when in a different vehicle:**

< >

**Are you ready for full autonomous driving vehicles:** < >

**Would you trust automation to drive without a steering wheel and brake peddle:**

< >

## References

- CAN in Automation, CiA (2020). <https://www.can-cia.org/can-knowledge/can/can-history/>. Accessed Feb 2020
- CDC (2018). <https://www.cdc.gov/media/releases/2018/p0816-disability.html>. Accessed Feb 2020
- Fitts, P.M. (ed.): Human Engineering for an Effective Air Navigation and Traffic Control System. National Research Council, Washington (1951)
- Guardigli, M.: Hacking your Car (2012). <https://marco.guardigli.it/2010/10/hacking-your-car.html>. Accessed Feb 2020
- Jenness, J.W., Lerner, N.D., Mazor, S., Osberg, J.S., Tefft, B.C.: Use of advanced in-vehicle technology by young and older early adopters. Survey Results on Headlamp Systems. Report No. DOT HS 810 902. Washington, DC: National Highway Traffic Safety Administration (2008)
- Lerner, N., et al.: Human factors for connected vehicles: effective warning interface research findings (Report No. DOT HS 812 068). National Highway Traffic Safety Administration, Washington, DC, September 2014
- NHTSA: Innovation (2020a). <https://www.nhtsa.gov/technology-innovation>. Accessed Jan 2020
- NHTSA: Traffic Deaths (2020b). <https://www.nhtsa.gov/traffic-deaths-2018>. Accessed Feb 2020
- Vehicle Cybersecurity, NHTSA. <https://www.nhtsa.gov/technology-innovation/vehicle-cybersecurity>. Accessed Feb 2020
- Vehicle Privacy, NHTSA (2020c). <https://www.nhtsa.gov/technology-innovation/vehicle-data-privacy>. Accessed Feb 2020
- Vehicle Safety, NHTSA (2020d). <https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety>. Accessed Feb 2020
- V2V, NHTSA Vehicle to Vehicle. <https://www.nhtsa.gov/technology-innovation/vehicle-vehicle-communication>. Accessed Feb 2020



# Emotions in Online Gambling Communities: A Multilevel Sentiment Analysis

Markus Kaakinen<sup>1</sup> , Atte Oksanen<sup>2</sup> , Anu Sirola<sup>2</sup> , Iina Savolainen<sup>2</sup> ,  
and David Garcia<sup>3,4</sup> 

<sup>1</sup> University of Helsinki, 00014 Helsinki, Finland

markus.kaakinen@helsinki.fi

<sup>2</sup> Tampere University, Tampere, Finland

<sup>3</sup> Medical University of Vienna, Vienna, Austria

<sup>4</sup> Complexity Science Hub Vienna, Vienna, Austria

**Abstract.** In this study, we analyzed whether interaction dynamics are related to emotional expressions within online gambling communities. As data, we used 8452 comments posted on Reddit gambling communities. The data were analyzed with sentiment analysis tool VADER and multilevel regression analysis. Results showed that comments were more positive when they were directed to other users and made by users with more interactive commenting behavior. Comments were less positive in those discussions that were most active and in those that mainly involved replies to other users. We also found that more positive posts received more positive commenting and negative posts received more negative comments. Overall, the activity and interactivity of communication and emotional correlation are associated with positive emotional expression in online communication. For negative emotions, we found evidence only for emotional correlation. Future studies should explore how interaction dynamics together with more contextual factors shape emotional expressions within online communities.

**Keywords:** Gambling · Online communities · Emotions · Sentiment analysis · Online interaction

## 1 Introduction

Online communities are virtual social networks that can be based on various online platforms and on mutual interests and identities [1, 2] or peer support [3], for example. These online communities are increasingly important contexts for sharing personal experiences and narratives and gaining social support from others [4–8]. Especially for individuals with insufficient social relationships offline, online social ties can offer social support and a sense of belonging [8]. Social support can promote wellbeing directly but also by buffering distressing life events [9, 10]. Some studies have suggested, however, that online social ties might not be as effective wellbeing buffers as more traditional offline ones [11, 12].

In addition to the beneficial consequences, online communities can promote adverse ideals [2]. On social media, users can easily search for others who share their interests

and attitudes which can help normalize harmful behaviors [2, 13]. Due to the selective nature of social media, those already interested in risky behaviors are also the most likely to access potentially harmful online content [14]. In line with this, earlier studies have stressed the role of online communities in hate speech [2, 15–17], eating disorders [18–20] and gambling problems [21, 22], to mention a few.

Gambling communities are online groups that are based on social networking sites and discussion forums [21, 22]. According to earlier research, the discussions within gambling communities are mostly about gambling experiences and tips, as well as gambling in general [22]. Thus, gambling communities can be considered as part of online gambling ecosystem, consisting of gambling sites and social contexts where users can interact with other gamblers and those interested in gambling [23]. Individuals who visit online gambling communities are also more likely to show more at-risk and probable pathological gambling and only a small proportion of users enter these communities to discuss gambling problems and recovery [22, 24].

Personal narratives shared within online communities are often emotional and cover intimate and even traumatic life events [5, 6, 20]. Especially negative sentiments are common in online recovery and peer support communities, as users with high psychological distress tend to express more negative emotions in their online interaction [25]. Furthermore, emotional expressions in online interaction are usually correlated. Negative messages commonly generate negative responses, while the opposite is true for positive ones. This phenomenon is typically explained by emotional contagion or emotional homophily [26].

Emotional contagion refers to a phenomenon where positive and negative emotions are transferred from one individual to another [26, 27]. In their large-scale experiment, Kramer and colleagues [27] found that a decrease in positive emotional expression within a social networking site made users' negative commenting more likely and positive commenting less likely. Emotional homophily, in turn, suggests that emotions do not just spread in online environment but they shape online interaction and social networks [28]. In other words, users tend to prefer interaction with others who share their emotional reactions and, thus, online interaction is concentrated around cliques with similar emotional valence [29]. Both emotional contagion and emotional homophily predict that emotional valence of online messages should shape the way they are responded to.

Emotional expressions are also shaped by the interaction within online communities [19]. Currently, however, there is a lack of research analyzing which characteristics of interaction dynamics are the ones that contribute to emotional expressions within online communities. According to Gonzalez-Bailon and colleagues [30], the hierarchical features of online discussion structures can be used in analysis of the quality of online interaction dynamics. The most interactive online interaction leads to deep discussions consisting of a long chain of comments referring back to each other. In other words, deep discussion chains represent complementary forms of interaction where the conversation flows and latter comments are made in reference to earlier comments. This kind of dialogical interaction supports elaborated argumentation between users and has been found to facilitate solidarity in offline groups [30, 31]. Conversations are wide, in turn, if they have a high level of activity and participation [30]. This does not mean, however, that there is much explicit interaction between participants as it may include mainly



solitary comments that are made without reference to other users (or even reading other's comments) [32]. Here, we use the terms *activity* to refer to the amount of commenting and *interactivity* to refer to the degree of explicit references to other users (i.e. comments that refer to other comments).

In this study, we focus on the role of interaction dynamics in emotional expressions within online gambling communities. More specifically, we analyze whether more active and interactive commenting and emotional correlation can explain positive and negative commenting within online gambling communities.

## 2 Methods

### 2.1 The Sample Selection and Data Extraction

The data was collected from Reddit (<https://www.reddit.com>). Reddit is a popular online discussion forum that provides a platform for different online communities (i.e. “subreddits”) where users can anonymously share content and discuss various subjects. The platform hosts a huge variety of communities based on common interests or identities [e.g. 33–35]. Within these communities, users can post contents that other community members are then able to rate and comment [35]. Most of the communities are free to join but some require subscribing. Previous studies have analyzed Reddit communities' role, for example, in recovery from eating disorders [36] and addictive behaviors [37].

Our analyses are based on a dataset consisting of a full history of interaction in 43 gambling-related communities (subreddits) that was automatically retrieved using Reddit platform's interface. No data were collected from subreddits that required subscribing. The data includes 19,942 posts and 258,043 comments of 44,490 users. The gambling communities were selected by using 84 gambling related search terms.

The search terms included such terms as “gambling”, “problem gambling”, “betting”, “poker”, and “addiction”. The search terms were constructed on the basis of review on gambling terminology and subreddit searches on the Reddit platform. In addition to the research team, five social psychology MA students participated in the search of gambling-related subreddits. The team of MA students was supervised by the research team and received explicit instructions for the task.

For this study, we selected two gambling subreddits for further analysis. The subreddits were *gambling* and *problemgambling*. According to subreddit descriptions, the gambling subreddit is a community that “promotes healthy and responsible gambling” (<https://www.reddit.com/r/gambling/>). Those users who are worried about their gambling behavior are directed to the *problemgambling* subreddit. The *problemgambling* subreddit is described as a community that serves as a “resource for individuals who have struggled - or know somebody who has struggled - with a gambling problem” (<https://www.reddit.com/r/problemgambling/>). The first subreddit hosts an array of discussions related to gambling and gambling tips while the latter is mainly focusing on social support for the recovery from problematic gambling. These two selected gambling subreddits allow us to analyze two different communities with a contrasting stance towards gambling.

All the comments included in the analysis contained full information of the comment (i.e. the comment text and time of posting) and its author and reference to the discussion

it was posted to. The final data consist of 8452 comments (2999 to the gambling subreddit and 5474 to the problemgambling subreddit) from 1817 individual users in 1249 discussions.

## 2.2 Measures

**Outcome Measures.** Our outcome measures included the positivity and negativity of comments. These emotional characteristics were quantified using the VADER tool designed for sentiment analysis of social media texts [38]. VADER is a sentiment analysis method that combines lexical and rule-based approaches and performs especially well with social media texts. It estimates the valence and arousal expressed in textual material and produces separate measures for positivity and negativity (ranging from 0 to 1 with higher value indicating more intense sentiment).

**Comment Level Measures.** Comment level predictors measured whether the comment was a response to other users' comment or not (0 = no and 1 = yes) and the depth of the comment within a discussion. Higher values of comment depth indicate that the comment was placed deep within an interaction chain.

**User Level Measures.** Our user level predictors included the total number of comments made by the user and the user ratio of replies. Ratio of replies refers to the number of user's replies to others' comments divided by the total number of user comments. Thus, higher values indicate that most of the user's comments were replies to others' comments (more interactive approach) while lower values indicate that the user mainly wrote comments with no reference to other users.

**Discussion Level Predictors.** Discussion level predictors measured the total number of comments and the deepest discussion chain within the discussion, the discussion's reply ratio, and the positivity and negativity of the original post (that started the discussion) and its title. Similarly to the user ratio of replies, the discussion reply ratio counted the number of all comments replying to other comments divided by the total number of comments within a discussion. Thus, high values of the discussion reply ratio indicate the discussion to be more interactive, as majority of comments were replies to other users' comments. The positivity and negativity of the post and its title were measured using the VADER tool (see Outcome measures). In addition, we included a measure indicating whether the comment was posted to gambling or problemgambling discussion.

For further analysis, all our study variables (except the reply to others variable that was dichotomous) were standardized. Mean value and standard deviation for all the study variables (before standardization) are presented in Table 1.

## 2.3 Statistical Technique

For descriptive statistics, we calculated mean values and standard deviations for our study variables before standardization (Table 1). This was done for the whole data set and separately for gambling and problemgambling discussions. The associations

**Table 1.** Mean values and standard deviations for our study variables before standardization.

| Fixed part                                    | Combined |        | Gambling |        | Problem gambling |        |
|---|----------|--------|----------|--------|------------------|--------|
|   | M        | Std.   | M        | Std.   | M                | Std.   |
| Comment positivity <sup>a</sup>               | 0.17     | 0.17   | 0.17     | 0.19   | 0.17             | 0.16   |
| Comment negativity <sup>a</sup>               | 0.08     | 0.11   | 0.07     | 0.12   | 0.09             | 0.10   |
| Reply to others <sup>b</sup>                  | 0.43     | 0.50   | 0.51     | 0.50   | 0.39             | 0.49   |
| Comment depth <sup>c</sup>                    | 1.92     | 1.51   | 2.26     | 1.81   | 1.73             | 1.29   |
| Number of user's comments <sup>d</sup>        | 85.39    | 159.54 | 61.14    | 149.41 | 98.63            | 163.31 |
| User's reply ratio <sup>a</sup>               | 0.43     | 0.31   | 0.51     | 0.33   | 0.39             | 0.29   |
| Number of comments in discussion <sup>e</sup> | 16.42    | 25.01  | 20.41    | 26.90  | 14.24            | 23.63  |
| Discussion's reply ratio <sup>a</sup>         | 0.43     | 0.25   | 0.51     | 0.23   | 0.39             | 0.26   |
| Depth of discussion <sup>c</sup>              | 2.46     | 2.11   | 2.96     | 2.51   | 2.19             | 1.80   |
| Title positivity <sup>a</sup>                 | 0.12     | 0.20   | 0.14     | 0.19   | 0.12             | 0.20   |
| Post positivity <sup>a</sup>                  | 0.12     | 0.09   | 0.11     | 0.11   | 0.12             | 0.08   |
| Title negativity <sup>a</sup>                 | 0.11     | 0.19   | 0.05     | 0.12   | 0.14             | 0.21   |
| Post negativity <sup>a</sup>                  | 0.09     | 0.08   | 0.04     | 0.06   | 0.11             | 0.08   |

Note. <sup>a</sup>range from 0 to 1. <sup>b</sup>0 = no, 1 = yes. <sup>c</sup>range from 1 to 10. <sup>d</sup>range from 1 to 515. <sup>e</sup>Range from 1 to 143.

between our outcome variables and predictors were analyzed using multilevel random coefficient regression modelling and maximum likelihood estimation. This method was used to account for the hierarchical structure of the data.

Multilevel models were estimated separately for positive and negative emotions. Each model included all our predictors and random intercepts at the user and discussion level. For our models (Table 2), we report regression coefficients with corresponding 95% confidence intervals (based on Wald method). For random part we report the standard deviations of the random intercepts.

### 3 Results

According to our multilevel models, replies to other users' comments were more positive than other comments ( $b = 0.14$ , 95% CI [0.07, 0.21]). In addition, users who were more active in replying to others wrote more positive comments ( $b = 0.03$ , 95% CI [0.00, 0.06]). Comments were less positive in discussions with highest number of comments ( $b = -0.07$ , 95% CI [-0.14, -0.01]) and high proportion of replies ( $b = -0.06$ , 95% CI [-0.10, -0.02]). Comments to posts with positive titles were more positive ( $b = 0.05$ , 95% CI [0.02, 0.08]). Comments expressed fewer positive emotions in the case of negative posts ( $b = -0.05$ , 95% CI [0.02, 0.08]) and posts with negative titles ( $b = -0.04$ , 95% CI [-0.08, -0.02]). All the other estimated associations were nonsignificant (the 95% confidence interval included zero).

**Table 2.** Linear multilevel regression models predicting comment positivity and negativity.

| Fixed part                       | Comment positivity |                | Comment negativity |               |
|----------------------------------|--------------------|----------------|--------------------|---------------|
|                                  | b                  | 95%CI          | b                  | 95%CI         |
| Problemgambling subreddit        | 0.01               | [-0.06, 0.09]  | 0.15               | [0.08, 0.22]  |
| Reply to others                  | 0.14               | [0.07, 0.21]   | -0.02              | [-0.10, 0.04] |
| Comment depth                    | 0.03               | [-0.00, 0.06]  | 0.01               | [-0.02, 0.04] |
| Number of user's comments        | -0.01              | [-0.12, 0.10]  | -0.00              | [-0.10, 0.09] |
| User's reply ratio               | 0.03               | [0.00, 0.06]   | -0.01              | [-0.04, 0.02] |
| Number of comments in discussion | -0.07              | [-0.14, -0.01] | 0.02               | [-0.03, 0.07] |
| Discussion's reply ratio         | -0.06              | [-0.10, -0.02] | 0.00               | [-0.03, 0.04] |
| Depth of discussion              | -0.04              | [-0.08, 0.01]  | 0.01               | [-0.02, 0.05] |
| Title positivity                 | 0.05               | [0.02, 0.08]   | -0.02              | [-0.04, 0.01] |
| Post positivity                  | 0.02               | [-0.01, 0.05]  | -0.01              | [-0.04, 0.01] |
| Title negativity                 | -0.04              | [-0.07, -0.01] | 0.05               | [0.03, 0.08]  |
| Post negativity                  | -0.05              | [-0.08, -0.02] | 0.06               | [0.03, 0.09]  |
| Intercept                        | -0.07              | [-0.15, 0.02]  | -0.08              | [-0.05, 0.07] |
| Random part                      |                    | Std.           |                    | Std.          |
| User                             |                    | 0.27           |                    | 0.23          |
| Post                             |                    | 0.29           |                    | 0.18          |
| Residual                         |                    | 0.91           |                    | 0.95          |

Note. *b* = unstandardized regression coefficient. 95%CI = 95% confidence interval. Reply ratio = number of comments replying to other users/number of all comments.

Comments in the problemgambling subreddit were more negative than comments in the gambling subreddit ( $b = 0.15$ , 95% CI [0.08, 0.22]). In addition to subreddit, only significant predictors related to emotional correlation, as negative posts ( $b = 0.06$ , 95% CI [0.03, 0.09]) and posts with negative titles ( $b = 0.05$ , 95% CI [0.03, 0.08]) received more negative comments. All other estimated relationships were nonsignificant (the 95% confidence interval included zero).

## 4 Discussion

In this study, we analyzed whether interaction dynamics can explain emotional expressions within online gambling communities. According to our findings, comments posted to a problemgambling community were more negative on average than comments posted to a gambling community. This is in line with earlier studies stating that distressed online community members tend to express more negative emotions in their online interaction [21]. There was no difference in positive emotions. This implies that communication within a community focused on gambling problems includes positive discussions and narratives as well.

Comments made directly to other users were more positive on average. In addition, comments made by users with more interactive commenting behavior were more positive. This implies that positive emotions within these communities are expressed more in dialogues. However, comments in most active discussions and more interactive commenting (involving mostly comments replying to others' comments) were slightly less positive than others. This can be considered surprising as comments replying to other users' comments were more positive on average. In online communities, users often post intimate and emotional narratives for others to see and interact with [5, 6]. The most active discussions and discussions consisting mainly of commenters' replies to each other may be more general (and less intimate) in nature, and remain distant to the original post and the narrative it shared. This could explain why comments were less positive in these discussions.

Both positive and negative emotional expressions were partly explained by emotional correlation [26]. That is, positive posts generated more positive commenting while comments to negative posts were more negative and less positive on average. This finding may imply either emotional transfer or homophily [26]. It is possible that the emotional contents shared in the posts induce similar emotional valence in other users, which is then reflected in correlated emotional expressions [27]. Users may also self-select the conversations that match with their personal emotional valence [e.g. 28–29].

Our analyses were based on data collected from two gambling-related online communities. Involving more communities might have brought new information on the relationship between emotional expressions and interaction dynamics. In addition, gambling and especially gambling problems are emotionally charged themes with distinctive terminology. This might limit the generalization of our results to other online communities. The effect sizes of our models imply that contextual factors are likely of major importance in determining online emotions. However, given the size and complexity of our naturally occurring data, these results can be considered realistic and valuable.

Online communities are increasingly important social contexts for social interaction and social support [5, 6]. Within these communities users can share even the most traumatic life events [3] and, thus, the emotional response they induce can be of great importance. Future studies should concentrate more on how contextual factors (such as the characteristics of shared narratives) and interaction dynamics within online communities interact in shaping emotional expressions.

**Acknowledgements.** This study was funded by the Finnish Foundation for Alcohol Studies (Problem Gambling and Social Media Project, 2017–2019, PI: Atte Oksanen). David Garcia was funded by the Vienna Science and Technology Fund through the project “Emotional Well-Being in the Digital Society” (Grant No. VRG16-005).

## References

1. Mikal, J.P., Rice, R.E., Kent, R.G., Uchino, B.N.: 100 million strong: a case study of group identification and deindividuation on Imgur.com. *New Media Soc.* **18**(11), 2485–2506 (2016)
2. Keipi, T., Näsi, M.J., Oksanen, A., Räsänen, P.: *Online Hate and Harmful Content: Cross-National Perspectives*. Routledge, London (2017)

3. Robinson, C., Pond, R.: Do online support groups for grief benefit the bereaved? systematic review of the quantitative and qualitative literature. *Comput. Hum. Behav.* **100**, 48–59 (2019)
4. Mudry, T.E., Strong, T.: Doing recovery online. *Qual. Health Res.* **23**(3), 313–325 (2013)
5. Shim, M., Cappella, J.N., Han, J.Y.: How does insightful and emotional disclosure bring potential health benefits? study based on online support groups for women with breast cancer. *J. Commun.* **61**(3), 432–454 (2011)
6. Verberne, S., Batenburg, A., Sanders, R., Das, E., Lambooi, M.S.: Analyzing empowerment processes among cancer patients in an online community: A text mining approach. *J. Med. Internet Res.* **21**(4), e9887 (2019)
7. Wang, Y.-C., Kraut, R.E., Levine, J.M.: Eliciting and receiving online support: using computer-aided content analysis to examine the dynamics of online social support. *J. Med. Internet Res.* **17**(4), e99 (2015)
8. Cole, D.A., Nick, E.A., Zelkowitz, R.L., Roeder, K.M., Spinelli, T.: Online social support for young people: does it recapitulate in-person social support; can it help? *Comput. Hum. Behav.* **68**, 456–464 (2017)
9. Cohen, S., Wills, T.A.: Stress, social support, and the buffering hypothesis. *Psychol. Bull.* **98**(2), 310–357 (1985)
10. Thoits, P.: Mechanisms linking social ties and support to physical and mental health. *J. Health Soc. Behav.* **52**(2), 145–161 (2011)
11. Minkinen, J., Oksanen, A., Näsi, M., Keipi, T., Kaakinen, M., Räsänen, P.: Does social belonging to primary groups protect young people from the effects of pro-suicide sites? a comparative study of four countries. *Crisis* **37**(1), 31–41 (2016)
12. Kaakinen, M., Keipi, T., Räsänen, P., Oksanen, A.: Cybercrime victimization and subjective well-being: an examination of the buffering effect hypothesis among adolescents and young adults. *Cyberpsychology Behav. Soc. Network.* **21**(2), 129–137 (2018)
13. Kaakinen, M., Sirola, A., Savolainen, I., Oksanen, A.: Shared identity and shared information in social media: development and validation of the identity bubble reinforcement scale. *Media Psychol.* **23**(1), 25–51 (2020)
14. Kaakinen, M., Sirola, A., Savolainen, I., Oksanen, A.: Young people and gambling content in social media: an experimental insight. *Drug and Alcohol Review*, online first (2020)
15. Baele, S.J., Brace, L., Coan, T.G.: From “Incel” to “Saint”: analyzing the violent worldview behind the 2018 Toronto attack. *Terrorism and Political Violence*, online first (2019)
16. Bliuc, A.-M., Betts, J., Vergani, M., Iqbal, M., Dunn, K.: Collective Identity Changes in Far-right Online Communities: The Role of Offline Intergroup Conflict. *New Media Soc.* **21**(8), 1770–1786 (2019)
17. Kaakinen, M., Räsänen, P., Näsi, M., Minkinen, J., Keipi, T., Oksanen, A.: Social capital and online hate production: a four country survey. *Crime, Law Soc. Change* **69**, 25–39 (2018)
18. Borzekowski, D.L.G., Summer, S., Wilson, J.L., Peebles, R.: e-Ana and e-Mia: A Content Analysis of Pro-Eating Disorder Web Sites. *Am. J. Public Health* **100**(8), 1526–1534 (2010)
19. Oksanen, A., et al.: Pro-anorexia and anti-pro-anorexia videos on YouTube: sentiment analysis of user responses. *J. Med. Internet Res.* **17**(11), e256 (2015)
20. Sirola, A., Kaakinen, M., Turja, T., Oksanen, A.: (Un)doing deviance: social categorization in user reactions to proanorexia videos on YouTube. In: *Digital technology: Advances in Research and Applications*, pp. 231–261. Nova Science Publishers, New York (2019)
21. Savolainen, I., Sirola, A., Kaakinen, M., Oksanen, A.: Peer group identification as determinant of youth behavior and the role of perceived social support in problem gambling. *J. Gambl. Stud.* **35**(1), 15–30 (2019)
22. Sirola, A., Kaakinen, M., Oksanen, A.: Excessive gambling and online gambling communities. *J. Gambl. Stud.* **34**(4), 1313–1325 (2018)
23. O’Leary, K., Carroll, C.: The online poker sub-culture: Dialogues, interactions and networks. *J. Gambl. Stud.* **29**(4), 613–630 (2013)

24. Parke, A., Griffiths, M.D.: Poker gambling virtual communities: the use of Computer-Mediated Communication to develop cognitive poker gambling skills. *Int. J. Cyber Behav. Psychol. Learn. (IJCPL)* **1**(2), 31–44 (2011)
25. Lyons, M., Aksayli, N.D., Brewer, G.: Mental distress and language use: Linguistic analysis of discussion forum posts. *Comput. Hum. Behav.* **87**, 207–211 (2018)
26. Rosenbusch, H., Evans, A.M., Zeelenberg, M.: Multilevel emotion transfer on YouTube: disentangling the effects of emotional contagion and homophily on video audiences. *Soc. Psychol. Pers. Sci.* **10**(8), 1028–1035 (2019)
27. Kramer, A.D.I., Guillory, J.E., Hancock, J.T.: Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Nat. Acad. Sci.* **111**, 8788–8790 (2014). <https://doi.org/10.1073/pnas.1320040111>
28. Song, Y., Dai, X.-Y., Wang, J.: Not all emotions are created equal: expressive behavior of the networked public on China's social media site. *Comput. Hum. Behav.* **60**, 525–533 (2016)
29. Himelboim, I., Cameron, K., Sweetser, K.D., Danelo, M., West, K.: Valence-based homophily on twitter: network analysis of emotions and political talk in the 2012 presidential election. *New Media Soc.* **18**(7), 1382–1400 (2016)
30. Gonzalez-Bailon, S., Kaltenbrunner, A., Banchs, R.E.: The structure of political discussion networks: A model for the analysis of online deliberation. *J. Inf. Technol.* **25**(2), 230–243 (2010)
31. Koudenburg, N., Postmes, T., Gordijn, E.H., Van Mourik Broekman, A.: Uniform and complementary social interaction: distinct pathways to solidarity. *PLoS ONE* **10**(6), e0129061 (2015)
32. Zhang, J., Danescu-Niculescu-Mizil, C., Sauper, C., Taylor, S.J.: Characterizing online public discussions through patterns of participant interactions. In: *Proceedings of the ACM on Human-Computer Interaction 2(CSCW)*, p. 198 (2018)
33. Gray, S.L., Lockridge, L., Peleaux, R.: Social media, online communities, connection and coping: contextual considerations within the developmental period of emerging adulthood. In: *Digital technology: Advances in Research and Applications*, pp. 125–144. Nova Science Publishers, New York (2019)
34. Triggs, A.H., Møller, K., Neumayer, C.: Context collapse and anonymity among queer Reddit users. *New Media Soc.* Online first (2019)
35. Medvedev, A.N., Lambiotte, R., Delvenne, J.-C.: The anatomy of reddit: an overview of academic research. In: Ghanbarnejad, F., Saha Roy, R., Karimi, F., Delvenne, J.-C., Mitra, B. (eds.) *DOOCN 2017. SPC*, pp. 183–204. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-14683-2\\_9](https://doi.org/10.1007/978-3-030-14683-2_9)
36. Bohrer, B.K., Foye, U., Jewell, T.: Recovery as a process: Exploring definitions of recovery in the context of eating-disorder-related social media forums. *Int. J. Eating Disord.* Online first (2020)
37. D'Agostino, A.R., Optican, A.R., Sowles, S.J., Krauss, M.J., Lee, K.E., Cavazos-Rehg, P.A.: Social networking online to recover from opioid use disorder: a study of community interactions. *Drug Alcohol Depend.* **181**, 5–10 (2017)
38. Hutto, C.J., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, pp. 216–225 (2014)



# Analysis of the Exposing Media Pattern that Affect Accessing Own Website

Yuho Katagiri<sup>1</sup> (✉), Kohei Otake<sup>2</sup>, and Takashi Namatame<sup>3</sup>

<sup>1</sup> Graduate School of Science and Engineering, Chuo University, 1-13-27, Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

a16.g5fm@g.chuo-u.ac.jp

<sup>2</sup> School of Information and Telecommunication Engineering, Tokai University, 2-3-23, Takanawa, Minato-ku, Tokyo 108-8619, Japan

otake@tsc.u-tokai.ac.jp

<sup>3</sup> Faculty of Science and Engineering, Chuo University, 1-13-27, Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

nama@indsys.chuo-u.ac.jp

**Abstract.** Recently, due to the expansion of TV or smart phone, the frequency of exposing media has increased. Also, the way of interacting with the media is diversifying by life stages or life balance. In such situations, people get much information about products and services on media. Therefore, it is important to select the best media for advertisement. In this study, we analyze the characteristics of exposing media using media exposure data. First, we performed Non-negative Matrix Factorization (NMF) to extract pattern of exposing media. Second, we used random forest to analyze the characteristics of the exposing media pattern. From our result, we discussed how to advertise on TV and website.

**Keywords:** Exposing media pattern · Nonnegative matrix factorization · Random forest

## 1 Introduction

Recently, due to the expansion of media like TV and smart phone, the frequency of exposing media has increased. According to Institute of Media Environment, people use various media 411.6 min in a day in 2019, especially using time of smart phone and TV increased from the previous year [1]. In such situations, people tend to get many information about products and services on TV and internet. Therefore, it is important to select the best TV programs or websites for each advertisement.

One of the purposes of advertising is to let consumers know about the products and services to increase sales. Therefore, it is important to achieve exposure as many people as possible. However, it is necessary to advertise efficiently because it requires some expenses to advertise. For this reason, advertisers need to select the most appropriate media from a variety of media to advertise to consumers who are potential customers.



In this study, we break situation of utilization media down into patterns using monitors' demographic and media exposure data. Furthermore, we analyze the media exposure pattern of some websites' users. From our result, we discussed how to advertise effectively on TV and the websites about each websites.

## 2 Datasets

In this study, we used i-SSP data provided by INTAGE Holdings Inc.<sup>1</sup>, in Japan. This is the data of media exposure during April in 2016 collected from the same monitors.

We used some parts of the data. The following is the summary.

### – Media data

- Real time TV view: monitor ID, date, day, holiday flag, time (in hours), channel, duration (sec.), genre
- PC, Smart phone browsing: monitor ID, date, day, holiday flag, time (in hours), duration(sec.), genre

In this study, we divided the time period of day into 4 period every 6 h from 5 a.m.

### – Monitors data

Information of monitors attributes such as monitor ID, gender, age and so on.

We extracted monitors who use both “TV and smart phone”, “TV and PC” or “TV, smart phone and PC”. Then, we excluded the monitors who used each media abnormally. Finally, we extracted 1,641 monitors as a dataset.

Also, we had 11 TV-genre and 24 website-genre. Among those genres, we focus on genre of shopping sites because the number of sites have increased [2]. Therefore, we analyzed about Amazon, general shopping sites (exclusive of Amazon and Rakuten, 76 sites) and fashion shopping sites (86 sites).

## 3 Analysis of Monitors

In this section, we explain our analyzing procedure.

### 3.1 Non-negative Matrix Factorization (NMF)

We performed NMF to find media exposure patterns. NMF is a matrix decomposition method to extract potential patterns among data. Given a non-negative matrix  $V$ , we find non-negative matrix factors  $W$  and  $H$  such that Eq. (1)

$$V \approx WH \quad (1)$$

<sup>1</sup> <https://www.nii.ac.jp/dsc/idr/intage/>.

We use  $V$  that consisted watching time of TV-genres and website-genres every monitor. We aggregated watching time (minutes) in time period every monitor. Therefore, the  $V$  was matrix of 1641 rows (monitors) and 140 columns (genres by time period). To calculate  $W$  and  $H$ , we minimize following Eq. (2).

$$F = \|V - WH\|^2 \quad (2)$$

Then,  $W$  is matrix of monitors rows and cluster columns, we obtain contribution of monitors in each cluster. Also  $H$  is matrix of cluster rows and TV-genres and website-genres columns, we obtain characteristics of each cluster.

Also, the update formulas of  $W$  and  $H$  are Eqs. (3), (4).

$$W_{ia} \leftarrow W_{ia} \frac{(VH^T)_{ia}}{(WHH^T)_{ia}} \quad (3)$$

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T WH)_{a\mu}} \quad (4)$$

By initializing  $W$  and  $H$  with random nonnegative values, and we update  $W$  and  $H$  by Eqs. (3) and (4) alternately to converge to optimal solution.

In this study, we divided data into 80% training and 20% test before NMF. First, we performed NMF to training, and found some media exposure patterns. Next, we performed NMF to test with the result of training NMF. We fixed  $H$  that result of training NMF, and calculated  $W_{test}$  using Eq. (5).

$$F = \|V_{test} - W_{test}H\|^2 \quad (5)$$

By calculating Eq. (5), we could get contribution of the new monitor at each existing clusters that we got from trained NMF [3].

### 3.2 Random Forest (RF)

Next, we performed Random Forest to clarify the characteristics of media exposure pattern regarding each of the websites to be analyzed.

RF is one of ensemble learning methods for classification or regression. In RF, plural ordinal decision trees with subset of data used as weak learning machine, and decide estimated class these results were integrated [4]. We used CART method for each decision tree. CART is developed by Breiman, Friedman, Olshen, & Stone in 1984 [5]. In this algorithm, the impurity measure used in building decision tree is Gini Index. The formula of Gini Index is Eq. (6).

$$Gini = 1 - \sum_{t=0}^k P_t^2 \quad (6)$$

- $P_t$ : Proportion of observations with target variable consist ratio  $t$

Also, we obtain importance of variables and the marginal effect of a variable on the class probability. The importance is calculated from the total decrease in node impurities from splitting on the variable averaged over all trees. The formula of the importance is Eq. (7).

$$n_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{7}$$

- $n_j$ : importance of node  $j$
- $w_j$ : weighted of samples reaching node  $j$
- $C_j$ : the impurity value of node  $j$
- $left(j)$ : child node of left split on node  $j$
- $right(j)$ : child node of right split on node  $j$

In this study, we label monitors who accessed to each of the target websites more than 4 times a month as 1, and the others as 0. As explanatory variables used in the model construction, we used the contribution of monitors in each cluster from NMF and the demographic data of monitors.

Also, we excluded cluster variables that close to objective variable. In model of Amazon, we excluded variables of cluster 2, 7, 13 and 21. Moreover, we excluded variables of cluster 11, 13, 16 and 17 in model of general shopping sites. Finally, we excluded variables of cluster 5, 13, 14 and 21 in model of fashion shopping sites. Details of the explanatory variables are shown in Table 1.

**Table 1.** Cluster variables and demographic variables used in the model construction

| Type of variable     |                        | Variable name                | Data type |
|----------------------|------------------------|------------------------------|-----------|
| Objective variable   |                        | Access                       | 0 or 1    |
| Explanatory variable | Cluster                | Contribution of each cluster | Integer   |
|                      | Demographic data       | Age                          | Category  |
|                      |                        | Gender                       | Category  |
|                      |                        | Marital status               | Category  |
|                      |                        | Occupation                   | Category  |
|                      |                        | Family type                  | Category  |
|                      |                        | Family income                | Category  |
|                      | Have children under 17 | 0 or 1                       |           |

### 3.3 Dataset and Evaluation Indicator

From 1641 monitors, we sampled randomly when constructed model. The datasets is Table 2.

**Table 2.** Datasets used in the model construction

| Genre name             | Label | Training data | Test data | Total |
|------------------------|-------|---------------|-----------|-------|
| Amazon                 | 0     | 582           | 170       | 752   |
|                        | 1     | 582           | 159       | 741   |
|                        | Total | 1164          | 329       | 1493  |
| General shopping sites | 0     | 500           | 194       | 694   |
|                        | 1     | 500           | 135       | 635   |
|                        | Total | 1000          | 329       | 1329  |
| Fashion shopping sites | 0     | 643           | 249       | 892   |
|                        | 1     | 645           | 80        | 725   |
|                        | Total | 1288          | 329       | 1617  |

In order to confirm the prediction accuracy of the constructed model, we performed hold-out validation by using the training data and test data. Specifically, we created a confusion matrix like a following table and calculated prediction accuracy of the constructed model by using following equations (Table 3).

**Table 3.** Confusion matrix

|              |          | Predicted class     |                     |
|--------------|----------|---------------------|---------------------|
|              |          | Positive            | Negative            |
| Actual class | Positive | True Positive (TP)  | True Negative (TN)  |
|              | Negative | False Negative (FP) | False Negative (FN) |

- Accuracy (ACC): Ratio of the total number correctly predicted among the total number predicted.

$$ACC = \frac{TP + TN}{FP + FN + TP + TN}$$

- Precision (PRE): Ratio of the total number that is a positive class actually among the total number predicted positive class.

$$PRE = \frac{TP}{TP + FP}$$

- Recall (REC): Ratio of the total number predicted positive class among the total number that is a positive class actually

$$REC = \frac{TP}{FN + TP}$$

- F-measure: harmonic mean of PRE and REC

$$F - \text{measure} = 2 \times \frac{PRE \times REC}{PRE + REC}$$

## 4 Results and Discussions

In this section, we summarize our results.

### 4.1 Result of Nonnegative Matrix Factorization (NMF)

We found 22 clusters from NMF based on cophenetic correlation and understandability [6]. Table 4 shows that top 3 contributed variables in each cluster.

**Table 4.** Top of contributed variables in each cluster

| Cluster | 1st                            | 2nd                             | 3rd                                 |
|---------|--------------------------------|---------------------------------|-------------------------------------|
| V1      | Q&A_morning                    | Ameba_morning                   | Ameba_daytime                       |
| V2      | Amazon_daytime                 | Amazon_evening                  | Amazon_morning                      |
| V3      | talk show_daytime              | variety show_daytime            | news show_daytime                   |
| V4      | variety show_midnight          | news show_midnight              | sports programs<br>midnight         |
| V5      | Rakuten_evening                | Rakuten_daytime                 | Rakuten_morning                     |
| V6      | Anime_morning                  | Anime_evening                   | music programs<br>morning           |
| V7      | free video sites evening       | free video sites midnight       | free video sites daytime            |
| V8      | SNS_evening                    | SNS_midnight                    | SNS_daytime                         |
| V9      | news show evening              | documentary evening             | sports programs<br>evening          |
| V10     | community sites morning        | community sites evening         | community sites<br>daytime          |
| V11     | online game_evening            | online game_morning             | online game_daytime                 |
| V12     | pay video sites_evening        | pay video sites_daytime         | pay video sites morning             |
| V13     | general shopping sites evening | general shopping sites daytime  | household shopping<br>sites_morning |
| V14     | fashion shopping sites evening | fashion shopping sites daytime  | fashion shopping<br>sites_morning   |
| V15     | Google_evening                 | Google_morning                  | Google_daytime                      |
| V16     | portal_evening                 | portal_morning                  | portal_daytime                      |
| V17     | finance sites_evening          | food shopping sites_morning     | finance sites_daytime               |
| V18     | Yahoo_evening                  | Yahoo_morning                   | Yahoo_daytime                       |
| V19     | news show morning              | documentary morning             | TV drama morning                    |
| V20     | review sites_evening           | cosmetic shopping sites morning | cosmetic shopping<br>sites_evening  |
| V21     | company sites midnight         | news sites midnight             | review sites midnight               |
| V22     | variety show evening           | movie programs evening          | TV drama evening                    |

As shown in Table 4, for example, although both cluster 9 and 19 watch news programs and documentary, the time period that they watch them is different. As described we obtained 22 different media exposure patterns from NMF.

Table 5 shows clusters list that brought together the characteristic of each cluster.

**Table 5.** Clusters list

| Cluster | Characteristics  |
|---------|--|
| V1      | Check the ameba and Q&A sites                                    |
| V2      | Check the amazon and home appliances shopping sites              |
| V3      | Watch tv in the daytime  |
| V4      | Watch tv in midnight   |
| V5      | Check the rakuten  |
| V6      | Watch entertainment program like anime and music                 |
| V7      | Check the free video streaming sites                             |
| V8      | Check the sns like twitter, instagram and facebook               |
| V9      | Watch news programs in the evening or night                      |
| V10     | Check the blog sites   |
| V11     | Check the online game sites                                      |
| V12     | Check the pay video streaming sites and video shopping sites     |
| V13     | Check the various shopping sites                                 |
| V14     | Check the fashion shopping sites and various information sites   |
| V15     | Check the google and news/weather forecast sites                 |
| V16     | Check the portal (except for yahoo and google) and company sites |
| V17     | Check the finance service sites and foods shopping sites         |
| V18     | Check the yahoo and news/weather forecast sites                  |
| V19     | Watch tv in the morning  |
| V20     | Check the review site and cosmetics shopping sites               |
| V21     | Check website at midnight  |
| V22     | Watch tv dramas and variety show in the evening or night         |

## 4.2 Result of Random Forest (RF)

To clarify the characteristics of users' media exposure pattern, we built a model that predicts access to each of the target websites more than 4 times a month using RF.

Figures 1, 2, and 3 show the importance of variables which were calculated based on Gini decreasing of the websites.

As shown in Figs. 1, 2 and 3, we found that variables of cluster had higher importance than demographic variables in all model. From this result, it follows from that it is

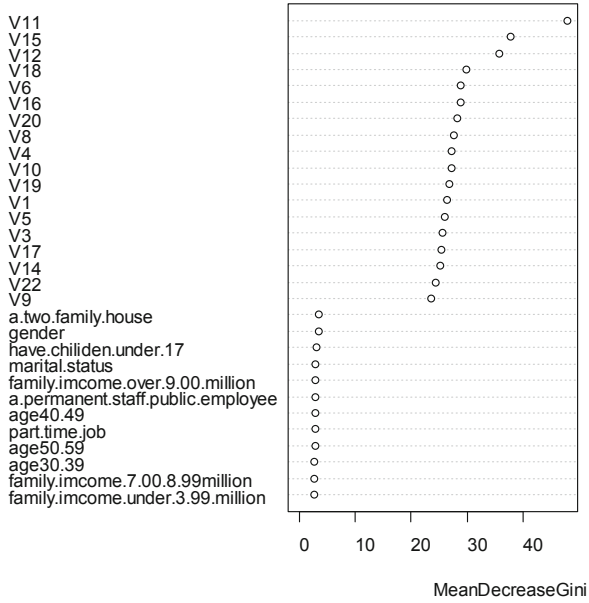


Fig. 1. Importance of variables (Amazon)

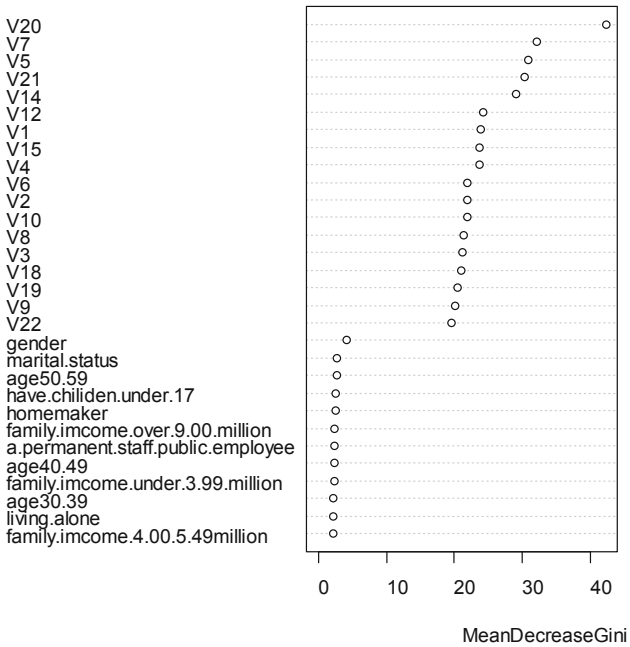
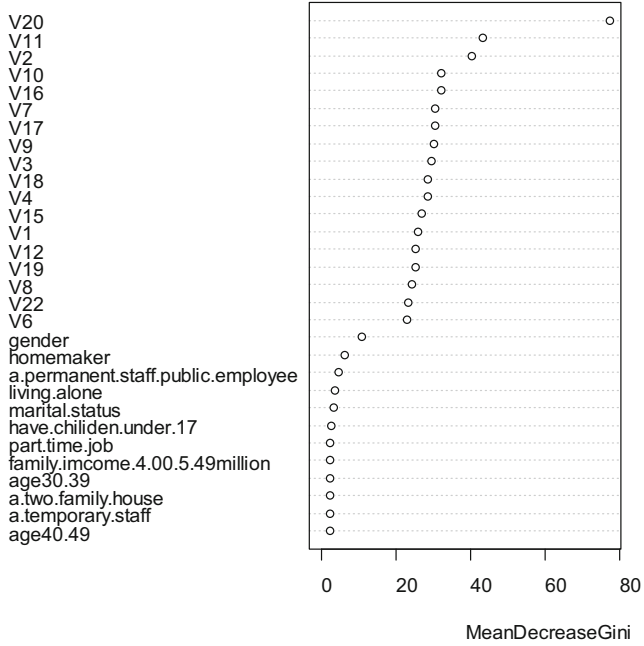


Fig. 2. Importance of variables (general shopping sites)

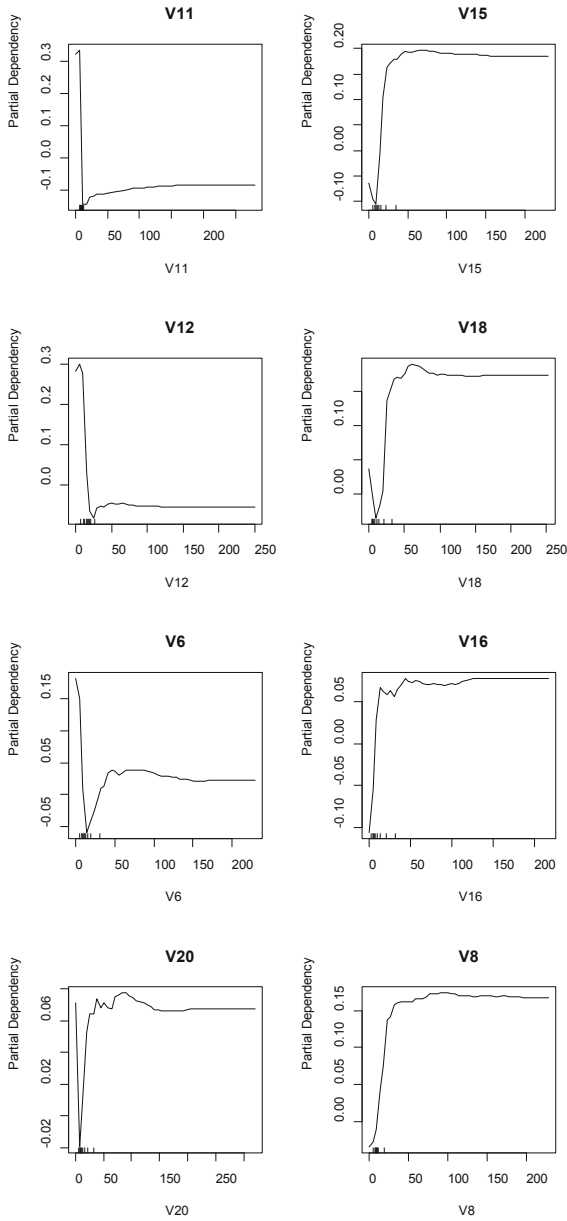


**Fig. 3.** Importance of variables (fashion shopping sites)

necessary to consider not only monitors' demographic like gender or age but also media exposure of users to advertise efficiently.

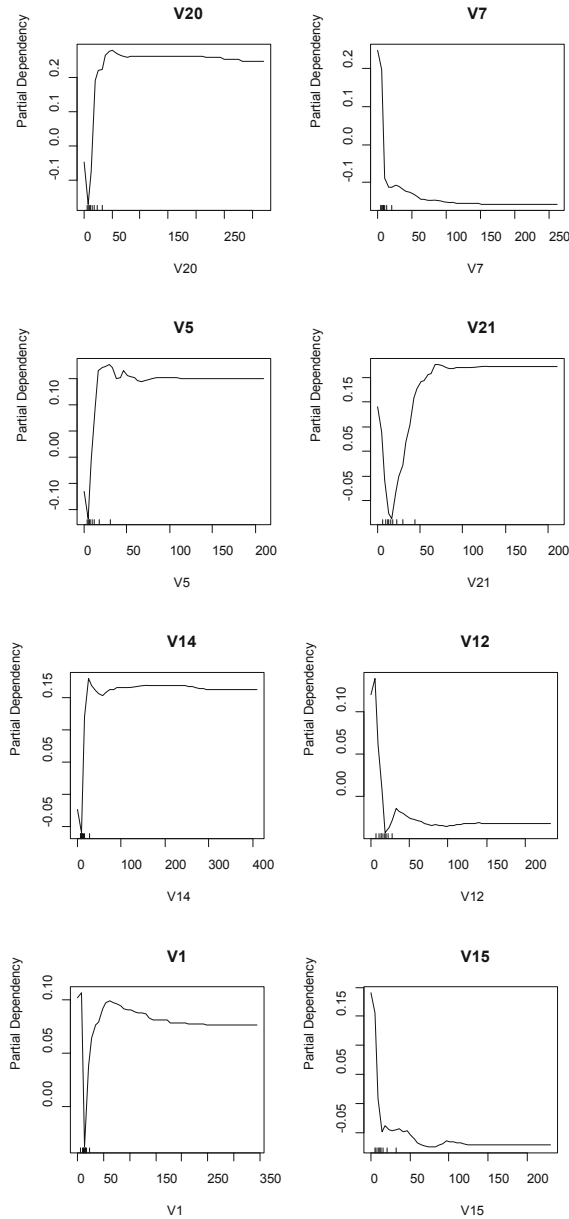
Also, Figs. 4, 5 and 6 show the marginal effect of top 8 variables on each class probability.





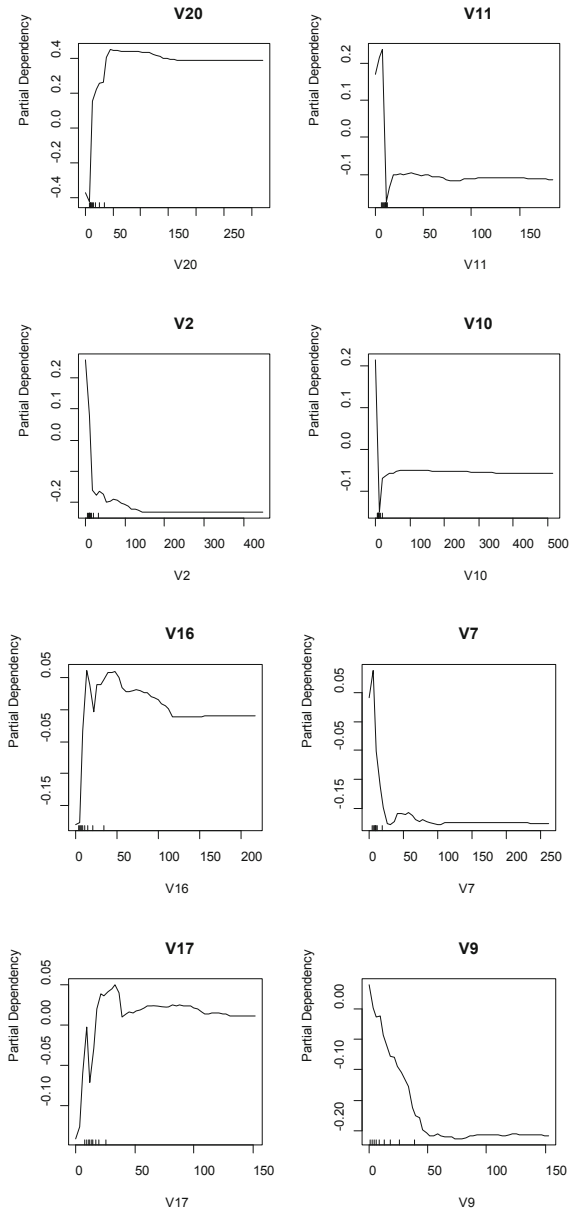
**Fig. 4.** The marginal effect of top (Amazon)

From Fig. 4, we found that the graph of variables V8, V15, V16, V18 and V20 were positively sloped curve. From this result, it speculated that variables of cluster 8, 15, 16, 18 and 20 were characteristics of class 1. Also, V6, V11 and V12 were negatively sloped curve. Therefore, they were characteristics of class 0.



**Fig. 5.** The marginal effect of top (general shopping sites)

From Fig. 5, we found that the graph of variables V1, V5, V14, V20 and V21 were positively sloped curve. From this result, it speculated that variables of cluster 1, 5, 14, 20 and 21 were characteristics of class 1. Also, V7, V12 and V15 were negatively sloped curve. Therefore, they were characteristics of class 0.



**Fig. 6.** The marginal effect of top (fashion shopping sites)

From Fig. 6, we found that the graph of variables V16, V17 and V20 were positively sloped curve. From this result, it speculate that variables of cluster 16, 17 and 20 were characteristics of class 1. Also, V2, V7, V9, V10 and V11 were negatively sloped curve. Therefore, they were characteristics of class 0.

Tables 6, 7, 8 and 9 show the confusion matrix and the evaluation indicator for the training data and the test data of Amazon.

**Table 6.** Confusion matrix of Amazon model for the train data

|              |          | Predicted class |          |
|--------------|----------|-----------------|----------|
|              |          | Positive        | Negative |
| Actual class | Positive | 364             | 218      |
|              | Negative | 184             | 398      |

**Table 7.** Evaluation indicator of Amazon model for the train data (%)

| ACC  | PRE  | REC  | F-measure |
|------|------|------|-----------|
| 65.5 | 64.6 | 68.4 | 66.4      |

**Table 8.** Confusion matrix of Amazon model for the test data

|              |          | Predicted class |          |
|--------------|----------|-----------------|----------|
|              |          | Positive        | Negative |
| Actual class | Positive | 99              | 71       |
|              | Negative | 44              | 115      |

**Table 9.** Evaluation indicator of Amazon model for the test data (%)

| ACC  | PRE  | REC  | F-measure |
|------|------|------|-----------|
| 65.0 | 61.8 | 72.3 | 66.7      |

From Tables 6, 7, 8 and 9, model of Amazon had well balanced evaluation indicator. Tables 10, 11, 12 and 13 show the confusion matrix and the evaluation indicator for the training data and the test data of general shopping sites.

**Table 10.** Confusion matrix of general shopping model for the train data

|              |          | Predicted class |          |
|--------------|----------|-----------------|----------|
|              |          | Positive        | Negative |
| Actual class | Positive | 317             | 183      |
|              | Negative | 167             | 333      |

**Table 11.** Evaluation indicator of general shopping model for the train data (%)

| ACC  | PRE  | REC  | F-measure |
|------|------|------|-----------|
| 65.0 | 64.5 | 66.6 | 65.6      |

**Table 12.** Confusion matrix of general shopping model for the test data

|              |          | Predicted class |          |
|--------------|----------|-----------------|----------|
|              |          | Positive        | Negative |
| Actual class | Positive | 117             | 77       |
|              | Negative | 31              | 104      |

**Table 13.** Evaluation indicator of general shopping model for the test data (%)

| ACC  | PRE  | REC  | F-measure |
|------|------|------|-----------|
| 67.2 | 57.5 | 77.0 | 65.8      |

From Tables 10, 11, 12 and 13, model of general shopping sites had well balanced evaluation indicator as with model of Amazon.

Tables 14, 15, 16 and 17 show the confusion matrix and the evaluation indicator for the training data and the test data of fashion shopping sites.

**Table 14.** Confusion matrix of fashion shopping model for the train data

|              |          | Predicted class |          |
|--------------|----------|-----------------|----------|
|              |          | Positive        | Negative |
| Actual class | Positive | 524             | 119      |
|              | Negative | 98              | 547      |

**Table 15.** Evaluation indicator of fashion shopping model for the train data (%)

| ACC  | PRE  | REC  | F-measure |
|------|------|------|-----------|
| 83.2 | 82.1 | 84.8 | 83.4      |

**Table 16.** Confusion matrix of fashion shopping model for the test data

|              |          | Predicted class |          |
|--------------|----------|-----------------|----------|
|              |          | Positive        | Negative |
| Actual class | Positive | 160             | 89       |
|              | Negative | 28              | 52       |

**Table 17.** Evaluation indicator of fashion shopping model for the test data (%)

| ACC  | PRE  | REC  | F-measure |
|------|------|------|-----------|
| 64.4 | 36.9 | 65.0 | 47.1      |

From Tables 6, 7, 8, 9, 10, 11, 12 and 13, models of Amazon and general shopping sites had well balanced evaluation indicator, however, from Tables 14, 15, 16 and 17, we found that prediction of fashion shopping sites' model tend to negative, precision was relatively low.

## 5 Discussions

In this section, we discuss about the results of RF and consider how to advertise on TV and websites.

As shown in Figs. 1, 2 and 3, we found that variables of cluster had high importance in all model. From this result, we confirmed that we have to consider media exposure of users to advertise efficiently.

Also, as shown in Figs. 4, 5 and 6, it turned out that some common variables were very important among these 3 target websites. First, we found that users of target websites often check review sites and cosmetics shopping sites (V20). It seems that they check reviews about items before they buy them. Therefore, we considered that the advertisement in the review sites was reached easily. Additionally, cosmetics' advertisement will make monitors interested certainly because they also check cosmetic shopping sites. In addition, I found that V12 was top of importance in Amazon and general shopping sites, V11 and V16 were top in Amazon and fashion shopping sites, V7 was top in general shopping sites and fashion shopping sites.

Also, each of target websites had different characteristics. V8 and V18 were characteristics of class 1 in Amazon. V8 is cluster that check the SNS. Therefore, the advertisement used SNS like Twitter, Facebook and Instagram will be effective. Furthermore,

we found that Amazon's user often use Yahoo! (V18). According to this result, listing advert on Yahoo! will exposure a lot of potential customer. On the other hands, V6 was characteristic of class 0. V6 is cluster that watching entertainment program. From this results, we confirmed that Amazon's users normally do not watch entertainment program. Hence, it may not be expected that commercial on entertainment program is effective.

In general shopping sites, V1, V5, V14 and V21 were characteristics of class 1. V5 is "check Rakuten shopping site" cluster and V14 is "check the fashion shopping sites" cluster. From cluster 5 and 14, we confirmed that general shopping sites' users check other shopping sites as well. Therefore, it is necessary to advertise good price to wide variety of items and so on to differentiate. On the other hands, V15 was characteristic of class 0. V15 is cluster that check the Google and news/weather forecast sites. We found that general shopping sites' users did not use Google relatively. From this results, we assumed that we have to advertise on portal site except Google mainly.

In fashion shopping sites, V17 was characteristic of class 1, and V2, V9 and V10 were characteristics of class 0. V2 is "check the Amazon and home appliances shopping sites" cluster and V17 is "check the finance service sites and foods shopping sites" cluster. From cluster 2 and 17, users of fashion shopping sites users tend to check various shopping sites. Therefore, it is important to advertise assortment unique to fashion shopping site. Moreover, V9 is cluster that watch news programs in the evening or night. From cluster 9, we found that fashion shopping sites' users did not watch news program in the evening or the night. Therefore, we assumed that we had to reduce the ratio of advertisement in news program in the evening or the night.

## 6 Conclusion

In this study, from media exposure data, we found potential media exposure patterns through NMF. Then, we constructed RF using result of NMF and monitors' demographic data to identify the characteristics of the users' media exposure pattern. From our result, we found some TV programs and websites that Amazon, general shopping sites and fashion shopping sites had to advertise.

In the RF model constructed in this research, although it is possible to clarify the characteristics of media exposure pattern, the accuracy of the model is not very satisfactory. Therefore there is room for improvement.

Moreover, in this study we found media exposure patterns through NMF. However it is difficult to find more than 3 characteristics. Therefore we need to consider more information like day by using tensor decomposition method for our future work.

**Acknowledgment.** We thank INTAGE Holdings Inc. for permission to use valuable datasets and for useful comments. This work was supported by JSPS KAKENHI Grant Number 19K01945 and 17K13809.

## References

1. Institute of Media Environment. Media Fixed Point Survey 2019 (2019). (in Japanese)

2. Ministry of Economy, Trade and Industry. Results of FY2019 E-Commerce Market Survey (2019). (in Japanese)
3. Hasumoto, K., Kumoi, G., Goto, M.: A prediction of customer lifetime value in a platform business using nonnegative matrix factorization. *IPSJ J.* **60**(7), 1283–1293 (2018). (in Japanese)
4. Nagahashi, K.: Introduction of Machine Learning with R. Impress (2017). (in Japanese)
5. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman & Hall, New York (1984)
6. Brunet, J., Tamayo, P., Golub, T., Mesirov, J.: Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci.* **101**(12), 4164–4169 (2004)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)





# Dynamic Properties of Information Diffusion Networks During the 2019 Halle Terror Attack on Twitter

Philipp Kessling<sup>(✉)</sup>, Bastian Kiessling, Steffen Burkhardt, and Christian Stöcker

Department of Information Competence Center Communication,  
Hamburg University of Applied Sciences, Hamburg, Germany  
{philipp.kessling,bastian.kiessling,steffen.burkhardt,  
christian.stoecker}@haw-hamburg.de

**Abstract.** On 9th October 2019 an armed nationalist murdered two bypassing citizens in a killing spree directed at a synagogue in Halle, Germany. Instantly, a broad diffusion of information unfolded on Twitter. Traffic for tweets mentioning Halle reached a peak as high as a hundred times to average value of the days before. In this study we examine this immense increase in communication and observe temporal diffusion patterns in crisis communication. We compare the traffic of tweets generated in this incident against Twitter traffic persisted during two other crisis events as well as a regular trending topic. A discussion of information diffusion based on network theoretic measures during crisis and terror events is presented. Results show that active user's behavior changes with the onset of the incidents in all events in focus—while popular user's metrics are consistent throughout the data sets. actively tweeting and repeatedly engaging users are detected only in two data sets.

**Keywords:** Information diffusion · Network dynamics · Terror attack

## 1 Introduction

Crisis and emergency situation events such as natural disasters and terror acts result in an immediate and massive response in social media. In these kinds of situations, Twitter is relevant due to its fast-paced nature, especially during developing events [5, 13]. The microblogging service has become a platform for breaking news, including the coverage of crises and disasters. It provides an indicator of emerging and developing events on a national and international scale with its trending topic feature. While a crisis is defined as an immediate threat to people or organizations and their key values or functions, hence, leading to situations of uncertainty [22], terrorism is politically motivated violence against noncombatants. The latter is trying to create an additional fearful state of mind among the public [23]. The user's motivation to contribute to a discussion regarding a terror attack ranges from sharing relevant information [14]

and expressing condolences to spreading anger and fear [12]. Twitter is furthermore an important communication channel for security forces and emergency authorities to create situational awareness: in times of crises the public needs to be informed and instructed about ongoing endangerments and to avoid panic. Twitter's environment, in which information can be efficiently shared, provides an opportunity for all participating parties to spread their messages during crisis [4]. Traditional mass media such as newspapers or TV channels are no longer in control of information dissemination.

A majority of research regarding crisis communication in social media analyzes public response and user behavior (e.g. [17, 19, 25]) or communication efforts and crisis management from an authoritative perspective (e.g. [4, 12, 31]). While some work focuses on the characteristics of information dissemination in crisis situations [7, 30], there still is a research gap in the specifics of information diffusion in social media networks during terror attacks. Especially, the dynamic properties in the diffusion process needs to be further studied. In this case study, we analyze information diffusion in the aftermath of terror attacks in form of retweet networks—Twitter's retweet function allows to share information effortlessly,—thus, resulting in large proportions of the total traffic being comprised by this replication-based content. We conduct the study on a total of four data sets, three of those connected to terror and crises and another trending topic on Twitter.

In Sect. 1, we commence with a literature overview regarding the fields of social media response to terror and crises events and information diffusion processes in social media. Thus, we deduce research questions thereof in Sect. 2. In Sect. 3, we present our methodology and data sets, discuss the results in Sect. 4 and conclude in Sect. 5.

### 1.1 The Aftermath of Terror Attacks and Crises Situations on Twitter

The ever-increasing usage and mass adoption of social media has reshaped the perception of news-worthy events: now, social media platforms have become platforms for breaking news which had previously been the playing field of mass media. Events such as terror acts or natural disasters typically result in an immediate response. Social media, especially Twitter, provide important and useful eyewitness information in the aftermath of crises [17, 19]. Twitter is extremely relevant in this context due to its real-time communication environment that allows users to quickly spread information about developing events [4]. The service provides a trending topic feature as an indicator for emerging events on a national and international scale. According to Eriksson Twitter can serve as a platform to establish counter themes contrasting mainstream media [10], while it may also serve as an environment for mainstream media during crisis [26].

Crisis events take a special role due to their abrupt appearance, leading to an emerging interest in communication research. Studies in this field focus on the potential of Twitter for authorities to handle crisis event communication [4, 32] as well as communication processes between users aiming to understand

the impact of crises on the public [30]. A majority of previous studies analyze the impact of single instances of terrorist attacks, while few compare different terror acts or a variety of crisis events [6, 19].

A common characteristic of information cascades during terror attacks is the immediate and steep increasing rate of messages referencing the event. Within a few days, the tweet volume in connection to the event strongly decreases, returning to the original message frequency [6]. The rising utilization of social media has increased the number of users who participate in discussions about crisis events [4]. While the dissemination of URLs in tweets increases more slowly, these users retweet more often and increasingly use hashtags. Toriumi et al. found that users change their behavior by spreading more situation-related information and by decreasing non-related tweets during crises to avoid interrupting important information [30]. On the other hand, Buntain et al. stress that terror events do not significantly impact the general usage of Twitter [6]. Crisis events are commonly discussed by referring to hashtags that are adapted by the participating users over time. These hashtags are used as geographical and temporal markers (e.g. #halle, #hl0910) to show support, resilience and cohesiveness [25, 26].

People use social media to make sense of the event, validate their worldview and keep their self-esteem. Calls for tolerance as well as nationalistic views can be found, compared to hostility toward different values expressed by other participants in the online debate. Overall, people tend to express more worry and support for the victims instead of spreading anger and fear. Especially in the first days following a terror attack, information sharing and positive social behavior dominate the connected discussion [12].

The chaotic situation caused by terror attacks typically leads to a lack of information and understanding of the event. To counter unverified information or even disinformation circulating on social media users, media and authorities constantly provide updates about the ongoing event [25]. Official government organizations such as the police use Twitter to create situational awareness in the aftermath of terror attacks. These accounts are important for reducing the impact of rumors and misinformation during crisis events [4]. Different actors suggest avoiding the spread of rumors and misinformation [12], pointing out the achievement of situation awareness as one of the key challenges during crises for all involved parties. News media and authorities emerge as central actors in social media debates following crises and terror events, while authorities as well as non-governmental organizations are less active in crisis situations than media actors, yet their published content is shared more often [31]. If authorities such as police departments are not participating on social media, the information vacuum is likely to be filled by the media [6]. Unexpected accounts such as local politicians, institutions or media can also become central communicators [26] due to a geographical information advantage.

## 1.2 Information Diffusion in Twitter Networks

Crises and terror events increase the motivation for social media users to seek and share information on social media, and, as mentioned above, especially use

Twitter [14]. In crisis discussions, the retweet ratio (number of retweets against the number of all tweets in the data set) significantly increases [24]: users share information with their followers by publishing original content or retweeting content. The diffusion of information on Twitter is mainly governed by the social network’s features [11]—who follows, i.e. who gets whose information. This rapid information diffusion is one of the central elements of the microblogging service [26] and crucial to reduce the uncertainty caused by a crisis situation.

A series of retweets, or retweet cascades, often involves a burst, a sharp spike in retweet frequency while the cascading process itself may run for weeks [11]. Taxidou et al. define retweets as explicit interactions [29], inducing retweet cascades that spread information through the network [8]. Next to favorites, retweets are the main factor to measure the popularity of a single tweet. The ratio of retweets to tweets is an indicator of spreading topics [28]. Further connections can be created through the mention and reply functions, where users highlight others by mentioning their usernames or comment on content by replying to tweets. Yoo et al. found that information originating from within the network spreads faster than information from external sources [32].

Burnap et al. [7] analyzed the flow of information following the Woolwich terrorist attack that took place in London in 2013. They define information flow as information spreading via the retweet function by considering the frequency of retweeting (size) and the duration between the first and the last retweet (survival). The authors find that negative content is shared less often than positive and supporting content. Brief time intervals between retweets, the use of URLs and hashtags as well as positive sentiment of the original content are constructive predictors of the information flow following crisis events. However, according to the study, the most influential factor for a fast information flow is the popularity of the author measured by the number of followers and tweets. Information published by influential users diffuse more quickly. That same pattern was observed for information that was published at an earlier stage of the crisis [32]. For an extensive overview of the properties of the retweet network refer to [3].

## 2 Research Questions and Contributions

From information diffusion theory and previous studies in the field, we learn that information diffusion generally adheres to what is called *bursty* behavior [2, 15, 18]. Information is spread in networks along ties between nodes—in our case Twitter accounts. Here, these ties render to the “follower” and “followee” relationships between accounts. Information is passed between the accounts according to the accordance of the platform: tweets are original pieces of information posted by users, retweets are unaltered replications of tweets and other interactions include commenting and replying to tweets with another tweet and lastly tagging or mentioning other accounts in one’s tweet. Additionally, users have the possibility to like tweets, a function which was originally implemented to bookmark tweets, but is now a popularity index. In total, a large share of the interactive possibilities depends on retweets and likes—as these interactions are

most accessible to users, requiring little work to execute [27]. Previous research has shown that the vast majority of traffic is published in the first few days following the event [6]. This effect is even stronger in this context, as crisis events are most commonly instants in time, leading to a sharp rise in tweet-traffic.

A question that has been left vacant in the above listed research, is how users engage in discussions in the aftermath of terror attacks: are the information cascades driven by a multitude of users which engage singularly or rather a small and active set of users? Secondly, how do users engage with one another in communication streams connected to terror attack?

Our contributions are the following:

(a) We explore the retweet network created alongside of crisis and terror events in a network theoretical manner. We base our discussion jointly on user activity and communication authority. We employ standard graph measures to determine topographic properties of user-based interactions.

(b) While the former presented research analyzed information networks during crisis events as a whole, we argue that longitudinal networks that evolve over time represent the dynamic process of information diffusion more accurately. This process potentially changes the structure of the social network and is also strongly affected by emerging topics [18].

### 3 Data Samples and Method

On 9th October 2019, a shooting took place in Halle, the largest city in the state Saxony-Anhalt in Germany. The terrorist tried and failed to enter the synagogue during the service on the Jewish holiday of Yom Kippur, instead killing two uninvolved people with self-built firearms at another location. He was identified as a male 27-year-old male, who had an antisemitic and nationalistic motivation for the crimes committed, according to the federal investigators. The attacker streamed his actions online, displaying his weapons, the murders committed and his rationale as a denier of the Holocaust.

In order to analyze the impact of and information diffusion during the terror attack in Halle we extracted data via Twitter’s search API with the search term *Halle*. We gathered tweets spanning three days before and three days after the activity peak of the Halle shooting, resulting in a sample of 518,922 tweets from 187,603 active users. To give a broader insight on the issue of the information diffusion dynamics, we compare the Twitter traffic surrounding the (a) Halle shooting with three other, distinct events: (b) a clash between police riot forces and radical-left protesters; (c) another shooting where a man killed his family members and; (d) as a baseline example for a trending topic New Year’s Eve 2019.

The second event happened in the night of 1st January 2020, when protesters clashed with police forces in the alternative left Leipzig district of Connewitz (search term “*Connewitz*”). It initiated a debate on left-wing violence in Germany and trust and confidence issues in police forces.

**Table 1.** Overview of the data sets: user and tweet counts are unique entities per data set, the events starting time.

| Incident  | # Tweets | # Users | $t_0$            |
|-----------|----------|---------|------------------|
| Halle     | 518,922  | 187,603 | 2019/10/09 12:00 |
| Connewitz | 48,806   | 14,113  | 2020/01/01 09:00 |
| Silvester | 26,786   | 17,660  | 2019/12/31 23:59 |
| Rotamsee  | 13,239   | 9,267   | 2020/01/24 12:00 |

The third data set covers a familicide that took place on 24th January 2020 in Rot am See, a small South-West German city (search term “*Rot am See*”). A 26-year-old man killed six of his family members, including his parents. While his motivation is still under investigation the incident immediately aroused conversation traffic on social media.

Finally, tweets in connection with the German-language debate on New Year’s Eve 2019 were finally gathered to include a recurring event in the analysis (search term “*Silvester*”). Table 1 gives an overview of all data sets including the number of gathered tweets and involved users for each case as well as the center of activity. We restricted the data sets to same time intervals and defined a temporal zero to ensure comparability between the four different events. In total we persisted and analyzed a sum of 607,753 tweets from 228,643 active users.

For each of the four instances of attacks and trending topics, we limited the data set to an extent of three days before and after the initial action. We determined the individual starting point  $t_0$  as the point of origin for each information cascade—for the terror attacks it is the known time of the first attack, in the case of the trending topic it is New Year’s Eve. As the retweet network is the outcome of the information diffusion process, we investigate the questions laid out beforehand, by an analysis of the retweet graph implicitly contained in our data sets. At first, we extract a dynamic network from each data set in order to examine how information diffusion is shaped during such a crisis event.

For the last ten years, retweeting behavior has been a subject of researcher’s scrutiny: a number of studies have been conducted on the general features of these networks as well as predicting single retweet cascades. The most approachable tool is a static network analysis. Albeit powerful, it is not feasible for our set of questions as the nature of the processes behind the user communication is suspected to be highly dynamic. Thus, viewing the diffusion network from a static point of view would yield an inaccurate representation.

We solve this problem with the utilization of longitudinal networks—a series of networks, each representing the status of the retweet stream at the end of each time bin. We partition the data sets into rectangular time windows to obtain these series of static networks. The window size we use as a default is one day, windows overlap 48 times, thus, the distance between windows amounts to 30 min. This wide window size has the advantage of cancelling out circadian

effects in tweet frequency. In the following we state the indicators employed in our study and how these are linked to the observed phenomena to interpret their results:

**Network Size and Order.** Navigating uncanonical and partially contradicting nomenclature, we doubt the number of edges in a network order and the network's number of nodes, network size. We are not only interested in their absolute values and also take their ratio into account.

**Degree Distributions.** In our case, the dynamic network represents retweet interactions between Twitter accounts, following Bild et al. [3] we assume this network to be scale-free [1]. Since multiple interactions between accounts are possible, we obtain directed multigraphs, relaxing structural cutoff constraints. For each network, we determine the degree distribution for incoming as well as for outgoing directed edges and their power law coefficients. The degree distribution yields information on the structure of the network—we are especially interested in how dominant hubs (the most active and popular accounts respectively) are emerging in the network.

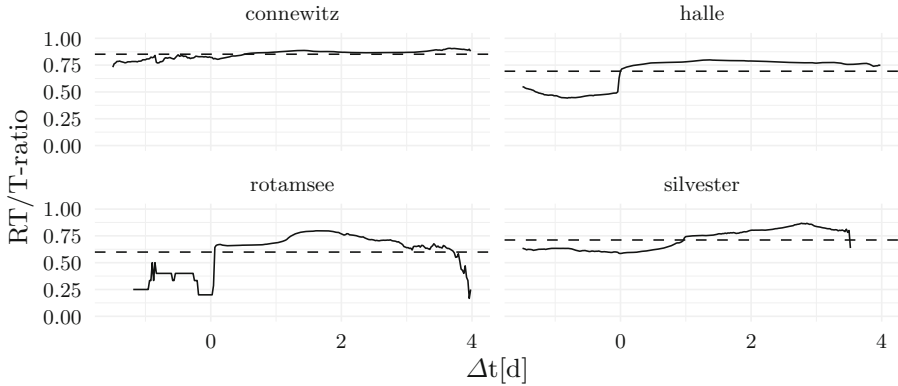
**Degree Dynamics.** We determine for each node the change of its degree over time as the network is growing. To uncover fast growing accounts, we determine the number of tweets in the duration from a given node's first to its last tweet in the data set. This average tweet frequency and its distribution give interesting insight into the account's behavior and engagement: who are the most-active accounts, as well as the most popular—when do they start? We calculate a node attribute which determines the temporal degree dynamics: we utilize the average tweet frequency for each account. We also calculate this measure in a directed manner as well. Thus, determining an average activity and popularity score.

**Degree Correlation.** Assortativity is the measure of correlation between the connecting node's incoming and outgoing degrees. Here, we apply Spearman's rank correlation, as suggested by [16] to measure assortativity between networks of different sizes.

All analysis was done with R 3.6.1 [21], utilized packages include igraph [9] and tidygraph [20].

## 4 Results and Discussion

The incident in Halle, Germany on 9th October 2019 created an instant and immense burst in Twitter traffic containing the keyword *Halle*. The tweet frequency increased from 100 t/h to a maximum of 25,000 t/h when the attacker was arrested by police forces. The three other incidents created less traffic by far, although the incident in Rot am See generated a peak of 5,000 t/h, while being the smallest data set in terms of total tweets. All three incident-based data sets (a, b, c) exhibit a large burst in traffic—a feature the New Year's Eve data set (d) lacks. All data sets contain a large share of replication-based traffic, i.e. retweets. As the overall tweet frequency rose, the RT-ratio of the overall traffic



**Fig. 1.** Retweet-Tweet ratio for each network.

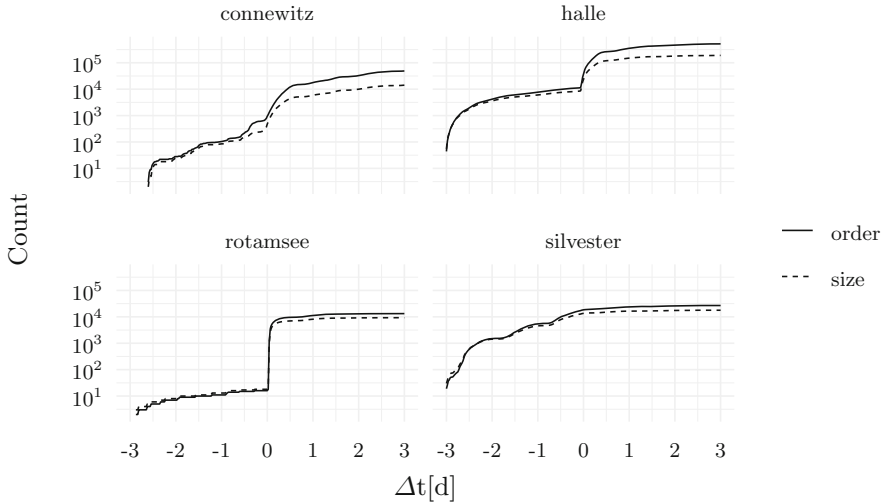
increased by 25%, as well, averaging over the entire time span 69.3%, with 48.8% before and 77.6% after the attack, see Fig. 1.

For data set (a) police forces—both the Halle police department, as well as Saxony state police—were initially retweeted the most, both in absolute amount, as well as retweet frequency. The local police Twitter account accumulated over 7,000 retweets in first 24 h after the attack. Other accounts, whose tweets were virally retweeted fall into distinct categories: Twitter personalities, organizational accounts (e.g. @ZentralratJuden, the Central Council of Jews in Germany) and celebrities. Overall the degree trajectories of these accounts show rapid increases—a few even multiple increases, due to multiple tweets.

Figure 1 gives the RT/T-ratio—defined as number of retweets in the data set divided by the total number of tweets—for the four data sets over time, as well as the average RT/T-ratio for the entire data set. An RT/T-ratio of 1.0 indicates traffic consisting only of retweets while 0.0 would indicate no present retweets. For all data sets, the RT/T-ratio is quite high with averages for (a) 69.3%, (b) 85.2%, (c) 59.9% and (d) 71.1%. Data sets (a, c) show a transient with data set (a) increasing by approx. 25% at  $t_0$  and (c) increasing by nearly 50% at  $t_0$  with the limitation that for (c) the traffic before  $t_0$  was comprised of singular tweets. Interestingly, data set (b) shows an already high RT/T-ratio before  $t_0$ , as does (d) where such behavior was expected.

Figure 2 shows the accumulated networks size and order—the number of nodes and edges, respectively. The incidents in Halle as well as Rot am See (a, c) display a massive surge of tweets during the events onset, while passing  $t_0$ . The incident in Connewitz (b) exhibits this phenomenon in a blurred fashion. The data set (d)—a suspected *normal* trending topic—does not show such transient behavior, but rather a continuous growth, which slows after the passing  $t_0$ . An explanation is surely the events background and overall importance: the antisemitic incident in



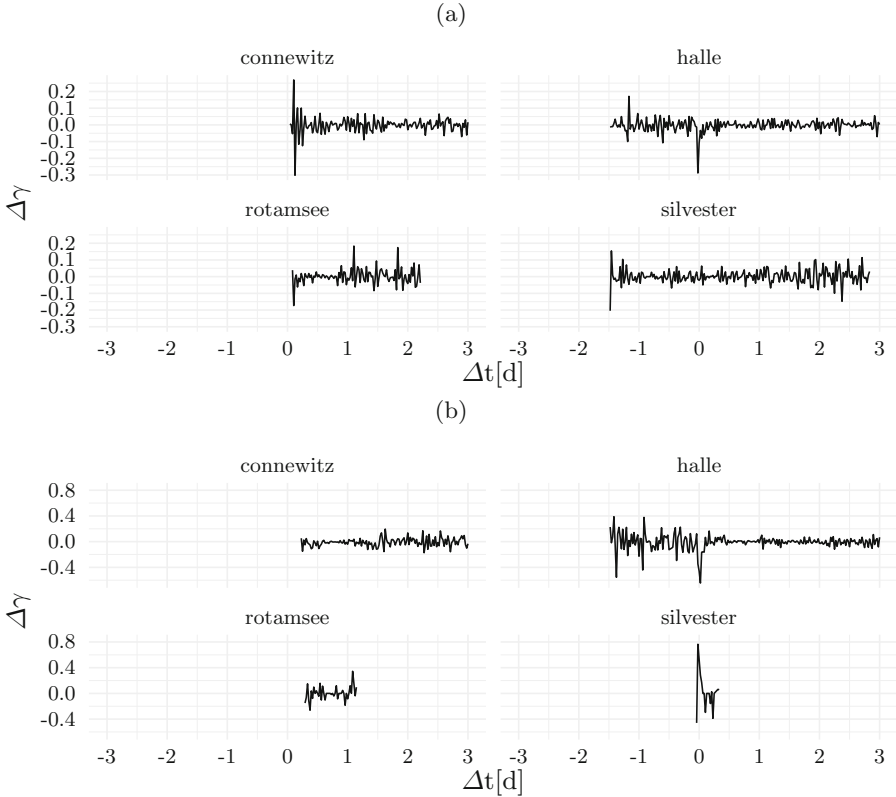


**Fig. 2.** Network size (number of nodes) and order (number of edges) per window, window size is 30 min, overlap 0x

Halle was an event of global scale, causing reactions and news reports from international news agencies and associations. In data sets (a) and (b) few users replicate tweets in large numbers. In contrast data sets (c) and (d) show no sign of such hyperactive users. Given that the crass political impact of (a) and the discussion surrounding (b) and the fact that (d) is a benign base-line example, we assume this a sign of attempted manipulation. Another indicator is the ratio of the networks order to the network size, giving an estimate of overall repeated user activity. Incidents (a) and (c) exhibit a large gap between the total number of interactions and the users involved in this, hinting at repeated user interaction. Data sets (b, d) on the other hand show only a much smaller gap, thus, less repeated user interactions.

A networks structure can be determined from its degree distribution: Fig. 3 indicates the degree distribution's power law exponent's change over time. Change in the network's out degree distribution indicates a change in the way users interact and replicate tweets. A decline in the coefficient's value indicates a larger proportion of low degree nodes. For data set (a), we detect a sudden change in the in degrees and out degrees, both decreasing by 0.3 and 0.6 respectively, indicating increased traffic from low degree users. In data sets (b) and (c) the onset similarly exhibits a spike, although the data sets limitation to nodes of a degree  $\geq 10$  may induce such artifacts.

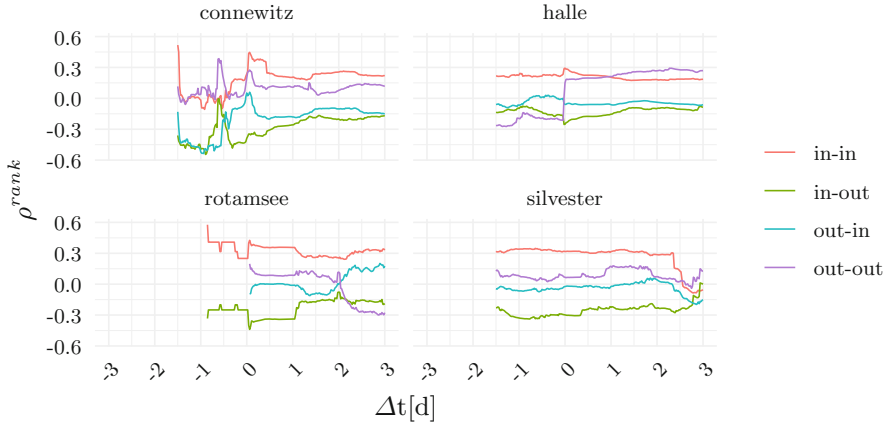
Figure 4 shows the pairwise directed degree correlation, e.g. *in-out* indicates the Spearman correlation between the users in degree and the retweeted users out degree. Data set (a) shows an interesting change at  $t_0$ : the correlation between out degree and the connected node's out degree increases radically, crossing



**Fig. 3.** (a) Estimated power law coefficient's change for the networks in degree distribution, (b) The same for the out degree distribution.

zero. Our base line data set (d) shows quite a different picture. All four assortativity measures are quite stable over time—even New Year's Eve,  $t_0$ , passes without a significant change in the network's topology. Overall, the correlation between in degree and in degree appear positively correlated and in degree and out degree negatively correlated, suggesting assortativity in the following manner: an account with high in degree will retweet another node with high in degree, while in the same instance the second node will have a low out degree. This confirms common assumptions of the behavior of popular accounts.

To extract temporal features of the information diffusion process during terror attacks, we obtain the average tweeting frequency for every account in the data sets. That being the number of tweets divided by the duration from first to the last tweet, respectively retweet. For each data set and edge direction, Fig. 5 shows the ECDF for the average frequency. The results for incoming edges—accounts being retweeted—is expectedly quite similar among the data sets.



**Fig. 4.** The networks assortativity per time bin.

We attribute this to the affordances of the platform. Interestingly, for the outgoing edges we are only able to calculate results for data sets (a) and (b). Accounts in this data sets were not multiple times, thus, not exposing a degree change over time. Contrary in data sets (a) and (b) we find a subset of accounts that massively contribute in an active manner. Top tweeters reached final degrees of maximally 666 tweets in three days for data set (a) and 409 tweets for data set (b). Data set (c) yielded a top max. final degree of 61.

In summary, data set (a)—the social media response to a nationalistically motivated shooting—yields insight into information diffusion in terror attacks. As events unfolded in Halle, traffic mentioning the city’s name rose almost instantly by factor 1,500. After this initial spike, traffic slowly declined over days as predicted by previous studies. Studying the degree dynamics present in the longitudinal network, we find that the popular and retweeted subset of the account population is similarly in structure as suggest by previous studies. Single accounts gained the most popularity and are connected to local authorities and media. Highly active accounts are present in the sample: the activity of 25 most active accounts accumulates to 6,343 tweets in three days. Comparing data set (a) to data sets (b, c, d) yields interesting observations. In data sets (c, d) there are no highly active users present. We speculate that is due to a lessened political impact and importance of both events. An inversion of this argument stands out: in both (a) and (b) highly active users are present and data set (c) shares many characteristics with data set (a) in terms of absolute message frequency rise in the onset (see Fig. 2), average message frequency distribution (see Fig. 5).

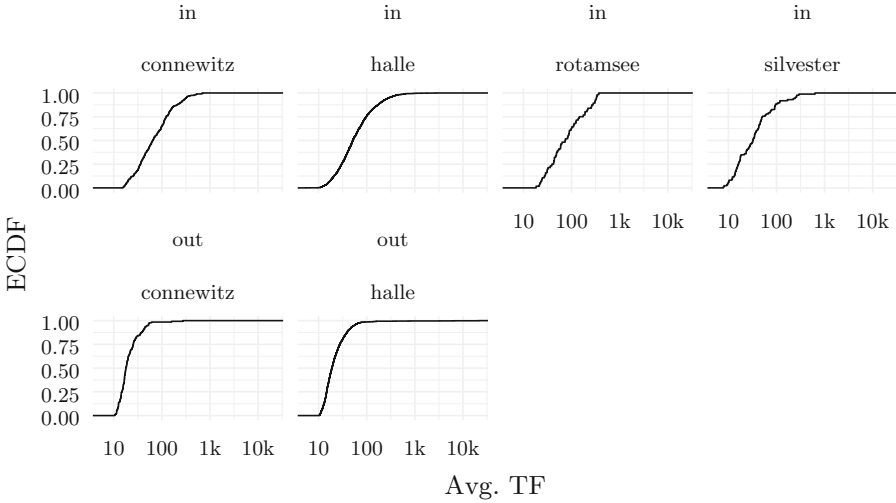


Fig. 5. ECDF of each account’s average tweeting frequency.

## 5 Conclusion

In this case study, we retrieved and analyzed a set of over 500,000 tweets using the keywords *Halle*, *Connewitz*, *Rot am See* and *Silvester* parallelly to three terror attacks and crisis events as well as one trending topic. We compared the traffic generated in the incident in Halle, Germany, against traffic persisted during two other crisis events as well as a regular trending topic. A discussion of information diffusion during crisis and terror events, based on network theoretic measures, is presented.

Results show that a high level of retweet-based traffic is present in all data sets, with RT/T-ratios consistently  $\geq 0.6$ . Active user’s retweeting behavior changes with the incident onset in all persisted events: on the onset of the incident, participating users start to share more information instead of publishing their own content during terror attacks and crisis events. This is consistent with previous research [30], whereas users spread more situation-related information and decrease non-related content to avoid the flow of relevant information. Local authorities—such as police forces and local journalists and newsrooms—also arise as central actors on social media in the aftermath of terror events, as already proposed by Wang and Zhuang [31]. The dynamic network’s size and order rise sharply, as interactions in form of retweets immediately increase with the onset of the crisis events. A massive network growth was observed especially in the aftermath of the Halle terror attack and the shooting in Rot am See. It is conceivable that both events were interpreted similarly on social media before the latter incident turned out to be family tragedy without a political motive behind it. The massive surge of the network size regarding the Halle terror attack can be explained with the incident itself, as an event of global scale that resulted in

a traffic spike and hyperactive users that furthermore increased the network size by actively replicating tweets.

Further studying the dynamic networks, we obtained in this study, we observe that for the passive side of each network—things user endure or experience due to others user's behavior—our indicators are quite consistent for all data sets, including our baseline example. In contrast, the active side—the behavior of single users—shows differences, especially in data sets connected to the incidents in Halle and Connewitz. This can be attributed to the presence of hyper-active accounts in these both data sets as well as the absence of higher degree nodes in two other data sets.

## References

1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47 (2002)
2. Barabási, A.L.: The origin of bursts and heavy tails in human dynamics. *Nature* **435**(7039), 207–211 (2005). <https://doi.org/10.1038/nature03459>
3. Bild, D.R., Liu, Y., Dick, R.P., Mao, Z.M., Wallach, D.S.: Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. *ACM Trans. Internet Technol.* **15**(1), 1–24 (2015). <https://doi.org/10.1145/2700060>
4. Bruns, A., Burgess, J.: Crisis communication in natural disasters: The Queensland floods and Christchurch earthquakes. In: Weller, K., Bruns, A., Burgess, J., Mahrt, M., Puschmann, C. (eds.) *Twitter and Society*, vol. 89, pp. 373–384. Peter Lang, New York (2014)
5. Bruns, A., Hanusch, F.: Conflict imagery in a connective environment: Audio-visual content on Twitter following the 2015/2016 terror attacks in Paris and Brussels. *Media Cult. Soc.* **39**(8), 1122–1141 (2017). <https://doi.org/10.1177/0163443717725574>
6. Buntain, C., Golbeck, J., Liu, B., LaFree, G.: Evaluating public response to the Boston marathon bombing and other acts of terrorism through Twitter. In: Tenth International AAAI Conference on Web and Social Media, March 2016
7. Burnap, P., et al.: Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Soc. Netw. Anal. Min.* **4**(1), 1–14 (2014). <https://doi.org/10.1007/s13278-014-0206-4>
8. Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., Leskovec, J.: Can cascades be predicted? In: Proceedings of the 23rd International Conference on World Wide Web - WWW 2014, Seoul, Korea, pp. 925–936. ACM Press (2014). <https://doi.org/10.1145/2566486.2567997>
9. Csardi, G., Nepusz, T.: The igraph software package for complex network research. *InterJournal Complex Syst.* **1695**(5), 1–9 (2006)
10. Eriksson, M.: Managing collective trauma on social media: The role of Twitter after the 2011 Norway attacks. *Media Cult. Soc.* **38**(3), 365–380 (2016). <https://doi.org/10.1177/0163443715608259>
11. Farajtabar, M., Wang, Y., Gomez-Rodriguez, M., Li, S., Zha, H., Song, L.: COE-VOLVE: A joint point process model for information diffusion and network evolution. *J. Mach. Learn. Res.* **18**(1), 1305–1353 (2017)

12. Fischer-Preßler, D., Schwemmer, C., Fischbach, K.: Collective sense-making in times of crisis: Connecting terror management theory with Twitter user reactions to the Berlin terrorist attack. *Comput. Hum. Behav.* **100**, 138–151 (2019). <https://doi.org/10.1016/j.chb.2019.05.012>
13. Hornmoen, H., Backholm, K. (eds.): *Social Media Use in Crisis and Risk Communication: Emergencies, Concerns and Awareness*, 1st edn. Emerald Publishing (2018). oCLC: on1032581075
14. Jin, Y., Fraustino, J.D., Liu, B.F.: The scared, the outraged, and the anxious: How crisis emotions, involvement, and demographics predict publics' conative coping. *Int. J. Strateg. Commun.* **10**(4), 289–308 (2016). <https://doi.org/10.1080/1553118X.2016.1160401>
15. Kleinberg, J.: Bursty and Hierarchical Structure in Streams, p. 25 (2002)
16. Litvak, N., van der Hofstad, R.: Uncovering disassortativity in large scale-free networks. *Phys. Rev. E* **87**(2), 022801 (2013). <https://doi.org/10.1103/PhysRevE.87.022801>
17. Liu, B.F., Fraustino, J.D., Jin, Y.: Social media use during disasters: How information form and source influence intended behavioral responses. *Commun. Res.* **43**(5), 626–646 (2016). <https://doi.org/10.1177/0093650214565917>
18. Myers, S.A., Leskovec, J.: The bursty dynamics of the Twitter information network. In: *Proceedings of the 23rd International Conference on World Wide Web, WWW 2014, Seoul, Korea*, pp. 913–924. ACM (2014). <https://doi.org/10.1145/2566486.2568043>
19. Olteanu, A., Vieweg, S., Castillo, C.: What to expect when the unexpected happens: social media communications across crises. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW 2015, Vancouver, BC, Canada*, pp. 994–1009. ACM Press (2015). <https://doi.org/10.1145/2675133.2675242>
20. Pedersen, T.L.: Tidygraph: A Tidy API for Graph Manipulation (2019)
21. R Core Team R: *A Language and Environment for Statistical Computing*. Vienna, Austria (2019)
22. Rosenthal, U., Boin, A., Comfort, L.K. (eds.): *Managing Crises: Threats, Dilemmas. Opportunities*. Charles C Thomas, Springfield (2001)
23. Ruby, C.L.: The definition of terrorism. *Analyses Soc. Issues Publ. Policy* **2**(1), 9–14 (2002). <https://doi.org/10.1111/j.1530-2415.2002.00021.x>
24. Sakaki, T., et al.: Regional analysis of user interactions on social media in times of disaster. In: *Proceedings of the 22nd International Conference on World Wide Web - WWW 2013 Companion, Rio de Janeiro, Brazil*, pp. 235–236. ACM Press (2013). <https://doi.org/10.1145/2487788.2487909>
25. Simon, T., Goldberg, A., Aharonson-Daniel, L., Leykin, D., Adini, B.: Twitter in the cross fire—the use of social media in the Westgate Mall terror attack in Kenya. *PLoS ONE* **9**(8), e104136 (2014). <https://doi.org/10.1371/journal.pone.0104136>
26. Steensen, S.: Tweeting terror: An analysis of the Norwegian Twitter-sphere during and in the aftermath of the 22 July 2011 terrorist attack. In: Hornmoen, H., Backholm, K. (eds.) *Social Media Use in Crisis and Risk Communication*, pp. 15–41. Emerald Publishing Limited, October 2018. <https://doi.org/10.1108/978-1-78756-269-120181006>
27. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network. In: *2010 IEEE Second International Conference on Social Computing, Minneapolis, MN, USA*, pp. 177–184. IEEE, August 2010. 10/dbpkkh

28. Takahashi, T., Igata, N.: Rumor detection on Twitter. In: The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems, Kobe, Japan, pp. 452–457. IEEE, November 2012. <https://doi.org/10.1109/SCIS-ISIS.2012.6505254>
29. Taxidou, I., Fischer, P.M., De Nies, T., Mannens, E., Van de Walle, R.: Information diffusion and provenance of interactions in Twitter: Is it only about Retweets? In: Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion, International World Wide Web Conferences Steering Committee, Montréal, Québec, Canada, pp. 113–114, April 2016. <https://doi.org/10.1145/2872518.2889393>
30. Toriumi, F., Sakaki, T., Shinoda, K., Kazama, K., Kurihara, S., Noda, I.: Information sharing on Twitter during the 2011 catastrophic earthquake. In: Proceedings of the 22nd International Conference on World Wide Web - WWW 2013 Companion, Rio de Janeiro, Brazil, pp. 1025–1028. ACM Press (2013). DOIurl-<https://doi.org/10.1145/2487788.2488110>
31. Wang, B., Zhuang, J.: Crisis information distribution on Twitter: a content analysis of tweets during Hurricane Sandy. *Nat. Hazards* **89**(1), 161–181 (2017). <https://doi.org/10.1007/s11069-017-2960-x>
32. Yoo, E., Rand, W., Eftekhari, M., Rabinovich, E.: Evaluating information diffusion speed and its determinants in social media networks during humanitarian crises. *J. Oper. Manage.* **45**(1), 123–133 (2016). <https://doi.org/10.1016/j.jom.2016.05.007>



# Cultural Factors as Powerful Moderators of Romanian Students' Adoption of Mobile Banking in Everyday Life

Valentin Mihai Leoveanu<sup>(✉)</sup>, Mihaela Cornelia Sandu, and Adela Coman

University of Bucharest, Bucharest, Romania

{valentin.leoveanu, mihaela.sandu, adela.coman}@faa.unibuc.ro

**Abstract.** The purpose of this research is represented by the need to highlight the trend regarding the access and use of mobile banking and mobile payment services by student consumers, in close connection with the cultural factors that can influence this trend. In this regard, the paper considers aspects such as observing the students' behavior regarding how mobile banking is used in the monitoring of bank accounts, taking financial decisions about saving and investments, as well as shopping. The research methodology addressed aimed at both qualitative and quantitative research, considering the mobile financial services offered by banks operating in Romania. The authors conducted a quantitative research based on a questionnaire, grouping selectively a multitude of questions on how students perceive mobile banking according to the specific interaction they had with it. The structural model that was used is called the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2), a combination of UTAUT developed in 2003 by Venkatesh et al. and Hofstede's cultural moderators. The qualitative analysis considered the placement of the factors analyzed in a cultural context linked with the Romanian students, trying to investigate the values, perceptions, desires, but also the stereotypes and prejudices taken from the family or other social organisms. The results of this research reveal how cultural factors determine the Romanian students attitude close to the use of mobile banking services. Factors that have a positive influence on behavioral intention are performance, facilitating conditions and habit. Habit has also a positive influence on use behaviour. Social influence has a negative effect on behavioral intention. There is no relationship between behavioral intention and effort expectancy, hedonic motivation and price value. Also, there is no association between use behavior and behavior intention. All cultural moderators signal a negative influence on the behavioral intention - use behavior relationship.

**Keywords:** Mobile banking · UTAUT2 · Cultural factors · Behavioral intention · Use behavior

## 1 Introduction

The multitude and breadth of the development of IT programs and applications that emerged as a result of the increasing trend of inventions and innovations in recent years

© Springer Nature Switzerland AG 2020

G. Meiselwitz (Ed.): HCII 2020, LNCS 12194, pp. 583–599, 2020.

[https://doi.org/10.1007/978-3-030-49570-1\\_41](https://doi.org/10.1007/978-3-030-49570-1_41)



has had no way of not reflecting on the field of financial services, with a particular impact on the banking businesses regarding the adaptation process to the client needs and the competitive evolution of the market. Not least, Romania, as a Member State of the European Union under the influence and in direct interaction with the evolutions on this unified market, has experienced in recent years a significant development in terms of the diversification of banking services, especially online and mobile ones, which welcomes any analytical approach in the field.

The purpose of this research is represented by the need to highlight the trend regarding the access and use of mobile banking and mobile payment services by student consumers, in close connection with the cultural factors that can influence this trend. In this regard, the paper considers aspects such as observing the students' behavior regarding how mobile banking is used in the monitoring of bank accounts, taking financial decisions about saving, investments, as well as shopping.

The motivation of this research is based on the fact that the explosive growth of online shopping in Romania - as a result of the economic growth based on the stimulation of consumption - has led to the pragmatism of choosing other forms of payment by consumers (especially young people and especially students), so that a study on the use of mobile banking and mobile payment by students is required in order to understand these trends related to students behavior in Romania.

The research methodology addressed aimed at both qualitative and quantitative research, considering the mobile financial services offered by banks operating in Romania. The authors conducted a quantitative research based on a questionnaire, grouping selectively a multitude of questions on how students perceive mobile banking according to the specific interaction they had with it. The survey was attended over a period of two weeks and targeted a sample of 250 students from the University of Bucharest - Romania, out of which 242 were respondents, with a specified socio-demographic distribution. The structural model that was used is called the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2), a combination of UTAUT developed in 2003 by Venkatesh et al. [23] and Hofstede's cultural moderators [10]. The qualitative analysis considered the placement of the factors analyzed in a cultural context linked with the Romanian students, trying to investigate the values, perceptions, desires, but also the stereotypes and prejudices learned/taken from the family or other social organisms.

The results of this research reveal how cultural factors determine the Romanian students attitude close to the use of mobile banking services. Factors that have a positive influence on behavioral intention are performance, facilitating conditions and habit. Habit has also a positive influence on use behaviour. Social influence has a negative effect on behavioral intention. There is no relationship between behavioral intention and effort expectancy, hedonic motivation and price value. Also, there is no association between use behavior and behavior intention. All cultural moderators signal a negative influence on the behavioral intention - use behavior relationship.

The limits of research in the authors' view are reflected in the size and selection of the research sample, considering a limited number of students from the University of Bucharest, but also in widening and deepening the analyze of the Hofstede's cultural moderators.

## 2 Literature Review

The researches in the sphere of mobile banking related to the dissemination, access and use, as well as the multitude of the factors that influence it, have met a diversification of the approaches and working methodologies according to the objectives that, at present, make the interdisciplinary investigation of the phenomenon in question necessary.

In the context of examining the human-computer interaction of the topic of mobile banking and the cultural factors that influence it, the present paper refers to a series of researches in this field that the authors have considered in their approach.

In analyzing the interaction between cultural factors and the adoption of mobile banking services by Romanian students, the definition of culture at the organizational or group level is naturally considered. Thus, Hofstede [10] designates culture as being “the collective programming of the human mind that distinguishes the members of one human group from those of another. Culture in this sense is a system of collectively held values”. Schein [18] shows that culture represents “the way in which people solve their problems/dilemmas or ways of solving which are verified and consolidated over time and which are transmitted to future generations in the form of values, ideas, symbols important for human nature”.

Concerning the cultural factors effect on mobile banking adoption, the authors consider as a benchmarking research for this paper the work of Baptista and Oliveira [4] titled “Understanding mobile banking: The unified theory of acceptance and use of technology combined with cultural moderators” which come up with “an innovative and comprehensive theoretical model that combines the extended unified theory of acceptance and use of technology (UTAUT2) of Venkatesh, Thong, and Xu (2012), with Hofstede’s cultural moderators, providing new insights into factors affecting the acceptance and how culture influences individual use behavior” [4]. As results of their research, the authors underline the significance of “performance expectancy, hedonic motivation, and habit” for behavior intention and “collectivism, uncertainty avoidance, short term and power distance” as being essential cultural moderators.

Buzamat [6] in her research “tries to underline the main components of the Romanian culture [...] in order to be able to quantify the Romanian culture based on the cultural dimensions of Geert Hofstede” showing numerous implications in Romanian economy and organizations.

Another paper worthy being cited is that of Mahfuz, Hu, and Khanam [15] where cultural dimensions are considered in analyzing the adoption and use of mobile banking services: “performance expectancy, facilitating condition and price value influences on behaviour intention but effort expectancy had no influence on m-banking adoption in this research” and “power distance had influence on m-banking adoption and masculinity and uncertainty avoidance had no influence on behavioral intention to adopt m-banking services”.

Merhia, Honea and Tarhinib [16] show in their research on “mobile banking adoption in Lebanon and England” that “behavioural intention towards adoption of mobile banking services was influenced by habit (HB), perceived security (PS), perceived privacy (PP) and trust (TR) for both the Lebanese and English consumers. In addition, performance expectancy (PE) was a significant predictor in Lebanon but not in England; whereas price value (PV) was significant in England but not in Lebanon. [...] Social Influence

(SI) and Hedonic Motivations (HM) were insignificant for both the Lebanese and English consumers”.

The paper “Does culture influence m-banking use and individual performance?” of Tam and Oliveira [19] considers “a combination of the task-technology fit (TTF) model and two of Hofstede’s cross-cultural dimensions scale” and shows that “individualism moderates the relationship between TTF and use, and uncertainty avoidance moderates the relationship between TTF and individual performance”.

Akhtar et al. [2] analyzed the mobile banking adoption in both Pakistan and China and discovered “the moderating role of cultural values was observed as dampening factor in positive relationship between social influence and individuals’ intentions”.

The authors of the paper “Apps for mobile banking and customer satisfaction: a cross-cultural study” [17] distinguish “how perceived justice moderates the relationship between the benefits offered by mobile banking and the consequences of satisfaction with mobile banking” in three contrasting countries from the perspective of cultural diversity and economic and social development: Brazil, India and the United States [17].

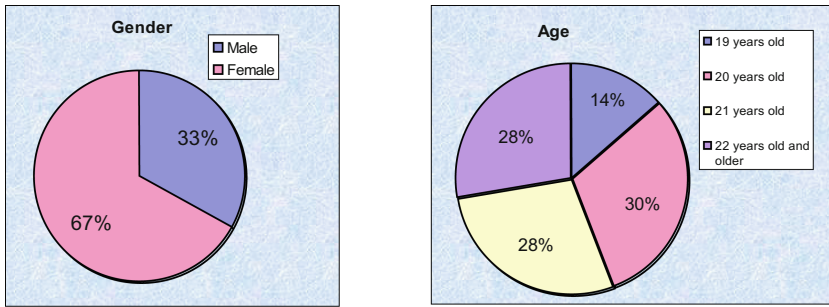
### 3 Cultural Factors Influence on Romanian Students Use of Mobile Banking

#### 3.1 Data and Methodology

The authors conducted a quantitative research based on a questionnaire, similar to the one used by Baptista & Oliveira [4], in order to obtain the data for this analysis. The survey comprises 49 questions with answers measured on 7-point Likert scale (from 7 - strongly agree to 1 - strongly disagree) and other questions about using mobile banking among students. We will consider these 49 questions to be measured variables that will form the latent variable used in the statistical analysis performed in the paper. Use behaviour, one of the latent variables, was marked from 1 (never) to 11 (several times per day), corresponding to the frequency of use of mobile banking services. The respondents were students from the first, second and third year of undergraduate study and students from master degree study within the Faculty of Administration and Business, University of Bucharest.

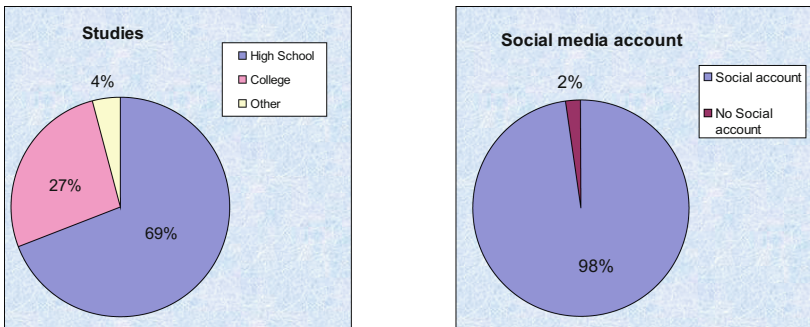
We have a sample of 237 respondents. Out of the total there were 66.7% female and 33.3% male respondents (Fig. 1). About the age, 30.8% of total were 20 years old, 27.8% were 21 years old, 13.5% were 19 years old and the rest of 27.9% were 22 years old or older. The majority of 69.6% have their residence in Bucharest, 15.2% have their residence in Muntenia (South-East of Romania), 5.9% in Moldova (East of Romania), 5.5% in Oltenia (South-West of Romania) and the rest of 3.8% in Dobrogea (South-South-East of Romania), Transilvania (Centre of Romania) and Moldavian Republic (Fig. 1).

Out of the total, 69.2% graduated the high school, 26.6% graduated college and the rest of 4.2% graduated another form of studies (Fig. 2). Majority of 30.8% have an income less than 999 lei (about 200 euros), 25.7% have an income between 1000 and 1999 lei (200–400 euros), 15.6% declared their income is 2000–2999 lei (400–600 euros), 12.7% declared an income bigger than 3000 lei (600 euros) and 15.2% didn’t



**Fig. 1.** Gender and age of the respondents (Source: authors calculations by using R Studio)

want to declare their income. All the respondents declared they have a smartphone, 97.9% said they have a social media account (Fig. 2) and 93.3% said they use Mobile Banking application (Fig. 3).



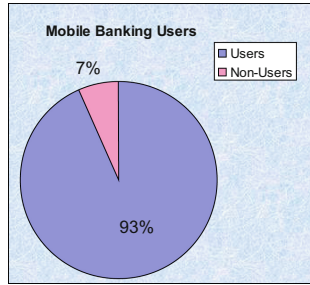
**Fig. 2.** Studies and social media account of the respondents (Source: authors calculations by using R Studio)

About the frequency of Mobile Banking using, 22.8% said they use it at 2–3 days, 18.6% daily, 13.1% once a week, 10.1% at 4–5 days or once a month, 14.8% rarely than once a week, 4.2% several times a day and 6.3% do not use it at all.

The questionnaire tested how students perceive mobile banking according to the specific interaction they had with it. Out of the 49 items, a number of latent variables were constructed: performance, social influence, effort, facility conditions, price value, hedonic motivation, habit, behavioural intention, individualism/collectivism, uncertainty avoidance, masculinity/femininity, power distance, long/short term orientation and indulgence [10–12].

### 3.2 Structural Equation Modelling for Mobile Banking

Just as correlation, regression and analysis of variance are general linear model, so is structural equation modelling, a technique introduced for the first time in the early 70's in behavioral research. Structural equation modelling (SEM) has the capacity to estimate



**Fig. 3.** Use of Mobile Banking by the respondents (Source: authors calculations in R Studio)

and test the connection among constructs or latent variable. Also, when we use SEM, we have to appraise multiple test statistics and a host of fit indices to find whether the model fits the data.

The structural model that we use in this paper and that also we want to test is called Unified Theory of Acceptance and Use of Technology 2 (UTAUT2). This model is a combination of Unified Theory of Acceptance and Use of Technology model (UTAUT) developed in 2003 [24] and Hofstede’s cultural moderators [10–12]: individualism/collectivism, power distance, uncertainty avoidance, masculinity/femininity, long/short term orientation and indulgence.

In the beginning we will calculate the confirmatory factor analysis because as mentioned before, the questionnaire was used in another study. We will determine if our factor analysis is suitable with our data (Kaiser-Meyer-Olkin Measure of Sampling Adequacy and Bartlett’s test of sphericity) and then we will test the reliability (Cronbach alpha coefficient). After this, we can see in Table 1 some indicators used to determine if the baseline model is stable.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) for these data is 0.956. We can say this value is very good from a statistical point of view and factor analysis can be suitable with data that we have. For Bartlett’s test of sphericity, we obtain a p-value of  $2.2e-16 < 0.05$  so we can say a factor analysis is useful with data.

Next, we will test the reliability and so we will develop Cronbach alpha coefficient. In this case we obtain that the value for this indicator is  $0.977 > 0.75$ , so we can say there is a good internal consistency.

We can see in Table 1 that almost all conditions are satisfied. To make some additional improvements to the model we will eliminate those variables that are not statistic significant according to the following criteria: the variables that have R-squared less than 0.4 or even 0.5 will be eliminated. In this way we remain with 47 measured variables used to construct 14 latent variables and further we can construct the structural equation model.

In order to construct the structural equation model, we will start with some hypotheses that we want to test in this paper (Table 2):

In Table 3 we have the values corresponding to regressions used to decide if the above hypothesis are accepted or rejected. For a positive estimate value in table above, we will say that we have a direct or positive relation between the variable. To determine

**Table 1.** CFA fitting values for baseline model

| Indicator                          | Expected value       | Value in the model  |
|------------------------------------|----------------------|---------------------|
| Convergence & number of iterations |                      | Yes, 139 iterations |
| Observations                       | As big as possible   | 237                 |
| Chi-square                         | > 0.05               | 0.000               |
| CFI                                | > 0.95               | 0.894               |
| TLI                                | > 0.95               | 0.882               |
| RMSEA                              | < 0.07               | 0.082               |
| 90% confident interval             | (0; 1)               | (0.078; 0.086)      |
| SRMR                               | < 0.08               | 0.068               |
| AIC                                | As small as possible | 32629.276           |

Source: authors calculations by using R Studio

**Table 2.** SEM hypothesis

|     |   |
|-----|---|
| H1  | Performance has a positive impact on behavioural intention                              |
| H2  | Effort has a positive impact on behavioural intention                                   |
| H3  | Social influence has a positive impact on behavioural intention                         |
| H4A | Facility conditions has a positive impact on behavioural intention                      |
| H4B | Facility conditions has a positive impact on use behaviour                              |
| H5  | Hedonic motivation has a positive impact on behavioural intention                       |
| H6  | Price value has a positive impact on behavioural intention                              |
| H7A | Habit has a positive impact on behavioural intention                                    |
| H7B | Habit has a positive impact on use behaviour  |
| H8  | Behavioural intention has a positive impact on use behaviour                            |
| H9  | Individualism/collectivism moderates the impact of behaviour intention on use behaviour |
| H10 | Uncertainty avoidance moderates the impact of behaviour intention on use behaviour      |
| H11 | Long/short term moderates the impact of behaviour intention on use behaviour            |
| H12 | Masculinity/femininity moderates the impact of behaviour intention on use behaviour     |
| H13 | Power distance moderates the impact of behaviour intention on use behaviour             |

Source: Baptista, G., Oliveira, T. (2015) [4]

if the relation between variable is statistic significant, we will compare p-value for each hypothesis with critical value 0.05 (for p-value is less than 0.05, the relation is statistic significant).

**Table 3.** SEM indices and decision for hypothesis tested

| Hypothesis | Estimate | Standard error | z-value | P     | Decision |
|------------|----------|----------------|---------|-------|----------|
| H1         | 0.132    | 0.067          | 1.977   | 0.048 | Accept   |
| H2         | 0.067    | 0.094          | 0.714   | 0.475 | Reject   |
| H3         | -0.149   | 0.056          | -2.662  | 0.008 | Accept   |
| H4A        | 0.342    | 0.124          | 2.755   | 0.006 | Accept   |
| H5         | 0.003    | 0.084          | 0.034   | 0.973 | Reject   |
| H6         | 0.021    | 0.072          | 0.289   | 0.773 | Reject   |
| H7A        | 0.656    | 0.065          | 10.060  | 0.000 | Accept   |
| H7B        | 1.010    | 0.450          | 2.244   | 0.025 | Accept   |
| H4B        | 0.647    | 0.345          | 1.872   | 0.061 | Reject   |
| H8         | -0.649   | 0.609          | -1.065  | 0.287 | Reject   |
| H9         | -0.080   | 0.033          | -2.437  | 0.015 | Accept   |
| H10        | -0.080   | 0.033          | -2.437  | 0.015 | Accept   |
| H11        | -0.080   | 0.033          | -2.437  | 0.015 | Accept   |
| H12        | -0.080   | 0.033          | -2.437  | 0.015 | Accept   |
| H13        | -0.080   | 0.033          | -2.437  | 0.015 | Accept   |

Source: authors calculations by using R Studio

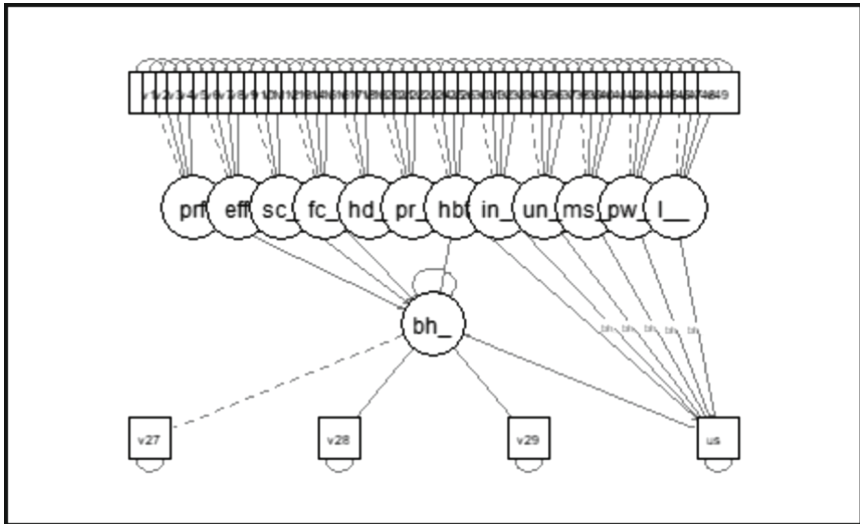
We can observe in Table 3 that out of all hypothesis tested, five of them will be rejected. In these cases, we can say that there will not be any relation between behavioural intention and following variables: effort, hedonic motivation and price value; there will not be any relation between use behaviour and the following variables: facility conditions and behavioural intention. Also, we determine that social influence has a negative effect on behavioural intention and all moderators have a negative influence.

After we eliminate the variable not significant according to previous table, we will construct the structural model for this case (Fig. 4).

The abbreviation in the image above are the following: prf = performance, eff = effort, sc\_ = social\_influence, fc\_ = facility\_conditions, hd\_ = hedonic\_motivation, pr\_ = price\_value, hbt = habit, in\_ = individualism\_collectivism, un\_ = uncertainty\_avoidance, ms\_ = masculinity\_femininity, pw\_ = power\_distance, l\_ = long\_short\_term, us = use\_behaviour, bh\_ = behavioural\_intention.

## 4 Results: Interpretation and Discussions

The results of this research reveal what cultural factors determine to a lesser or greater extent the Romanian students attitude close to the use of mobile banking services and how this factors could have a positive influence or a negative one on behavioral intention and use behavior of the student consumer.



**Fig. 4.** Graph of structural equation model for mobile banking (Source: authors calculations by using R Studio)

#### 4.1 Performance Expectancy, Facilitating Conditions and Habit Have a Positive Effect on Behavioral Intention

Performance expectancy (the degree to which the individual admit that the use of a particular technology, will bring benefits and increase its performance), facilitating conditions (the degree to which the individual accept that the technical framework needed to endorse the use of a technology in the organization exists) and habit (the natural behavior of an individual that can be seen either as a precursor to an act of purchase or as an automatism) exerts a positive effect on behavioral intention. Habit has also a positive influence on use behaviour.

Facilitating conditions influences the intention to use mobile banking services (hypothesis 4A accepted), and on the other hand, does not exercise any influence on the frequency of their use (hypothesis 4B rejected).

By processing of our data it is clear that the respondents want to use mobile banking services because they are aware of their usefulness, they know that the necessary infrastructure exists and, at the same time, they realize that they, as beneficiaries, have the skills and the resources needed to access them. On the other hand, there is no dependency between facilitating conditions and behavioral intention (hypothesis 4B rejected). In other words, the existence of the necessary resources (infrastructure, skills) did not cause the individuals to manifest the intention to use the mobile banking services, even if they are compatible with other technologies used by the respondents. Also, according to the data, the habit of calling on mobile banking services (habit) acts in some cases as a precursor of the intention to use these services (hypothesis 7A accepted) while in others, it acts as an automatism, also determining the frequency their use (hypothesis 7B accepted). Thus, we find a double manifestation exerted by facilitating conditions (positive on behavioral intention and indifference in relation to use behavior) and habits



(positive effect on behavioral intention, but also on use behavior) in the case of Romanian students.

## **4.2 Social Influence Has a Negative Leverage on Behavioral Intention**

This factor shows the intensity with which an individual perceives family or friends as influencing the decision to embrace a certain technology. In our case, the social influence has a negative effect on the intention to act or to behave/manifest in a certain way (behavioral intention).

According to Kelman [14], there are three different stages related to social influence which affects consumer behavior: “compliance, identification and internalization”. Confirmation implies that an individual adopts a certain behavior because he expects to thus obtain a reward or to avoid paying/bearing a sanction. The identification is associated with the acceptance of the influence coming from the family or the close ones because the individual wants to maintain a satisfactory relationship with other person/persons within the group.

Internalization occurs when an individual accepts to be influenced by the others due to the fact that these influences resonate with their own system of values. In our case, it has been proven that consumers seem to want to adopt mobile banking just because they want to comply with the new trend (carrier of benefits). In contrast, by adopting the mobile banking they do not necessarily want to identify with the group of belonging, and the new technology is not necessarily consistent with their own values. For these reasons, in this case, the social influence negatively impacts the behavioral intention as well as the consumer’s manifest behavior close to the mobile banking adoption.

## **4.3 There Is No Dependence Between Behavioral Intention and Effort Expectancy, Hedonic Motivation and Price Value**

Effort expectancy or the degree to which the individual distinguishes the use of technology as easy/accessible or difficult does not affect the intention to use such technology (hypothesis H2 rejected). Most of our respondents belong to the Z generation, meaning the people who were born with the technology “in their arms”. Therefore, everything related to the use of technology will not require effort, an observation that can be found in other studies [7]. Young people have the skills needed to access mobile banking services, but this is not an incentive for them to actually use them.

Hedonic motivation or the extent to which an individual experiences joy/pleasure when using a technology - does not exert any effect on the intention (behavioral intention) to use mobile banking services.

The hedonic goods are considered to be luxury goods, meant to bring the consumer pleasure and joy when purchasing them. This is mainly the difference from the utility goods that are purchased for daily use, and are intended to cover the basic needs of the consumer. For this reason, the consumer wants to spend more on hedonic goods - because he realizes that they produce joy, while the utilitarian goods, common goods, are bought out of necessity, and do not bring the consumer any joy. This explains why our respondents lack the motivation to engage in the purchase of mobile banking: because mobile banking services are utilitarian, common goods, and their acquisition is not

generating pleasure or joy (hypothesis 5 rejected). On the other hand, consumers may not feel joy when using mobile banking services, but they certainly appreciate the benefits of such services: reducing transportation costs to a fixed banking point, and eliminating time wasted by standing on a long tail.

#### **4.4 Price Value Has No Influence on Use Behavior**

Price value or “the trade-off between the cost of using the technology and the associated benefits” [9] does not exert any influence on the intention to use mobile banking services (hypothesis H6 rejected). In the case of mobile banking, the perceived costs are small (finding consistency with the literature), and the benefits are significant in terms of time. Therefore, the price of these services does not have a decisive role in the purchase decision and, therefore, in the purchase intention, as a precursor factor of the purchasing act.

#### **4.5 There Is No Correlation Between Use Behavior and Behavioral Intention**

Behavioral intention - a term that indicates how prepared an individual is to perform a certain activity (the previous phase of manifest behavior) and use behavior - which shows the frequency of service use - are extremely intense terms researched in the literature [4, 24]. In our study, hypothesis H8 according to which the intention to use the mobile banking services determines the frequency of their use is rejected. In line with the specialized literature [7] the intention to buy a product or service may result in one of the following situations: the consumer decides to buy the product/service; may delay the purchase or may not buy the product/service due to the non-attractive offer existing at that time or if the consumer prefers to wait for possible price reductions/discounts or the emergence of a new product/new service, corresponding to his needs [7]. In none of the above situations, the intention to use is a prognosticator of the frequency of purchase/use of the respective product/service. In the case of Romanian students, the answers received confirm that their intention (behavioral intention) to use mobile banking services is not a precursor of the frequency of their use (use behavior).

#### **4.6 All Cultural Moderators Have a Negative Influence on the Behavioral Intention (BI) - Use Behavior (UB) Relationship Subsection Sample**

The Hofstede's cultural moderators – individualism-collectivism, power distance, uncertainty avoidance, masculinity-femininity, long/short term orientation and indulgence - have a negative effect on the behavioral intention (BI) - use behavior (UB) relationship.

**Individualism-Collectivism.** This cultural factor refers to “the degree of interdependence a society maintains among its members. It has to do with whether people's self-image is defined in terms of ‘I’ or ‘We’” [13]. In individualistic societies, people are presumed to care only about their own person and their family. In collectivist societies, people group and take care of each other as a result of the loyalty they carry. Considering all that, an elevated score means that the person and his rights are supreme within the society.

Romania has a score of 30 [13] and is seen as a collectivist society. In collectivist societies, the offense entails shame and image alteration, the employee-employer relations are perceived in moral terms (as a family link), the decisions of employment and promotion take into account the group of the employee, and the management is one of group management [10]. Morality in a collectivist culture is much more contextual, and the supreme value is the good of the collectivity [20]. Therefore, to lie becomes a more acceptable behavior in collectivist cultures than in individualistic cultures, if the purpose of the lie is to save one's reputation or help one's own group. Trilling [22] argues that when people have a strong sense of self-determination (in individualistic cultures) - they seek sincerity and authenticity. In contrast, when they feel overwhelmed by traditions and obligations (in collectivist cultures) they do not put on authenticity. According to Triandis [20, 21] there is a greater tendency of interpersonal deception among the collectivities.

Also Triandis [20, 21] said that "in collectivist cultures, morality consists in doing what the group expects". If individuals deviate from morality, not only do individuals lose their reputation, but the group's reputation is also affected. When individuals interact with persons outside the group, it is not considered "immoral" to exploit them and deceive others. In other words, in collectivist cultures, morality is only applicable to members of their own group/in-group [20, 21].

The same author [21, 22] also shows that individuals in collectivist societies are less motivated in the situations in which they have to decide: being in a situation where someone else trusts the decision on their behalf, they get to activate the highest level of intrinsic motivation and performance.

The consequences of collectivism on the adoption of MB services by Romanian consumers can be summarized as follows: Romanians tend to be influenced by inherited or adopted habits; the mimicry is manifested, respectively the tendency to use the same financial and technological tools used by family and friends (the membership group). And because morality has a deeply contextual nature, in the case of Romanians, the decision to use or not use MB services is marked by the emotional side: if the consumer feels threatened by the possibility of being the victim of a deception, he will avoid getting involved and to access the services offered, preferring to stay in expectation. Thus, the consumer will protect himself and, implicitly, the group he belongs to, from potential inconveniences.

**Uncertainty Avoidance.** This factor highlights the perception of society regarding the uncertainties of the future: to try to change it or to simply let it happen?

The extent to which the members of a culture feel threatened by ambiguous or unknown situations and have created beliefs and institutions that try to avoid these" [13] places Romania, with the 90 points, among the countries with a high degree of uncertainty.

Often the behavior of the individual towards banks is based on errors of perception or analysis of the context [8]. Negligence, lack of information, inability to properly process the available information, fear again - all lead to the choice of inappropriate solutions. Thus, a study by Agarwal et al. [1] analyzes the results of an experiment based on the choice of the right type of credit card by the clients of a bank and observes that a large number of them made a wrong choice - which led to the making of some payments

greater than necessary. Some of these consumers, finding the error, fixed it at one point, opting for a more appropriate type of card. However, a small percentage of consumers persisted in the mistake, continuing to use the wrong card. Moreover, Ameriks et al. [3] show that there is a third type of consumers - the "distracted" - who fail to remember how they spend/manage their money.

The consequences of avoiding uncertainty about this dimension in the individual-bank and individual-MB relationship services are: the tendency to reject/adopt with difficulty the novelties in the financial field as well as to show preferences for the instruments whose use implies a low risk (such as debit cards).

**Long/Short Term Orientation.** "This dimension describes how every society has to maintain some links with its own past while dealing with the challenges of the present and future, and societies prioritise these two existential goals differently" [13]. According to Hofstede, the West is oriented in the short term, and Romania has an intermediate score of 52 in this regard.

The orientation towards the long term is part of the financial education of the population, in the sense that the attitudes, abilities and individual limits in the financial field are built, but also modified through education. According to Behrman et al. [5], book science in the financial field - defined as the ability to interpret economic information and make informed decisions about money management - contributes to the accumulation of wealth.

On the other hand, the complexity of the financial environment increases with the passing day: the number of people who make payments online at merchants, pay services to various suppliers using the mobile phone, make investments on the stock market or resort to sophisticated saving methods is increasing. In this context, Romanian students can be regarded as persons with a high degree of financial education, which should induce them the perception of a low risk associated with online operations, of any kind. But our study shows that things are not so: students perceive banking as risky. The mistrust and fear of deception prevail, and the experiences of individual nature explain their reluctance to engage in a long-term relationship with the bank. We return below the opinion of Adrian (second year student) in this regard: "My parents took out a mortgage loan from the bank to buy an apartment from my older sister. They guaranteed with their own house. Only the financial crisis of 2008 came, and my father lost his job. We could not repay the bank loan anymore, and in a short time we were threatened with eviction. Fortunately we didn't get to the street - and that's because my father and I managed to find new jobs. And yes, I hope we can pay off the credit. But, if we were to take from the beginning, we, as a family, do not think we would make the same mistake".

In other words, young people like Adrian, despite the high level of financial information and education, will tend to avoid long-term relationships with the bank, whether it is credit services or other services in general. This is because contextual (economic) changes can occur and may affect the smooth running of a long-term contract between the individual and the financial service provider.

**Power Distance.** The distance from power refers to "the extent to which the less powerful members of institutions and organisations within a country expect and accept that power is distributed inequally" [13].

Romania has a high score of 90 for this dimension, which shows the acceptance of a hierarchical order, in which each individual has his or her place.

As an impact on financial decisions, the distance from power can have the following consequences: manifesting preferences or easily accepting elements related to the field of personal finances imposed by the state or any other authority; easier acceptance of those instruments (financial, technological) “recommended” by the persons with authority.

**Masculinity-Feminity.** “The fundamental issue here is what motivates people, wanting to be the best (Masculine) or liking what you do (Feminine)” [13].

A high score, refers to masculine characteristics and points out that the society promotes competition, personal success (defined as being the best in the field), and a values system founded by the school and continued throughout the professional career. In contrast, a low, feminine score means that the dominant values are the concern for others and the quality of life [10, 13].

Romania, with a score of 42, is revealed as a society with feminine characteristics. In countries with such features, the focus is on “working for a living”, manager fight for consensus while men value solidarity and quality at work. The emphasis is on the well-being, and the social status is not displayed.

The consequences of manifesting this dimension at the individual level are: individuals seek to make those decisions (financial, technological) that could differentiate them from others; groups are far from homogeneous.

**Indulgence.** Defined in terms of tolerance versus restrictiveness, indulgence is the latest dimension added by Hofstede with an impact on consumer decisions. Tolerance refers to “the extent to which people try to control their desires and impulses, based on how they were raised” [10, 11]. Relatively weak control is associated with tolerant behavior, and relatively strong control is associated with restrictive behavior. Cultures can therefore be described as tolerant or restrictive.

With a very low score of 20, the Romanian culture is restrictive [13]: “Societies with a low score on this dimension have a tendency towards cynicism and pessimism”. Also, in opposition to tolerant societies, restrictive societies do not emphasize the way of spending their free time. People with such (restrictive) orientation have the perception that their actions are restricted by social norms and feel that tolerating their own desires is somehow wrong.

We mention that this dimension was not tested in our study. The authors have included it, however, in the discussion about moderating factors because it explains to some extent the context of mistrust that puts its mark on the consumers’ relationship with the bank.

**Romanians’ Popular Beliefs on Money and Banking.** The subconscious conceptions and the stereotypes related to the money of the Romanians have many variants of expression, the most common being: a) “Money is a bad thing”; b) “The rich are greedy, superficial and insensitive”; c) “Money is the eye of the devil”.

The Moral: “If you have money, this is a bad thing that has nothing to do with the spiritual world, but with something devilish. Wealthy people are supposed to have done something wrong to tighten their wealth. Therefore, if you have money, you are a bad person. Rich people are bad - says the collective mind - so if I get rich, I’ll be a bad man too”.

This extremely widespread conception in Romania can come from one of the parents, or from the bitterness related to raising a child in a poor environment. Some of our respondents witnessed the disintegration of their families when they could no longer pay their debts to the bank, or the destruction of their relationships because of what initially seemed like luck (example: winning the lottery).

People, who perceive reality in this way, deform it to support their own beliefs. There are many who say that money is the source of all evil. But, the Bible says that the love of money lies at the root of all evil, so money is not the problem, but the way we relate to them.

In these circumstances, we can say that the relationship of consumers with the bank and the services offered is tarnished by preconceptions that are emotional in nature and cannot be easily disassembled. As long as the Romanian associates the banks with the money, it can be assumed that there is a fear/resistance to what the banks can offer due to the destructive potential that individuals attribute to the money, and, consequently, to the institution that manipulates them: the bank. For this reason, consumers tend to reject MB services, or at least postpone their use in the near future, even if they perceive that the risks of using MB are (significantly) lower than the benefits.

## 5 Conclusions

Our study has analyzed and discussed aspects specific to how students/young people relate to mobile banking services. Thus, if from the perspective of the factors that influence the purchase intention (behavior intention) of mobile banking services and the frequency (use behavior) of their use, we have identified many similarities with the results of other research in the field [4, 23], the present study highlights numerous differences in the consumer behavior of mobile banking services, generated by the effect of cultural factors.

A summary of the results underline the followings:

- a) Factors that have a positive influence on behavioral intention are performance, facilitating conditions and habit. Habit has also a positive influence on use behaviour.
- b) Social influence has a negative impact on behavioral intention.
- c) There is no relationship between behavioral intention and effort expectancy, hedonic motivation and price value.
- d) There is no relationship between use behavior and behavior intention.
- e) All cultural moderators have a negative influence on the behavioral intention - use behavior relationship.

### 5.1 Original Contribution

The specific contribution of the authors in carrying out this research is given by the application of the UTAUT2 analysis methodology to the particular conditions of the social environment in Romania, more precisely on the student community within the Faculty of Public Administration and Business of the University of Bucharest.

Context plays an important role in the individual's relationship with the bank, Hofstede's cultural dimensions (individualism/collectivism, uncertainty avoidance, long/short

term orientation, power distance, masculinity/femininity, indulgence) exerting a negative effect on the relationship between intention to use (behavioral intention) and frequency of use (use behavior) of mobile banking services. This effect can be explained by the tendency accentuated towards collectivism, the low tolerance towards the risk, the orientation towards the short term, the tendency of submission towards the authority and the orientation towards the models promoted by the group. These cultural trends give Romanian students a rather restrictive/reserved behavior in relation to banks: despite the benefits that mobile banking services generate (and which consumers are aware of), young people access them with distrust. This is largely due to preconceptions inherited or acquired in the family.

## 5.2 Limitations

The limits of research in the authors' view are reflected in the size and selection of the research sample, considering a limited number of students from the University of Bucharest, but also in widening and deepening the analyze of the Hofstede's cultural moderators.

## 5.3 Vision on Future Research

The authors consider that, in the perspective of future research on this topic, it is necessary to take into account the limitations already highlighted in order to deepen the analysis and to obtain results as close to reality as possible and to pay greater attention to consumer behavior, a necessary study for providers/ creators of mobile banking financial services.

In this regard, a number of elements, such as: increasing the research sample by including students from other faculties of the university and even co-opting other universities in carrying out a larger study, with a greater impact in terms of results; considering the information needs and the necessary data to be made available to the providers/creators of financial services for mobile banking by adequately adapting the questions from the questionnaires already used, based on previous internationally recognized researches.

## References

1. Agarwal, S., Driscoll, J.C., Gabaix, X., Laibson, D.: The age of reason; financial decisions over the lifecycle. NBER Working Paper Series (13191) (2007). <https://www.nber.org/papers/w13191.pdf>
2. Akhtar, S., Irfan, M., Sarwar, A., Rashid, Q.U.A.: Factors influencing individuals' intention to adopt mobile banking in China and Pakistan: The moderating role of cultural values. *J. Public Aff.* **19**(1) (2019). <https://doi.org/10.1002/pa.1884>
3. Ameriks, J., Caplin, A., Leahy, J.: The absent-minded consumer. NBER Working Paper Series (10216) (2004). <https://www.nber.org/papers/w10216.pdf>
4. Baptista, G., Oliveira, T.: Understanding mobile banking: the unified theory of acceptance and use of technology combined with cultural moderators. *Comput. Hum. Behav.* **50**, 418–430 (2015). <https://www.sciencedirect.com/science/article/pii/S0747563215003118>

5. Behrman, J.R., Mitchell, O.S., Soo, C., Bravo, D.: Financial literacy, schooling and wealth accumulation. PARC Working Paper Series, WPS 10-06 (2010). [https://repository.upenn.edu/cgi/viewcontent.cgi?article=1031&context=parc\\_working\\_papers](https://repository.upenn.edu/cgi/viewcontent.cgi?article=1031&context=parc_working_papers)
6. Buzamat, G.: Study regarding cultural diversity and organizational behavior in Romania. In: Conference Proceedings 2nd International Scientific Conference ITEMA 2018 (2018). <https://doi.org/10.31410/itema.2018.874>
7. Cățoiu, I., Teodorescu, N.: Comportamentul Consumatorului: Abordare Instrumentală. Uranus Publishing House, Bucharest (2001)
8. Ciomara, T.: Caracteristici si factori de influenta ai deciziei financiare. Victor Slavescu Center for Financial and Monetary Reform, Bucharest (2012). [https://www.researchgate.net/publication/259557326\\_Capitol\\_3\\_Caracteristici\\_si\\_factori\\_de\\_influenta\\_ai\\_deciziei\\_financiare](https://www.researchgate.net/publication/259557326_Capitol_3_Caracteristici_si_factori_de_influenta_ai_deciziei_financiare)
9. Dhiman, N., Arora, N., Dogra, N., Gupta, A.: Consumer adoption of smartphone fitness apps: an extended UTAUT2 perspective. J. Indian Bus. Res. Vol. ahead-of-print No. ahead-of-print (2019). <https://doi.org/10.1108/jibr-05-2018-0158>
10. Hofstede, G.: Culture's Consequences: International Differences in Work-Related Values. Sage Publications, Beverly Hills (1980)
11. Hofstede, G., Bond, M.H.: The Confucius connection: from cultural roots to economic growth. Organ. Dyn. **16**(4), 5–21 (1988)
12. Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and Organizations: Software for the Mind. McGraw-Hill, New York (2010)
13. Hofstede Insights. <https://www.hofstede-insights.com/country/romania/>
14. Kelman, H.C.: Compliance, identification and internalization: three processes of attitude change? J. Confl. Resolut. **2**, 51–60 (1958)
15. Mafhuz, M.A., Hu, W., Khanam, L.: The influence of cultural dimensions and website quality on mbanking services adoption in Bangladesh: Applying the UTAUT2 Model Using PLS. In: Proceedings Wuhan International Conference on e-Business WHICEB 2016. 18 (2016). <http://aisel.aisnet.org/whiceb2016/18>
16. Merhia, M., Honea, K., Tarhinib, A.: A cross-cultural study of the intention to use mobile banking between Lebanese and British consumers: Extending UTAUT2 with security, privacy and trust. Technology in Society (2019). <https://www.sciencedirect.com/science/article/pii/S0160791X19300132>
17. Sampaio, C., Ladeira, W., Santini, F.: Apps for mobile banking and customer satisfaction: a cross-cultural study. Int. J. Bank Mark. **35**(7), 1133–1153 (2017). <https://doi.org/10.1108/IJBM-09-2015-0146>
18. Schein, E.H.: Organizational culture and leadership: a dynamic view. Organ. Stud. **7**, 199–201 (1985). <https://doi.org/10.1177/017084068600700208>
19. Tam, C., Oliveira, T.: Does culture influence m-banking use and individual performance? Inf. Manag. **56**(3), 356–363 (2019). <https://doi.org/10.1016/j.im.2018.07.009>
20. Triandis, H.C.: Individualism and collectivism. Westview, Boulder CO (1995)
21. Triandis, H.C., Suh, E.M.: Cultural influences on personality. Annu. Rev. Psychol. **53**, 133–160 (2002)
22. Trilling, L.: Sincerity and Authenticity. Harvard University Press, Boston (1972)
23. Venkatesh, V., Morris, M.G., Hall, M., Davis, G.B., Davis, F.D., Walton, S.M.: User acceptance of information technology: toward a unified view. MIS Q. **27**(3) (2003)





# Social Behaviour Understanding Using Deep Neural Networks: Development of Social Intelligence Systems

Ethan Lim Ding Feng<sup>1</sup>, Zhi-Wei Neo<sup>1</sup>, Aaron William De Silva<sup>1</sup>, Kellie Sim<sup>2</sup>, Hong-Ray Tan<sup>2</sup>, Thi-Thanh Nguyen<sup>3</sup>, Karen Wei Ling Koh<sup>4</sup>, Wenru Wang<sup>5</sup>, and Hoang D. Nguyen<sup>1</sup>(✉)

<sup>1</sup> University of Glasgow, Singapore, Singapore  
{2427232L, 2355362N, 2355348D}@student.gla.ac.uk,  
Harry.Nguyen@glasgow.ac.uk

<sup>2</sup> Ngee Ann Polytechnic, Singapore, Singapore  
{s10163148, s10177638}@connect.np.edu.sg

<sup>3</sup> National College for Education, Ho Chi Minh City, Vietnam  
thanhtw76@gmail.com

<sup>4</sup> National University Health System, Singapore, Singapore  
karen\_wl\_koh@nuhs.edu.sg

<sup>5</sup> National University of Singapore, Singapore, Singapore  
nurww@nus.edu.sg

**Abstract.** With the rapid development in artificial intelligence, social computing has evolved beyond social informatics toward the birth of social intelligence systems. This paper, therefore, takes initiatives to propose a social behaviour understanding framework with the use of deep neural networks for social and behavioural analysis. The integration of information fusion, person and object detection, social signal understanding, behaviour understanding, and context understanding plays a harmonious role to elicit social behaviours. Three systems, including depression detection, activity recognition and cognitive impairment screening, are developed to evidently demonstrate the importance of social intelligence. The study considerably contributes to the cumulative development of social computing and health informatics. It also provides a number of implications for academic bodies, healthcare practitioners, and developers of socially intelligent agents.

**Keywords:** Artificial Intelligence (AI) · Social intelligence · Deep neural networks · Social behaviours

## 1 Introduction

The landscape of social computing has evolved, with the proliferation of smaller, more powerful devices, such as mobile phones, tablets and wearable devices.

Having all these advanced technologies available at our fingertips, people are now using them in everyday life, introducing new habits and generating new forms of data.

The social computing paradigm has been moving beyond capturing information toward focusing on social intelligence [29]. As a vital facet of human intelligence, social intelligence is the capability to understand oneself and to understand others. The development of social intelligence systems entails recognising social and behavioural patterns from the new types of data and providing in-depth analysis of social signals for better human support. This paper aims to address major boundaries of social computing capabilities and social signal processing by introducing a social behaviour understanding platform with the use of deep neural networks.

With the rapid advancement of artificial intelligence (AI), deep learning utilises complex networks of artificial neurons to provide new ways to investigate human interactions in various contexts. We propose a deep learning framework for understanding social signals and behaviours from an individual or a group of people.

The niche nature of the previous generations of approaches and devices restricted the types and possibilities for social behaviour analysis. With state-of-the-art technologies today, we introduce the design and implementation of social intelligence systems for activity recognition, behavioural analysis, and health assessment. The paper demonstrate three use cases of social intelligence.

- **Depression detection** aims to develop a social intelligence system, that uses machine learning techniques, to classify vocal features present in a depressed individual’s voice. Utilising smartphone microphones, to determine if an individual suffers from depression through a mobile application.
- **Activity recognition** aims to utilise smartphones, activity trackers and smartwatches, to collect accelerometer sensor data, for the classification of human activities. Proceeded by machine learning and deep learning techniques, to predict patient activities, for a fall prevention mobile application.
- **Cognitive impairment screening** aims to build a tool to assess cognitive disorders based on individual’s writings and movements with the use of convolutional neural networks (CNN).

Based on academic foundations, the study contributes to the cumulative development of social intelligence and mobile health. It draws out many implications for academic theorists and healthcare practitioners.

The structure of the paper is as follows. Firstly, we review the literature background of our study in Sect. 2. Next, we present our social behaviour understanding architecture with the design concepts and three use cases of social intelligence systems. Lastly, the paper is concluded with findings and contributions.

## 2 Literature Background

### 2.1 Social Intelligence

Social intelligence is the ability to detect, interpret and react to human social and behavioural cues. These cues have many facets, ranging from physical appearance to vocal and facial features and gestures. Innately present in humans, social intelligence is a vital skill for understanding human behaviour, attributing significant impact on people’s lives, to this form of intelligence [1].

With the advance of Internet technologies, social computing has been moving beyond social informatics towards emphasising on social intelligence [29]. The field of implementing social intelligence in computers, is named social signal processing (SSP) [27]. Computers are relatively untrained to comprehend the aforementioned social signals in most current applications. With existing computing capabilities, context-independent tasks like arithmetic and retrieval operations can be performed without issue; however, the current state of computing is struggled to handle context-dependent tasks, such as virtual-reality applications. Incapable of realising the full potential of Internet-of-Things (IoT) networks as it is unable to utilise the data generated to predict actions or needs [27].

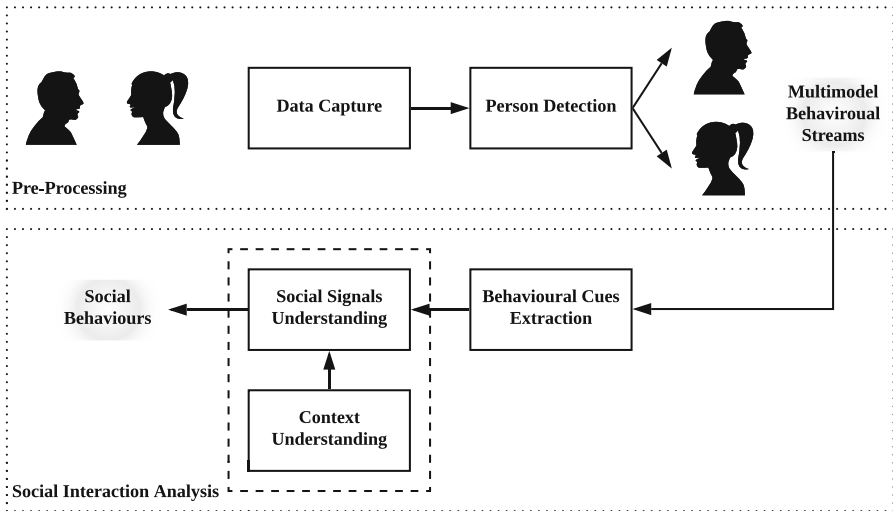


Fig. 1. Machine analysis of social signals (Vinciarelli et al.) [27]

On the other hand, in regards to the ability to observe social signals, human tend to have fluctuating performance, whereas computers have more consistent performance. This indicates that humans are not fully utilising present social and behavioural cues, relying on more scenario-oriented contextual cues. While machines are able to utilise the cues more extensively [5, 20]. Vinciarelli et al.

[27] proposed a popular framework for machine analysis of social signals and behaviours as shown in Fig. 1.

Achieving social intelligence will open up opportunities to a whole myriad of new applications. The development of social intelligence systems, therefore, becomes essential to stimulate greater availability of approaches and methodologies. To enable researchers and administrator to select the optimal approach, a guideline with clear objective and procedures will be invaluable.

## 2.2 Machine Learning for Cognitive and Social Behavioural Detection

The ever-growing popularity of artificial intelligence has led it to be applied in numerous fields of study. Empowering the discovery of novel applications, and thoroughly testing its limits. This trend has drawn focus into the utility of machine learning in health care [18].

Many researchers have investigated the accuracy and viability of incorporating or utilising machine learning, with existing methodologies. The research results have proven capability of artificial intelligence at diagnosing various ailments and disorders, displaying high levels of accuracy, with the opportunity for further improvement [21]. Wall et al. demonstrated an opportunity to improve healthcare methodology for diagnosing mental disorders and decrease healthcare costs [28]. The use of AI in computer games has also been explored to evaluate human behaviour [9]. It was bound to specific behavioural preferences with the ability to simulate a certain degree of human behaviours.

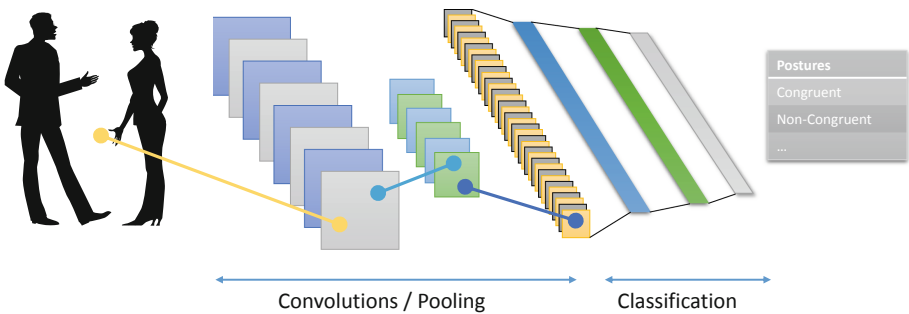


Fig. 2. Convolutional neural networks for posture detection

With recent breakthroughs in AI, deep learning has been well recognised as the next suitable wave of machine learning for social and behavioural analysis. Deep learning a multi-layered neural network, to steadily draws higher-level features from the input. Each deep neural network layer transforming the data to increasingly abstract representations of the input. A common deep learning implementation of interest is convolution neural networks (CNN), often used

for image and video analysis [13], as shown in Fig. 2. The CNN architecture consists of thousands to millions of artificial neurons in multiple layers, including convolutional layers, pooling layers, or fully connected activation layers. CNNs have been proven to perform with the better efficiency when it comes to vision and speech classification tasks, as shown in [8,10]

### 3 Social Behaviour Understanding Using Deep Neural Networks

In recent research, the use of artificial intelligence has been widely exploited to analyse the behavioural and social cues in social interactions [9,28]. This research trend has led to better understandings of human beings, where new knowledge and patterns of behavioural, social, and contextual cues are constantly discovered. With the new computing capabilities, we propose a framework for social behaviour understanding using state-of-the-art deep learning, as shown in Fig. 3. Multiple constructs are adapted based on the original framework for machine analysis of social and behavioural signal processing from Vinciarelli, Pantic and Bourlard [27].

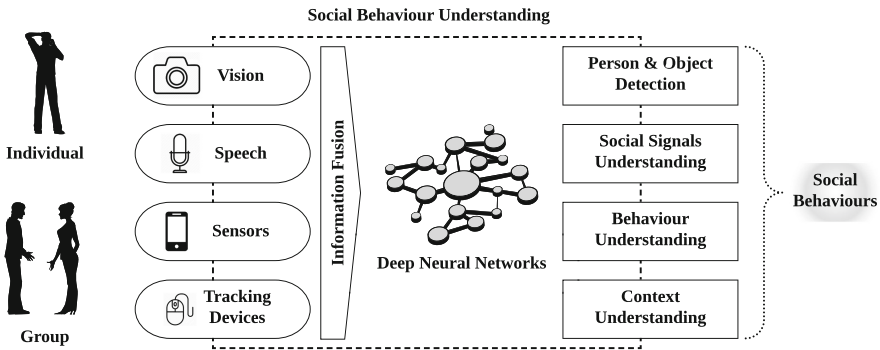


Fig. 3. Social behaviour understanding using deep neural networks

In our framework, the role of socially intelligent agents has been evolved to more closely emulate humans, thereby shortening the gap between machines and humans. We emphasise on fully realising the capabilities of artificial intelligence for robust and versatile detection of social and behavioural signals. The use of deep neural networks aims to simulate cerebral activities to create new ways of understanding data and making inferences. The proposed framework consists of FIVE (5) key components: (i) Information Fusion, (ii) Person and Object Detection, (iii) Social Signal Understanding, (iv) Behavioural Understanding, and (v) Context Understanding.

- **Information Fusion.** With the increasing ubiquity of Internet-of-Things (IoT) technology, new types of sensors, tracking devices, and mobile equipment have been widely introduced [17]. These capabilities allow data capture of multimodal inputs, including visual, audible and movement data. Information fusion strategies are required to eliminate uncertainty and reliability issues in such data. The process of information fusion integrates multiple data sources into a robust, accurate and consistent input body for deep learning. Lee et al. suggested a hierarchical decomposing method to handle the data at three different levels: raw sensor data fusion, feature level fusion, and decision level fusion [14].
- **Person and Object Detection.** Social intelligence entails interactions among multiple agents, including humans and objects. The traditional methods in person and object detection are typically developed based on limited feature extraction and shallow learning models [31]. The recent breakthroughs in deep learning have raised a new ground for detecting objects with high confidence in audios, images and videos. Convolutional neural network models perform distinguishably, with a variety of network architectures, training and optimisation strategies. It is also important to note that detection models can be integrated within a single multimodal neural network architecture.
- **Social Signal Understanding.** Social signals occur in everyday situations, which include many social cues such as attention, empathy, politeness, or agreement. Social signal processing has drawn huge research efforts to understand human interactions in an automated and continuous manner [7, 16, 22]. Deep evolutionary spatial-temporal networks were suggested to extract both temporal and spatial features of facial expressions, which outperformed traditional approaches in a large margin [30]. Similarly, deep learning has been used to learn social signals from appearance, gesture and posture [3].
- **Behavioural Understanding.** Human behaviours play a vital role in shaping the perception of human interactions. Investigating behavioural cues, hence, allows intelligent agents to elicit social signals with a higher degree of support. This is also applicable to individual behaviours, captured with or without social interactions, due to temporal dynamics of social behaviours. This framework suggests behavioural understanding component is a good supplement to develop social intelligence.
- **Context Understanding.** Understanding social and behavioural signals is not without contextual information such as location, time, or situation. The contexts are tightly associated with communicative intention; thus, it is critical to consider their dynamics in social behaviour analysis. With new mobile and sensor capabilities, the presence of context data can be embedded into multimodal deep neural networks in various ways [26].

With the recent development in deep neural networks, the fusion of multimodal understanding units opens new pathways to analyse and recognise social behaviours. Frequently, social, behavioural, and contextual dimensions of the data contain both unique and overlapped signals; thus, training using deep neural networks is a viable option for the development of intelligent agents.

Cross-modality transformers are increasingly explored to address the challenge of multifacet representation learning and pattern recognition [25], such as social intelligence.

## 4 Development of Social Intelligence Systems

Based on the Social Behaviour Understanding framework, the study takes an important step to develop three social intelligence systems for health assessment. They utilise deep neural networks to detect social and behavioural cues using real-time data for timely interventions.

The paper aims to bring collaborative care to the next level, where social behaviours are recognised and exchanged with social support agents.

### 4.1 Use Case 1: Depression Detection

Current methods of diagnosing depression are time-consuming and archaic, with only minor improvements being made regarding its process [2]. The process requires psychiatrists to initially screen patients through questionnaires utilising scales, including but not limited to: Centre for Epidemiological Studies Depression Scale (CES-D) [24], Beck Depression Inventory (BDI) [15] or PRIME-MD Patient Health Questionnaire [11, 12].

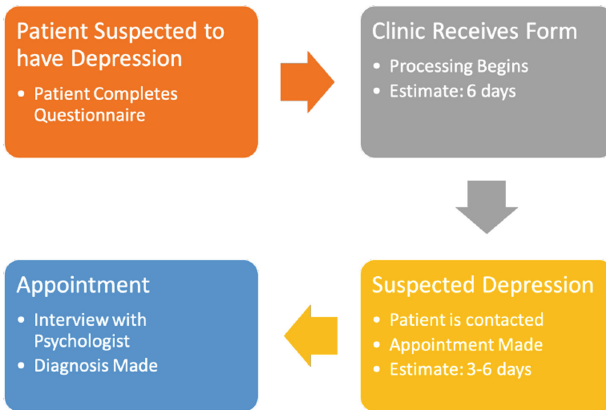


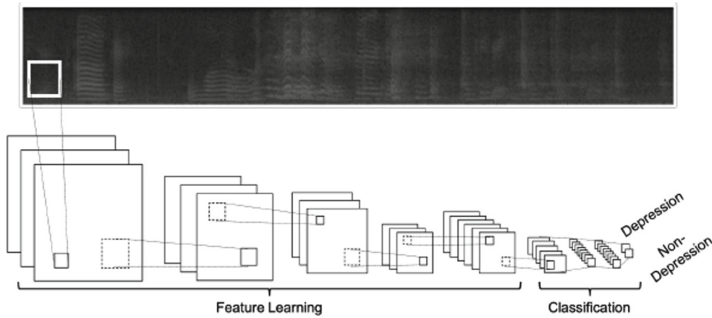
Fig. 4. Process of diagnosing depression. [2]

After screening, individuals suspected of suffering from depression would be contacted to arrange for an additional appointment to affirm the diagnosis [2]. This process could entail a minimum wait period of 2 weeks before patients are diagnosed and can begin treatment, as depicted in Fig. 4. Currently, much of the waiting period, is devoted to the processing of questionnaires, and the scheduling

for an appointment. Each of the patient's responses must be evaluated by a psychologist, after which, the result is only an indication of whether the patient suffers from depression.

Few studies and experiments have been conducted to evaluate the effectiveness of speech-based depression detection. However, 2 studies, Depression Speaks [4] and Depression Detect [23] used, The Distress Analysis Corpus-Wizard of Oz (DAIC-WOZ) database [6], to experiment with speech-based depression detection. Utilising machine and deep learning, respectively, to extract features and classify depression from speech audio. Nevertheless, there is no mobile application currently, that is able to detect symptoms of depression, based on their voice. By attempting to improve the medical industry through novel means, advances in the methods of mental health diagnosis could be made in the future.

This system aims to explore the feasibility of using a mobile application to detect patients with depression based on their vocal features. Allowing it to predict in real time, if an individual displays symptoms of depression.



**Fig. 5.** Convolutional neural networks for spectrogram

**Data Collection.** Data is collected from the built-in microphone of, in this case, a Google Pixel 2 XL. The output format is a Pulse Code Modulation (PCM) file, which is a file format that represents a digitization of analog audio. The sampling rate is 44100 Hz which means that there are 44100 samples of audio frequency per second.

The training dataset used for this project is The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) database by the University of Southern California (USC) [6]. It contains 189 clinical interview sessions designed to support the diagnosis of psychological distress conditions. Each session comprises of a transcript of the interview, an audio recording and the facial features of the participant.

**Development.** We processed the data into visual representations using convolutional neural networks in Fig. 5. And a prototype mobile application was



developed using the Android Platform. Upon opening the application, the user is prompted to enter the Name and ID of the patient before beginning the session. Once a session is started, the user can start the recording process to detect possibility of having depression, as shown in Fig. 6.

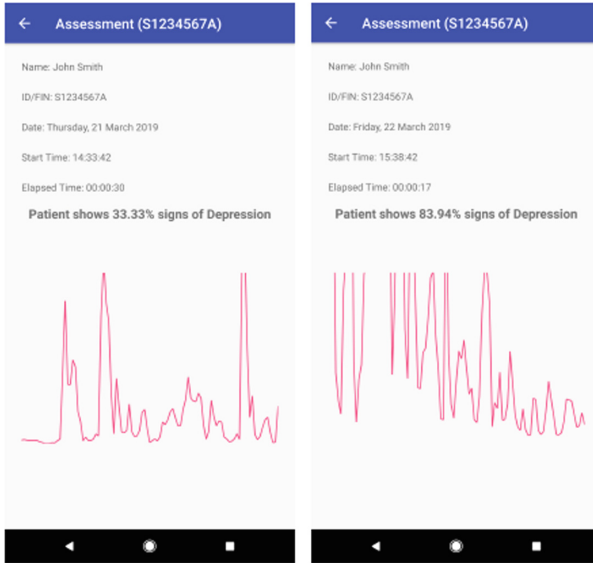


Fig. 6. Screenshots of the application processing live data of 2 different users.

#### 4.2 Use Case 2: Activity Recognition for Fall Detection and Prevention

Patient accidents in hospitals are of significant concern, especially if they occur with the elderly. Globally, a third of adults over 65 years old, falls once a year. These accidents could lead to additional harm, such as further injury, complications and loss of mobility. Therefore, this social intelligence system aims to

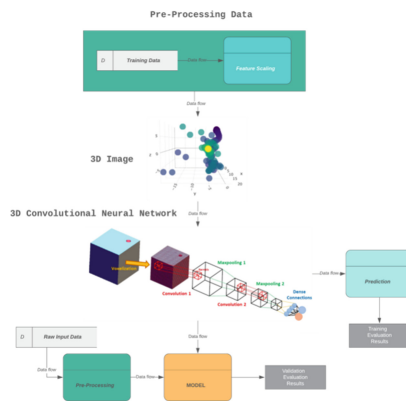


Fig. 7. Fall-risk wrist tag

recognise patients' activities for fall detection and prevention. High-risk patients are given green wrist tags, as shown in Fig. 7, and a green label, which are set at the panel of their beds, and they are required to always be continuously monitored in the system.

**Data Collection.** The widespread proliferation of smartphones has made low-cost smartphones equipped with a variety of sensors commonplace. This project would explore the use of mobile technologies to enhance the fall detection and prevention strategy further. The training dataset from UniMiB-SHAR was used as it was an open dataset available online. The UniMib-SHAR dataset consists of 17 different kinds of activities, divided into nine different types of daily activities such as walking, running, etc. and eight different types of falls such as fall forward, fall left, etc. There are a total of 7759 daily activities, and 4192 falls respectively.

The social intelligence would be performed in real-time with the assistance of a smartphone, with an in-built accelerometer. The patient would carry a smartphone, with the mobile application deployed to it and perform different activities. Logged accelerometer data of 1-s intervals would be sent through an API call to the server for processing.



**Fig. 8.** 3D convolution neural networks for behavioural analysis

**Development.** We employed 3D convolutional neural networks to analyse the behavioural data, as shown in Fig. 8. The 4D tensor would then be passed through the many convolutional, pooling, batch normalisation, flatten, and multi-perceptron layers to finally the activation layer. This would generate the 3D CNN model and show the training accuracy of the model. Firebase was used to provide real-time database as a backend service to store and return the information of the patient's name, activity and time of activity to be displayed on the clinician and patient applications.

### 4.3 Use Case 3: Cognitive Impairment Screening

75 million people are predicted to be affected by dementia by 2030 [19]. With individuals older than 65 years, at much greater risk of developing a form of cognitive impairment. Hospitals employ a battery of cognitive tests, to detect cognitive impairments. The tests commonly take the form of writing and drawing examinations, requiring the completion of tasks ranging from simple instructional writing, to complex memory-based drawings.

This social intelligence system aims to predict the risk of cognitive impairments with the use of hand writings and pen movements.

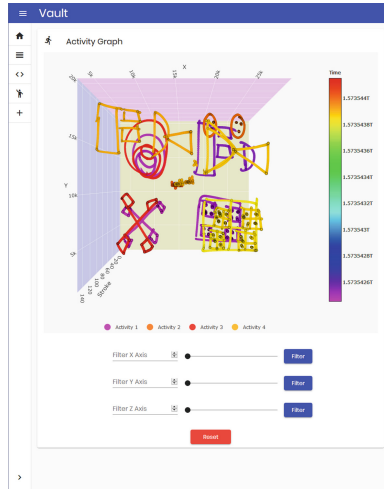


Fig. 9. 3D visualisation of subject data

**Data Collection.** In this study, participants undergo a series of cognitive manual handwriting tests; and electronic tablet and pen was used to capture writing and in-air (hover) trajectory.

**Development.** We developed 3D images from feature scaling the training and test data, will be represented by a 4D tensor. Then, the tensors would then be passed through a 3D deep neural networks for cognitive impairment detection. In our development, Angular 8 and JavaScript framework was utilised for the frontend, Flask and Python for the backend, and MongoDB for the database. Finally, after configuration, a 3D model will created from the variables selected, with the option to filter the X, Y and Z axis, for inspection as shown in Fig. 9.

## 5 Conclusion

Social intelligence systems are promising to revolutionise people's everyday life. Our study proposes a social behaviour understanding framework which performs recognition of social and behavioural signals for the development of socially intelligent systems. The framework consists of five key components: (i) Information Fusion, (ii) Person and Object Detection, (iii) Social Signal Understanding, (iv) Behavioural Understanding, and (v) Context Understanding. Cross-modality analysis of social, behavioural, and contextual information with the use of deep neural networks is suggested to bring social intelligence to the next level. Moreover, we developed three social intelligence systems for depression detection, activity recognition, and cognitive impairment screening.

Our study contributes to the cumulative theoretical development of social computing and artificial intelligence. The uniqueness of social intelligence is evidently demonstrated to shed light on new applications. We hope our social behaviour understanding framework provides meaningful guidelines on the development of new types of social computing systems. This paper is not an end, but rather a beginning of future research as we are looking into ways of further refining and evaluating our social intelligence systems.

## References

1. Albrecht, K.: *Social Intelligence: The New Science of Success*. Wiley, Hoboken (2006)
2. Beck, A., Ward, C., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measuring depression. *Arch. Gen. Psychiatr* **4**, 561–571 (1961)
3. Chen, H., Liu, X., Li, X., Shi, H., Zhao, G.: Analyze spontaneous gestures for emotional stress state recognition: a micro-gesture dataset and analysis with deep learning. In: *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–8. IEEE (2019)
4. Eyben, F., Weninger, F., Wöllmer, M., Shuller, B.: *Open-source media interpretation by large feature-space extraction*. TU Munchen, MMK (2016)
5. Gold, J.M., Tadin, D., Cook, S.C., Blake, R.: The efficiency of biological motion perception. *Perception Psychophys.* **70**(1), 88–95 (2008)
6. Gratch, J., et al.: The distress analysis interview corpus of human and computer interviews. In: *LREC*, pp. 3123–3128. Citeseer (2014)
7. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. *Int. J. Synth. Emotions (IJSE)* **1**(1), 68–99 (2010)
8. Hershey, S., et al.: CNN architectures for large-scale audio classification. In: *2017 IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135. IEEE (2017)
9. Hildmann, H.: Designing behavioural artificial intelligence to record, assess and evaluate human behaviour. *Multimodal Technol. Interact.* **2**(4), 63 (2018)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)
11. Kroenke, K., Spitzer, R.L.: The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr. Ann.* **32**(9), 509–515 (2002)

12. Kroenke, K., Spitzer, R.L., Williams, J.B.: The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16**(9), 606–613 (2001)
13. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
14. Lee, H., Park, K., Lee, B., Choi, J., Elmasri, R.: Issues in data fusion for health-care monitoring. In: *Proceedings of the 1st International Conference on Pervasive Technologies Related to Assistive Environments*, pp. 1–8 (2008)
15. Lewinsohn, P.M., Seeley, J.R., Roberts, R.E., Allen, N.B.: Center for epidemiologic studies depression scale (CES-D) as a screening instrument for depression among community-residing older adults. *Psychol. Aging* **12**(2), 277 (1997)
16. LiKamWa, R., Liu, Y., Lane, N.D., Zhong, L.: MoodScope: building a mood sensor from smartphone usage patterns. In: *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 389–402 (2013)
17. Nguyen, H.D., Jiang, Y., Eiring, Ø., Poo, D.C.C., Wang, W.: Gamification design framework for mobile health: designing a home-based self-management programme for patients with chronic heart failure. In: Meiselwitz, G. (ed.) *SCSM 2018. LNCS*, vol. 10914, pp. 81–98. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91485-5\\_6](https://doi.org/10.1007/978-3-319-91485-5_6)
18. Nguyen, H.D., Poo, D.C.C.: Automated mobile health: designing a social reasoning platform for remote health management. In: Meiselwitz, G. (ed.) *SCSM 2016. LNCS*, vol. 9742, pp. 34–46. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-39910-2\\_4](https://doi.org/10.1007/978-3-319-39910-2_4)
19. Organization, W.H., et al.: *Global action plan on the public health response to dementia 2017–2025* (2017)
20. Pollick, F.E., Lestou, V., Ryu, J., Cho, S.B.: Estimating the efficiency of recognizing gender and affect from biological motion. *Vis. Res.* **42**(20), 2345–2355 (2002)
21. Rutkowski, T.M., Abe, M.S., Koculak, M., Otake-Matsuura, M.: Cognitive assessment estimation from behavioral responses in emotional faces evaluation task-AI regression approach for dementia onset prediction in aging societies. *arXiv preprint arXiv:1911.12135* (2019)
22. Schuller, B., et al.: The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, Lyon, France (2013)
23. Scibelli, F., et al.: Depression speaks: automatic discrimination between depressed and non-depressed speakers based on nonverbal speech features. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6842–6846. IEEE (2018)
24. Spitzer, R.L., Kroenke, K., Williams, J.B., Patient Health Questionnaire Primary Care Study Group, et al.: Validation and utility of a self-report version of primemd: the PHQ primary care study. *JAMA* **282**(18), 1737–1744 (1999)
25. Tan, H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019)
26. Vahora, S., Chauhan, N.: Group activity recognition using deep autoencoder with temporal context descriptor. *Int. J. Next-Gener. Comput.* **9**(3) (2018)
27. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: survey of an emerging domain. *Image Vis. Comput.* **27**(12), 1743–1759 (2009)
28. Wall, D.P., Dally, R., Luyster, R., Jung, J.Y., DeLuca, T.F.: Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS ONE* **7**(8) (2012)
29. Wang, F.Y., Carley, K.M., Zeng, D., Mao, W.: Social computing: from social informatics to social intelligence. *IEEE Intell. Syst.* **22**(2), 79–83 (2007)

30. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutionary spatial-temporal networks. *IEEE Trans. Image Process.* **26**(9), 4193–4203 (2017)
31. Zhao, Z.Q., Zheng, P., Xu, S.T., Wu, X.: Object detection with deep learning: a review. *IEEE Trans. Neural Netw. Learn. Syst.* **30**(11), 3212–3232 (2019)



# Materialism and Facebook Usage: Could Materialistic and Non-materialistic Values Be Linked to Using Facebook Differently?

Roshan Rai<sup>(✉)</sup> , Jade Blocksidge, and Mei-I Cheng 

De Montfort University, Leicester, UK  
rrai@dmu.ac.uk

**Abstract.** Materialism is a set of human values that places importance on the symbolic value of money or material goods. Furthermore, materialistic values have been associated with Internet usage, and also social media usage. The current research investigates this relationship further by specifically examining whether those with *more materialistic* values might use social media (Facebook) in different ways to those with *less materialistic* values. Self-report measures were collected from 108 participants. It was found that the higher the importance (*extrinsic importance*) attached to materialistic values, the more time spent posting photos, but the less time spent chatting on Facebook messenger and less time posting links. The higher the perceived likelihood (*extrinsic likelihood*) of achieving materialistic values, the more reported time posting status updates, but the less time spent chatting on Facebook messenger and less time posting links. Conversely, the higher the importance attached to non-materialistic values (*intrinsic importance*) the more reported time chatting on Facebook messenger, more time spent posting links, but less time spent posting photos. And the higher the reported likelihood of achieving non-materialistic values (*intrinsic likelihood*) the more reported time chatting on Facebook messenger, more time spent posting links, but less time spent posting status updates. However, neither self-reported time checking Facebook, nor self-reported attention paid to advertising were related to either materialistic or non-materialistic values. Overall, the findings indicate that certain activities on Facebook can be associated with both materialistic and non-materialistic values.

**Keywords:** Social media · Facebook · Materialistic values · Non-materialistic values

## 1 Introduction

Social Media and Social Networking Sites have become an everyday activity for many people, providing an essential electronic medium for social interaction. Undoubtedly, one of the current leaders in Social Media is Facebook, which NBC reported as having 2.27 billion users worldwide as of 2018 [1]. Unlike text-based platforms like Twitter, and image-based platforms like Instagram, Facebook is a truly mixed-platform allowing for a range of activities: such as the sharing of text, images, and web links [2].

## 1.1 The Internet, Social Media and Materialism

Traditionally, materialism has been defined as the importance people place on worldly possessions [3, 4]; although materialism has also been defined as a belief in materialistic activities defining the self, and being linked to desirable symbolic values (e.g., power) [5, 6]. Kasser and Ryan [7] defined materialistic (extrinsic) values as centering around three main preoccupations – financial success, social recognition, and an appealing appearance. This contrasts with non-materialistic (intrinsic) values of self-acceptance, affiliation, community feeling, and physical fitness.

Research has demonstrated an association between the Internet and materialism, with greater self-reported time spent using the Internet (but not time spent watching TV, nor reading newspapers/magazines) positively predicted both materialistic values and better brand knowledge (being able to identify brand logos) [8]. Additionally, in both American and Chinese samples, the intensity of social media usage was found to be positively related to materialistic values [9]. And furthermore, more time on Facebook specifically was associated with materialistic values, with materialists using Facebook to satisfying materialistic goals [10].

These findings, naturally, raise the question of why the Internet and social media might be associated with materialism. As Gerbner [11] claimed in his cultivation theory, agents of mass communication (originally referring to television) transmit mass messages that profoundly affect people's perception and values. And as Cultivation theory further argued, marketing is a factor influencing the cultural environment, and subsequently people's values [12]. There is evidence that this could be applied to the Internet (as the new mass communication technology) and social media. Marketers have already recognized the importance of Facebook, with many top brands maintain a presence on Facebook [13]. And some argued that social media is a more effective method of advertising, compared to more traditional forms of media such as television or radio [14], with social media able to reach specific target audiences and place advertisements more cheaply than traditional media [15].

Furthermore, there is evidence that social media usage can be linked to materialism and consumption. Social media marketing has been shown to specifically affect adolescents' attitudes towards certain brands [16]. Advertising on social media can also raise consumer engagement [13]. And increased social media usage also has been associated with increased brand consciousness as well as an intention to buy luxury products in a sample of millennials [17]. Furthermore, some claimed that the various consumption-related messages online could be a factor in raising levels of materialism in young adults [18]. And others argued that consumerist messages transmitted through Chinese social media platform Weibo have influenced the growth of materialism and hedonism in China [19].

Given previous research [e.g., 10], the first purpose of the current research is to see whether time checking Facebook would be associated with higher levels of materialistic values in our sample (RQ1.).

The second purpose of the current research is to investigate whether paying greater attention to advertising on Facebook specifically will be associated with higher level of materialistic values (RQ2.).



## 1.2 Materialistic Values and Facebook Usage

Previous research demonstrated an association between materialistic values and greater time using Facebook, as it is claimed that materialists use Facebook to satisfy materialistic goals [10]. And if materialists use Facebook to satisfy their specific needs, then it could furthermore be the case that materialists use Facebook in different ways to less materialistic people. Given the nature of both materialistic values (extrinsic) and non-materialistic values (intrinsic) [7] it could be the case that some activities might be more attuned to extrinsic values and some more attuned with intrinsic values. For example, some materialists use Facebook to gain positive affect from positive self-presentation [10], which could be linked to the extrinsic (materialistic) values of seeking social recognition [7]. Conversely, research has found that the motivation to share on Facebook include the drive to share information with others and to interact with others [20], which could be related to the non-materialistic (intrinsic) values of affiliation and community feeling (drives to connect and help others). Overall, materialistic and non-materialistic values could be related to using Facebook in different ways, and the current research attempts to investigate any possible associations.

The current research will investigate whether different types of Facebook usage are associated with materialistic values (RQ 3.).

The current research will also investigate whether different types of Facebook usage are associated with non-materialistic values (RQ4.).

## 2 Method

### 2.1 Participants

One hundred and eight undergraduate psychology participants were recruited from a university in the Midlands of England, U.K. All took part in an online study for course credit. The sample comprised of 94 females and 14 males, with ages ranging from 18 to 50 ( $M = 20.57$ ,  $SD = 5.79$ ).

### 2.2 Materials

Participants were given a demographic questionnaire that asked their age, gender, as well as two questions about their social media usage: firstly, how many times on average that they check Facebook a day (free typing a response) and secondly how often they pay attention to advertising on Facebook (1 never, 2 rarely, 3 sometimes, 4 often, 5 very frequently).

Participants also filled in a questionnaire on the specific activities they most used Facebook for. We initially based our measure on Junco's [21] Facebook Usage Scale. After pilot work, seven items were adapted from Junco's original scale, asking participants how often they: posted status updates, shared links, posted photos, chatted on Facebook messenger, checked to see what someone is up to, commented on content, and viewed other people's photos. In addition, after pilot work, an additional item (not from Junco's original measure) was added asking participants how often they: added friends. This made a total of eight items. Participants indicated on a 5-point scale (1 never, 2

rarely, 3 sometimes, 4 often, 5 very frequently) how frequently they performed each activity.

Materialistic and non-materialistic values were measured using Kasser and Ryan's [7] 42-item Aspiration Index. Participants had to indicate both the importance and likelihood of achieving both extrinsic (materialistic) and intrinsic (non-materialistic) values on a 5-point scale (higher numbers equaling higher levels). Extrinsic values comprised of three subscales, measuring how much participants valued: financial success, social recognition, and an attractive appearance. Intrinsic values consisted of four sub-scales, measuring how much participants valued: affiliation, community feeling, physical fitness, and self-acceptance. Cronbach's alpha showed good internal consistency for extrinsic importance ( $\alpha = .87$ ), extrinsic likelihood ( $\alpha = .88$ ), intrinsic importance ( $\alpha = .87$ ), and intrinsic likelihood ( $\alpha = .89$ ).

### 2.3 Procedure

Participants completed the survey online, with the survey housed on the online research platform Qualtrics. Participants first filled out the Demographics Questionnaire, then the Facebook Usage Questionnaire, and lastly the Aspiration Index.

## 3 Results

Table 1 presents the means and standard deviations for all variables in the current study, and the zero-order correlations are shown in Table 2. Firstly, tests of assumptions were conducted for the regression analyses of the four models being investigated. The test of Cook's distance indicated there to be no outliers, due to the statistical value being smaller than 1. To assess multicollinearity the variance inflation factor analyses (VIF) was carried out for all four models, and displayed there to be no multicollinearity, as each VIF was less than 2 ( $VIF < 2$ ). The Durbin-Watson statistic demonstrated that adjacent residuals were uncorrelated, with the value for each model being close to 2, indicating that the assumption for independence of errors was met.

In the present study, four multiple linear regressions were conducted using the stepwise method. The stepwise method was applied due to the large number of predictor variables in the current study, and due to its utility in being able to identify the most significant relationships between variables.

### 3.1 Materialistic Values and Times Checking Facebook a Day, Attention to Advertising on Facebook, and Types of Facebook Usage

Two multiple linear regressions were used to examine whether times checking Facebook a day, attention to advertising on Facebook, and types of Facebook usage would predict materialistic values (both for extrinsic importance and extrinsic likelihood). The first multiple linear regression was used to examine whether times checking Facebook a day, attention to advertising on Facebook, and types of Facebook usage would predict extrinsic importance. Using the stepwise method, it was found that a three-predictor model accounted for 15.9% of the variance,  $F(3, 104) = 6.57, p < .001$ , with an

**Table 1.** Mean and standard deviations of measures

|  | Mean | SD   |
|--|------|------|
| 1. Time checking Facebook                | 6.57 | 7.40 |
| 2. Attention to advertising              | 2.28 | .91  |
| 3. Posting status updates                | 1.77 | .80  |
| 4. Sharing links                         | 2.33 | 1.05 |
| 5. Posting photos                        | 2.31 | .84  |
| 6. Chatting on Facebook messenger        | 3.81 | 1.15 |
| 7. Checking to see what someone is up to | 2.86 | 1.07 |
| 8. Commenting on content                 | 2.69 | 1.06 |
| 9. Viewing other people's photos         | 3.31 | .98  |
| 10. Adding friends                       | 2.69 | .82  |
| 11. Intrinsic importance                 | .61  | .27  |
| 12. Intrinsic likelihood                 | .35  | .18  |
| 13. Extrinsic importance                 | -.81 | .36  |
| 14. Extrinsic likelihood                 | -.47 | .24  |

$R^2$  of .16 (Adjusted  $R^2 = .14$ ). Further to this, Cohen's  $f^2 = .19$  suggested a medium effect size. Extrinsic importance (materialistic) was negatively associated with chatting on Facebook messenger ( $\beta = -.27$ ,  $t(104) = -2.88$ ,  $p = .005$ ), negatively associated with sharing links ( $\beta = -.28$ ,  $t(104) = -3.02$ ,  $p = .003$ ), and positively associated with posting photos ( $\beta = .28$ ,  $t(104) = 2.88$ ,  $p = .005$ ).

The second multiple linear regression was used to examine whether times checking Facebook a day, attention to advertising on Facebook, and types of Facebook usage would predict extrinsic likelihood. Using the stepwise method, it was found that a three-predictor model accounted for 17.6% of the variance,  $F(3, 104) = 7.42$ ,  $p < .001$ , with an  $R^2$  of .18 (Adjusted  $R^2 = .15$ ). Further to this, Cohen's  $f^2 = .21$  suggested a medium to large effect size. Sharing links was found to negatively predict extrinsic likelihood ( $\beta = -.34$ ,  $t(104) = -3.54$ ,  $p < .001$ ), chatting on Facebook messenger was found to negatively predict extrinsic likelihood ( $\beta = -.29$ ,  $t(104) = -2.55$ ,  $p = .012$ ), and posting status updates was found to positively predict extrinsic likelihood ( $\beta = .23$ ,  $t(104) = 2.44$ ,  $p = .016$ ).

### 3.2 Non-materialistic Values and Times Checking Facebook a Day, Attention to Advertising on Facebook, Types of Facebook Usage

Two further multiple linear regressions were used to examine whether times checking Facebook a day, attention to advertising on Facebook, and types of Facebook usage would predict non-materialistic values (both for intrinsic importance and intrinsic likelihood). The first multiple linear regression was to examine whether times checking Facebook

Table 2. Zero-order Correlations

|  | 1     | 2    | 3     | 4      | 5     | 6      | 7     | 8     | 9     | 10   | 11     | 12     | 13    |
|--|-------|------|-------|--------|-------|--------|-------|-------|-------|------|--------|--------|-------|
| 1. Time checking Facebook                | –     |      |       |        |       |        |       |       |       |      |        |        |       |
| 2. Attention to advertising              | .15   | –    |       |        |       |        |       |       |       |      |        |        |       |
| 3. Posting status updates                | .16   | .10  | –     |        |       |        |       |       |       |      |        |        |       |
| 4. Sharing links                         | .17   | –.02 | .35** | –      |       |        |       |       |       |      |        |        |       |
| 5. Posting photos                        | .36** | .11  | .54** | .30**  | –     |        |       |       |       |      |        |        |       |
| 6. Chatting on Facebook messenger        | .32** | .14  | –.01  | .07    | .22*  | –      |       |       |       |      |        |        |       |
| 7. Checking to see what someone is up to | .31** | .13  | .15   | .05    | .27** | .29**  | –     |       |       |      |        |        |       |
| 8. Commenting on content                 | .35** | .22* | .32** | .38**  | .41** | .33**  | .29** | –     |       |      |        |        |       |
| 9. Viewing other people's photos         | .19   | .08  | .01   | –.04   | .13   | .31**  | .58** | .25** | –     |      |        |        |       |
| 10. Adding friends                       | .23*  | .02  | .02   | .09    | .19   | .34**  | .28*  | .38** | .39** | –    |        |        |       |
| 11. Intrinsic importance                 | .07   | –.13 | –.12  | .22*   | –.13  | .22*   | –.04  | –.03  | –.07  | .03  | –      |        |       |
| 12. Intrinsic likelihood                 | .05   | –.11 | –.12  | .27**  | .09   | .25**  | .09   | .14   | .12   | .07  | .64**  | –      |       |
| 13. Extrinsic importance                 | –.07  | .13  | .12   | –.22*  | .13   | –.22*  | .04   | .03   | .07   | –.03 | 1.0**  | –.64** | –     |
| 14. Extrinsic likelihood                 | –.05  | .11  | .12   | –.27** | –.09  | –.25** | –.09  | –.14  | –.12  | –.07 | –.64** | –1.0** | .64** |

\* $p < .05$ , \*\* $p < .01$

a day, attention to advertising on Facebook and types of Facebook usage would predict intrinsic importance. Using the stepwise method, it was found that a three-predictor model accounted for 15.9% of the variance,  $F(3, 104) = 6.57, p < .001$ , with an  $R^2$  of .16 (Adjusted  $R^2 = .14$ ). Further to this, Cohen's  $f^2 = .19$  suggested a medium effect size. Chatting on Facebook messenger was found to positively predict intrinsic importance ( $\beta = .27, t(104) = 2.878, p = .005$ ), sharing links was found to positively predict intrinsic importance ( $\beta = .28, t(104) = 3.02, p = .003$ ), and posting photos was found to negatively predict intrinsic importance ( $\beta = -.28, t(104) = -2.88, p = .005$ ).

The final multiple linear regression was used to examine whether times checking Facebook a day, attention to advertising on Facebook, and types of Facebook usage would predict intrinsic likelihood. Using the stepwise method, it was found that a three-predictor model accounted for 17.6% of the variance,  $F(3, 104) = 7.42, p < .001$ , with an  $R^2$  of .18 (Adjusted  $R^2 = .15$ ). Further to this, Cohen's  $f^2 = .21$  suggested a medium to large effect size. Sharing links was found to positively predict intrinsic likelihood ( $\beta = .34, t(104) = 3.53, p < .001$ ), chatting on Facebook messenger was found to positively predict intrinsic likelihood ( $\beta = .29, t(104) = 2.55, p = .012$ ), and posting status updates was found to negatively predict intrinsic likelihood ( $\beta = -.23, t(104) = -2.44, p = .016$ ).

## 4 Discussion

Initially, in support of research question 3, it was found that materialistic values were associated with specific Facebook activities. However, these associations were slightly different according to whether participants were responding to how important (extrinsic importance) they thought these values to be, or how likely (extrinsic likeliness) they thought these materialistic values were to be achieved. Extrinsic importance was associated with more time spent posting photos on Facebook, but less time messaging others on Facebook messenger, and less time spent sharing links. Extrinsic likelihood was associated with more time spent posting status updates on Facebook, but less time messaging others on Facebook messenger, and less time spent sharing links. Conversely, and in support of research question 4, levels of non-materialistic (internal) values were also associated with specific Facebook activities. Although these associations were slightly different according to whether participants were responding to how important (intrinsic importance) they thought these values to be, or how likely (intrinsic likeliness) they thought these non-materialistic values were to be achieved. Intrinsic importance (importance attached to non-materialistic values) was associated with more time spent using Facebook messenger, more time spent sharing links, but less time spent posting photos. However, intrinsic likelihood was associated with more time spent using Facebook messenger, more time spent sharing links, but less time posting status updates. Surprisingly and in contrast to research question 1, self-reported time spent checking Facebook was not associated with materialistic values. And in contrast to research 2, greater self-reported attention to advertising was not associated with materialistic values. Research questions 1 and 2 were therefore not supported in the current study.

The current research found that (self-reported) higher levels of materialistic values were associated with reports of using Facebook in a specific way: namely spending

more time posting photos (for extrinsic importance) and posting more status updates (for extrinsic likelihood). Although there needs to be caution when interpreting causation from tests of association, this may not seem surprising as previous research has showed that materialists use Facebook to gain positive affect from positive self-presentation [10]. Although the current research does not specifically test these specific motivations, an obvious way to seek positive self-presentation on Facebook could (theoretically) be to post flattering photos of oneself, one's life and (when considering materialistic values) perhaps one's possessions. Furthermore, Facebook status updates could also be a way of promoting positive self-image, e.g., positing about positive events or achievements in one's life. Previous research claimed materialists use Facebook to satisfy their materialistic goals [10], and the posting of photos and status updates on Facebook could be examples of materialists doing exactly that. The current research goes beyond previous research though, to demonstrate that greater (self-reported) non-materialistic values were associated with spending less time positing photos (for intrinsic importance) and posting status updates (for intrinsic likelihood); this could possibly be a reflection that non-materialistic values are not greatly served by these particular activities.

In contrast to materialistic values, non-materialistic values were associated with greater self-reported time spent using Facebook messenger and posting links. Bearing in mind that caution should be applied to inferring causation from tests of association, one could speculate that this could possibly reflect how non-materialists use Facebook to satisfy their non-materialistic values. If central non-materialistic values include affiliation and community [7], then it might not seem so surprising that non-materialistic people spend more time using Facebook messenger to communicate with specific others, and use Facebook to share links with others. Furthermore, these findings might not be too dissimilar to previous research that has shown that motivations to share on Facebook include interacting with others and to share information [20]. Perhaps these motivations are more likely for those more intrinsically (non-materialistically) motivated. So, the current study goes beyond previous research to suggest that more materialistic individuals might use Facebook in a directly opposite manner to less materialistic individuals: with more materialistic individuals reporting less time spent using Facebook messenger, and sharing links. And future research could further investigate the specific underlying mechanisms being the association between materialists and non-materialists and using Facebook differently, which is particularly under-researched in the case of those higher in non-materialistic values.

Surprisingly, the current research did not find a link between self-reported time checking Facebook and materialistic values, in direct contrast to previous research showing such an association [10]. Additionally, the current research did not show a link between advertising and materialism, in contrast to previous research that suggested such a relationship [13, 16–19]. Both results were unexpected, but with the specific self-report questions employed (time spent on Facebook in a typical day, and attention paid to time on Facebook) these relationships did not materialize. It is not clear whether these associations do not always bear out in every sample, or whether the specific test questions employed nullified any potential associations; it could be noted that the test questions employed were rather brief. Perhaps more detailed measures of the attention paid to

Facebook advertising and time spent using Facebook might have produced different results. But future research could further investigate these questions.

Overall, the current research demonstrated that materialistic values and non-materialistic values are associated to different Facebook usage. The higher the reported importance of materialistic values (extrinsic importance) the more reported time posting photos. Whilst the higher the reported likelihood that these materialistic values (extrinsic likelihood) were to be achieved, the higher the reported time spent posting status updates. Furthermore, materialistic values (both extrinsic importance and extrinsic likelihood) were associated with less reported time spent chatting on Facebook messenger, and less reported time spent sharing links. Conversely, non-materialistic values (intrinsic importance and intrinsic likelihood) were associated with reporting more time spent sharing links, and more time chatting on Facebook messenger, but less time posting photos (for intrinsic importance) and less time posting status updates (for intrinsic likelihood). Future research could investigate the robustness of these associations, as well as further investigating the underlying mechanisms between the association between materialistic and non-materialistic values and the differing usage of Facebook functions.

## References

1. Abbruzzese, J.: Facebook hits 2.27 billion monthly active users as earnings stabilize (2018). <https://www.nbcnews.com/tech/tech-news/facebook-hits-2-27-billion-monthly-active-users-earnings-stabilize-n926391>
2. Pittman, M., Reich, B.: Social media and loneliness: why an Instagram picture may be worth more than a thousand Twitter words. *Comput. Hum. Behav.* **62**, 155–167 (2016). <https://doi.org/10.1016/j.chb.2016.03.084>
3. Belk, R.W.: Three scales to measure constructs related to materialism: reliability, validity, and relationships to other measures of happiness. In: Kinnear, T. (ed.) *Advances in Consumer Research*, vol. 11, pp. 291–297. Association for Consumer Research, Provo (1984)
4. Richins, M.L., Dawson, S.: A consumer values orientation for materialism and its measurement: scale development and validation. *J. Consum. Res.* **19**, 303–316 (1992). <https://doi.org/10.1086/209304>
5. Manchiraju, S.: Materialism in consumer behavior: a Review. *Manag. Market. Challenges Knowl. Soc.* **8**(2), 329–352 (2013)
6. Shrum, L.J., et al.: Reconceptualizing materialism as identity goal pursuits: functions, processes, and consequences. *J. Bus. Res.* **66**(8), 1179–1185 (2013). <https://doi.org/10.1016/j.jbusres.2012.08.010>
7. Kasser, T., Ryan, R.M.: Further examining the American dream: differential correlates of intrinsic and extrinsic goals. *Pers. Soc. Psychol. Bull.* **22**, 280–287 (1996). <https://doi.org/10.1177/0146167296223006>
8. Rai, R., Chauhan, C., Cheng, M.: Materialistic values, brand knowledge and the mass media: Hours spent on the Internet predicts materialistic values and brand knowledge. *Current Psychol.* (2018). <https://doi.org/10.1007/s12144-018-9900-0>
9. Chu, S.C., Windels, K., Kamal, S.: The influence of self-construal and materialism on social media intensity: a study of China and the United States. *J. Adv.* **35**(3), 569–588 (2016). <https://doi.org/10.1080/02650487.2015.1068425>
10. Ozimek, P., Baer, F., Förster, J.: Materialists on Facebook: the self-regulatory role of social comparisons and the objectification of Facebook friends. *Heliyon* **3**(11), e00449 (2017). <https://doi.org/10.1016/j.heliyon.2017.e00449>

11. Gerbner, G.: On content analysis and critical research in mass communication. *AV Commun. Rev.* **6**, 85–108 (1958)
12. Gerbner, G.: Cultivation analysis: an overview. *Mass Commun. Soc.* **1**(3/4), 175–194 (1998)
13. Lee, D., Hosanagar, K., Nair, H.S.: Advertising content and consumer engagement on social media: evidence from Facebook. *Manage. Sci.* **64**(11), 5105–5131 (2018). <https://doi.org/10.1287/mnsc.2017.2902>
14. Alhabash, S., Mundel, J., Hussain, S.A.: Social media advertising. In: Rodgers, S., Thorson, E. (eds.) *Digital Advertising: Theory and Research*, pp. 285–299. Routledge, New York (2017)
15. Kirtişa, A.K., Karahanbb, F.: To be or not to be in social media arena as the most cost-efficient marketing strategy after the global recession. *Procedia Soc. Behav. Sci.* **24**, 260–268 (2011). <https://doi.org/10.1016/j.sbspro.2011.09.083>
16. Yazdanparast, A., Joseph, M., Muniz, F.: Consumer based brand equity in the 21st century: an examination of the role of social media marketing. *Young Consum.* **17**(3), 243–255 (2016). <https://doi.org/10.1108/YC-03-2016-00590>
17. Chu, S.C., Kamal, S.: An investigation of social media usage, brand consciousness, and purchase intention towards luxury products among millennials. In: Okazaki, S. (ed.) *Advances in Advertising Research*, vol. 2. Gabler, Wiesbaden (2011)
18. Chu, S.C., Kamal, S., Kim, Y.: Understanding consumers' responses toward social media advertising and purchase intention toward luxury products. *J. Glob. Fashion Market.* **4**(3), 158–174 (2013)
19. Duan, J., Dholakia, N.: The reshaping of Chinese consumer values in the social media era: exploring the impact of Weibo. *Qual. Market Res. Int. J.* **18**(4), 409–426 (2015). <https://doi.org/10.1108/QMR-07-2014-0058>
20. Baek, K., Holton, A., Harp, D., Yaschur, C.: The links that bind: uncovering novel motivations for linking on Facebook. *Comput. Hum. Behav.* **27**(6), 2243–2248 (2011). <https://doi.org/10.1016/j.chb.2011.07.003>
21. Junco, R.: The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement. *Comput. Educ.* **58**(1), 162–171 (2012). <https://doi.org/10.1016/j.compedu.2011.08.004>





# Analyzing #LasTesis Feminist Movement in Twitter Using Topic Models

Sebastian Rodriguez, Héctor Allende-Cid<sup>(✉)</sup>, Cristian Gonzalez,  
Rodrigo Alfaro, Claudio Elortegui, Wenceslao Palma,  
and Pedro Santander

Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile  
{sebastian.rodriguez.o, hector.allende,  
cristian.gonzalez, rodrigo.alfaro, claudio.elortegui, wenceslao.palma,  
pedro.santander}@pucv.cl

**Abstract.** Nowadays, social networks have created a massive mean of communication, that was unthinkable many years ago. Informal communication, blogging, and online discussions have transformed the Web into a huge repository of remarks on numerous themes, producing a potential wellspring of data for various areas. In this paper we analyze, using Topic Models, a recent widespread feminist movement. Las Tesis is a feminist collective that initiated a protest against sexual abuse, and that was replicated in more than dozen different countries in matter of days. We use LDA and BTM to detect automatically the topics in over 627643 tweets that were gathered from the 25th November until the 5th January. The resulting topics obtained, from tweets in Spanish and English, show that these algorithms are able to capture the real-world events that occurred in Chile and Turkey.

**Keywords:** Topic Models · Twitter · LDA · BTM

## 1 Introduction

Nowadays, social networks have created a mean of communication, that was unprecedented years ago. Informal communication, blogging, and online discussions have transformed the Web into a huge archive of remarks on numerous themes, producing a potential wellspring of data for various areas. The accessibility of large-scale electronic social information from the Web and other electronic means is as of now changing how people nowadays communicate [14]. The social networks are also being used for other objectives, for example, consequently separating client opinions about products or brands [15], nowcasting earthquakes [12] and detecting suicidality [13].

Twitter, one of the most used Social Networks, can be depicted as a informal community website that captures messages of 280 characters. This micro-blogging service, provides users with a framework for writing brief, often-noisy

postings about different subjects. These posts are called “Tweets”. It is for blogging on the grounds that the focal action is posting short announcement messages (tweets) by means of the Web or handheld device. Twitter is additionally an interpersonal organization site since individuals have a profile page and those individuals can be associated with different individuals by “following” them. A common element of Twitter is retweeting: sending a tweet by posting it once more. The reposting of the equivalent (or comparative) data works since individuals will in general follow various arrangements of individuals, in spite of the fact that retweeting likewise fills different needs. For example, helping supporters to discover more established posts. Another element of Twitter (and other social networks) is the hashtag: a metatag starting with # that is intended to help other people discover a post, regularly by denoting the Tweet theme or its target group. This component appears to have been created by Twitter clients, in mid 2008 [8]. The utilization of hashtags stresses the significance of generally conveying data in Twitter. Conversely, the character is utilized to deliver a post to another enrolled Twitter client, permitting Twitter to be used successfully for discussions and coordinated effort.

In order to analyze and extract semantic information about this huge amount of data generated from this microblogging platform, automatic methods are necessary. In this sense, Topic Models are a very useful tool for this purpose. Topic models are statistically inspired and unravel the hidden structure in large collections of texts.

In this paper we analyze the social impact of the performance “A rapist in your path” (Un violador en tu camino) proposed by the feminist collective Las Tesis. Although the performance started in several cities in Chile, this performance has been also replicated in different cities around the world. Some of this cities were Paris, London, Barcelona, New York, Mexico City, Istanbul, Madrid, Berlin and Bogotá. This street art intervention greatly exceeded national borders and has brought together hundreds of women around the world, who have organized to replicate the choreography and song created by four women from Valparaíso, Daffne Valdés Vargas, Sibila Sotomayor Van Rysseghem, Paula Cometa Stange and Lea Cáceres Díaz.

The paper is organized as follows: In Sect. 2 we briefly describe Topic Models. In Sect. 3, we perform a descriptive analysis and apply two Topic Models to the data, namely LDA and BPM. In Sect. 4 we describe the results and in the last section we conclude and delineate future work.

## 2 Topic Models

Topic Models, in a very concise way, are a specific type of statistical language models used for unveiling hidden structure in large collections of texts. Intuitively, we can think of it in different aspects:

- Dimensionality Reduction, where rather than representing a text  $T$  in its feature space, you can represent it in a topic space.

- Unsupervised Learning, where it can be compared to clustering. The number of topics, like the number of clusters, is an output parameter. By doing topic modeling, we build clusters of words rather than clusters of texts. A text is thus a mixture of all the topics, each having a specific weight.
- Tagging, abstract “topics” that occur in a collection of documents that best represents the information in them.

There are several existing algorithms you can use to perform the topic modeling. The most common of it are, Latent Semantic Analysis (LSA/LSI) [4], Probabilistic Latent Semantic Analysis (pLSA) [7], and Latent Dirichlet Allocation (LDA) [2]. Topic modeling is the task of identifying topics automatically in a set of documents. This can be very useful for customer service automation, search engines and any other case where knowing the topics of documents is important. LDA [2] is a form of unsupervised learning that views documents as bags of words (where order does not matter). LDA works by first making a crucial assumption: the way a document was generated was by selecting a set of topics and then for each topic selecting a set of words. In order to do this it does the following for each document  $m$ :

- Assume there are  $k$  topics across all of the documents.
- Distribute these  $k$  topics across document  $m$  (this distribution is known as  $\alpha$  and can be symmetric or asymmetric) by assigning each word a topic.
- For each word  $w$  in document  $m$ , assume its topic is wrong but every other word is assigned the correct topic.
- Probabilistically assign word  $w$  a topic based on two things:
  - what topics are in document  $m$
  - how many times word  $w$  has been assigned a particular topic across all of the documents (this distribution is called  $\beta$ )
- Repeat this process a number of times for each document.

There have been several works on Topic Models applied to Twitter. LDA has been extended in several ways, and in particular for social networks and social media, a number of extensions to LDA have been proposed. For example, in [3] the authors proposed a novel probabilistic topic model to analyze text corpora and infer descriptions of the entities and of relationships between those entities on Wikipedia. The authors in [11] proposed a model to simultaneously discover groups among the entities and topics among the corresponding text. In [18] a model was introduced to incorporate LDA into a community detection process. In [10] and [17] we can find related work.

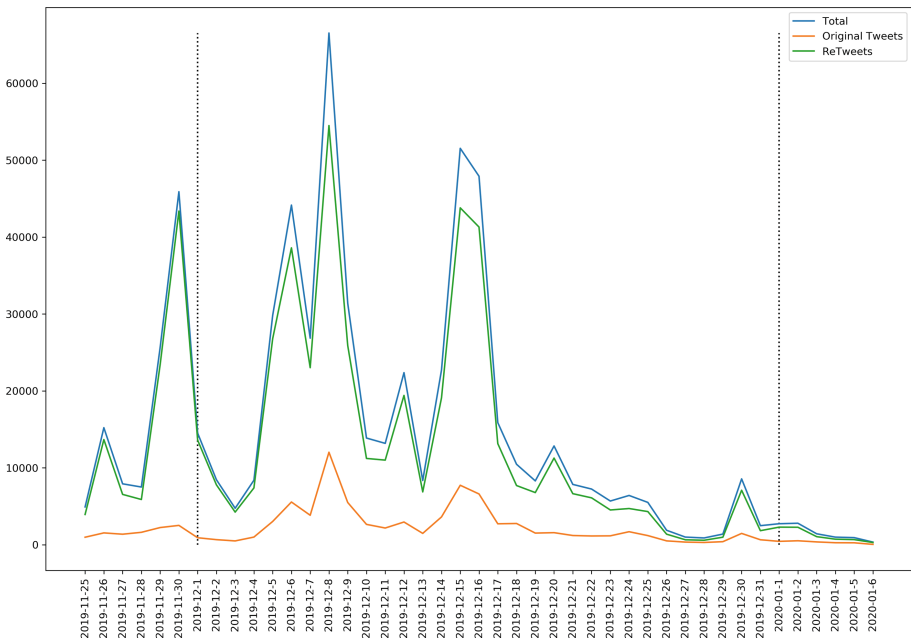
Uncovering the topics within short texts, such as tweets and instant messages, has become an important task for many content analysis applications. However, directly applying conventional topic models (e.g. LDA and PLSA) on such short texts may not work well. The main reason lies in that traditional topic models implicitly capture the document-level word co-occurrence patterns to reveal topics, and thus suffer from the severe data sparsity in short documents. In [16], the authors propose a novel way for modeling topics in short texts, referred as biterm topic model (BTM). Specifically, in BTM the topics are

learnt by directly modeling the generation of word co-occurrence patterns (i.e. bigrams) in the whole corpus. The major advantages of BTM are that 1) BTM explicitly models the word co-occurrence patterns to enhance the topic learning; and 2) BTM uses the aggregated patterns in the whole corpus for learning topics to solve the problem of sparse word co-occurrence patterns at document-level. The authors carry out extensive experiments on real-world short text collections. The results demonstrate that their approach can discover more prominent and coherent topics, and significantly outperform baseline methods on several evaluation metrics. Furthermore, they find that BTM can outperform LDA even on normal texts, showing the potential generality and wider usage of this new topic model.

### 3 Analysis

#### 3.1 Descriptive Analysis

The data used in this study was collected from the micro blogging platform Twitter. Several hashtags related to the event were used in order to capture 627643 tweets between the 25th November 2019 and the 5th January 2020. This sample was obtained with the paid Twitter API, so we got the entire number of tweets that were shared in those dates. In November the total number of tweets



**Fig. 1.** Number of tweets mentioning “#LasTesis” (and related words) from November 25, 2019 to January 5, 2020

were 111371 and the number of unique users was 54465. In December the number of tweets was 507193 with a total of 167464 users. In January we obtained 9079 from 7264 users. The total of unique users were 202797.

In Fig. 1, we can see the time series of the original messages and retweets. The time series has several peaks, achieving the maximum around the 8th of December. The highest peak in November is due to the replication of the performance of the feminist group in several cities in Chile. The highest peak in December was produced after the performance of the song in Turkey, were several woman were arrested by the police, due to the ‘crude’ language of the song. After that, there was a peak in the 16th of December, when women politicians of Turkey replicated the performance in the parliament. All the peaks are reflecting some activities of the real world, and we can see the backlash of this in this social network.

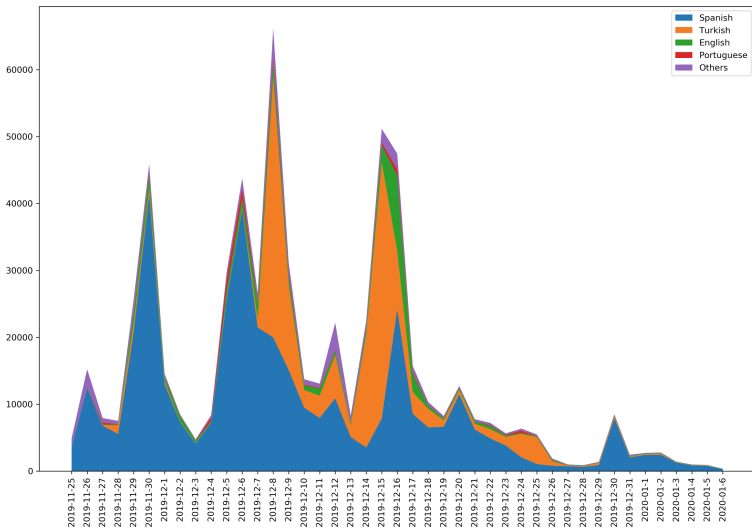


Fig. 2. Language distribution for the tweets from November 25, 2019 to January 5, 2020

In Fig. 2 we see the number of tweets and the language distribution. The majority of the tweets were written in Spanish, Turkish, English and Portuguese. We detected a total of 32 languages in the total tweets. Before the 7th of December the predominant language was Spanish, but after the performance in Istanbul, and the consequent violence from the police to the manifesters, the predominant language was Turkish. In Fig. 3 we can see the normalized graph where we can see that Spanish and Turkish were the most common languages.

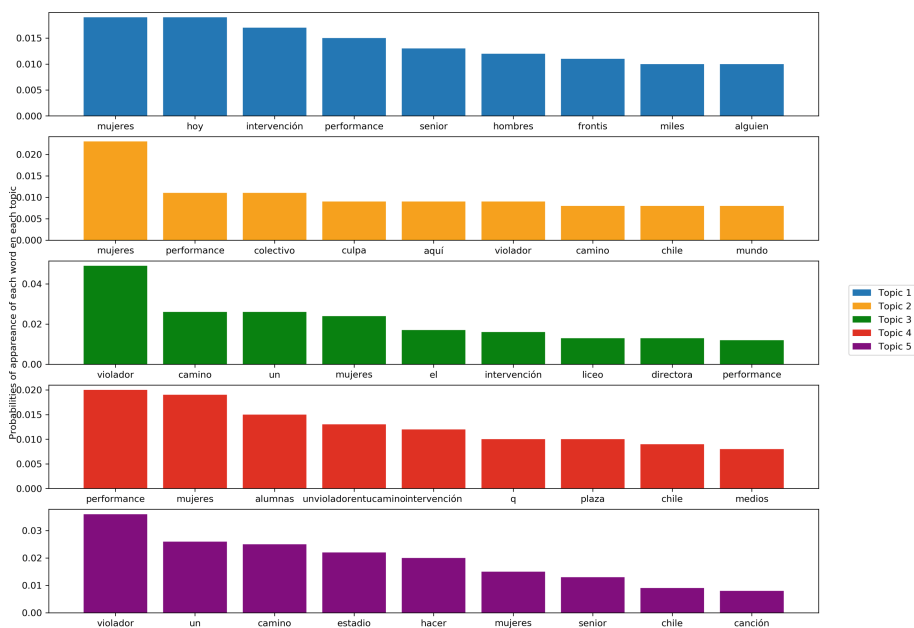
In Fig. 4 we observe a Word Cloud of the entire dataset. The most common words were “violador” (rapist), “mujeres” (women), “camino” (path), “Chile” and “kadnlar” (from kadınların, that means women in Turkish). In Fig. 5 we





different words that each of the results gave. With both methods there are some noticeable differences.

**Spanish.** In Figs. 7 and 8 we see the results of applying both algorithms, LDA and BTM, in the Spanish tweets corpus. Since BTM is more suited for short messages, we observe that the topics obtained with the latter algorithm are more related with real world events. The events are related with the Chilean context, since they refer to performances made in schools (“liceo”) and the one made by older women in front of Estadio Nacional, Santiago, Chile. In one topic there is also a reference on how this performance was replicated in other parts of the world.



**Fig. 7.** Five topics obtained with LDA for Spanish Tweets. The words are ordered by their probability of appearance in a given topic.

**English.** In Figs. 9 and 10 we see the results obtained with both algorithms in the English twitter corpus. The topics produced refer to the context in Turkey (results obtained from LDA and BTM) and France (result obtained from BTM). The events that are discussed in the twitter corpus mainly refer to the violent repression of the performance in the streets of Istanbul and after, the performance made in the parliament by women politicians.



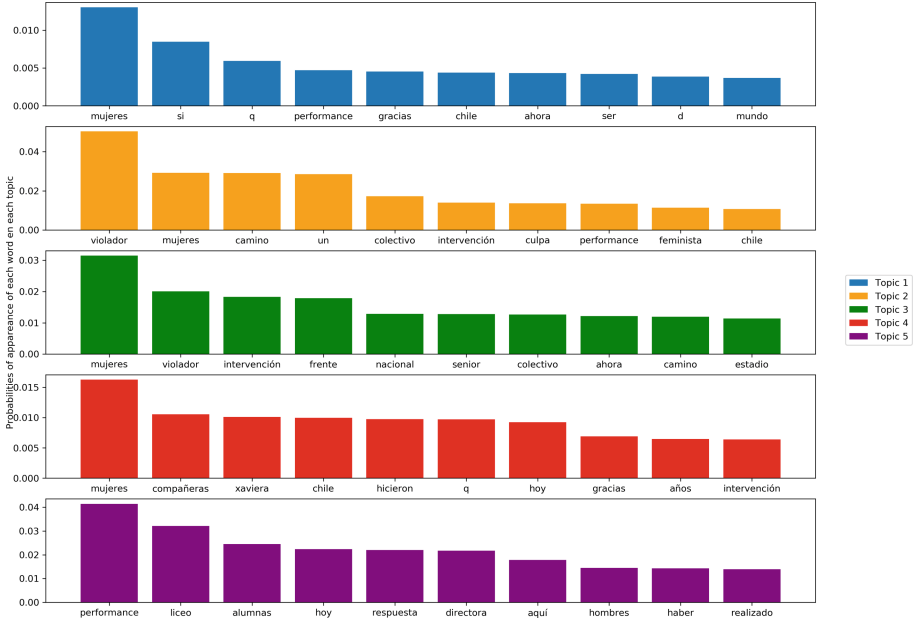
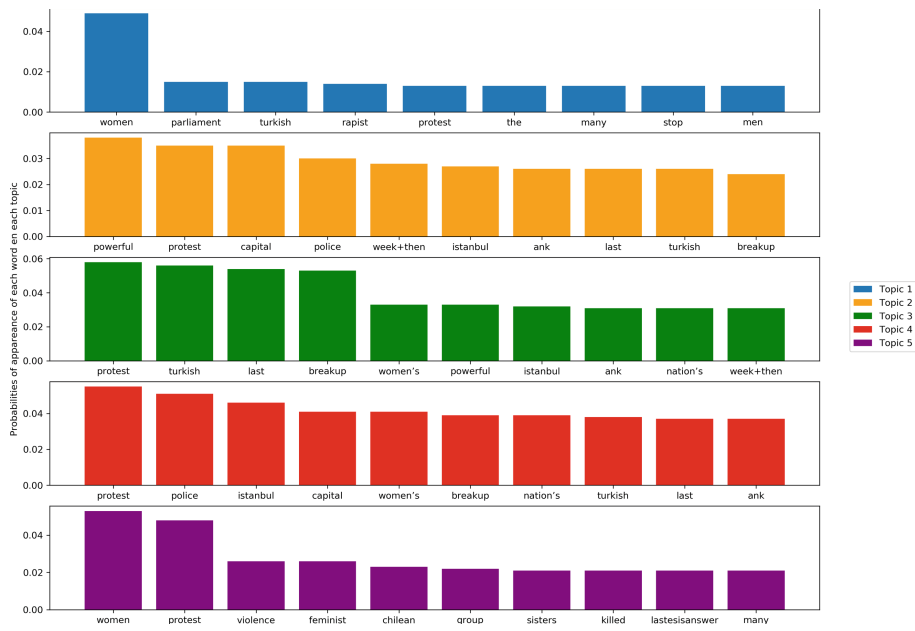


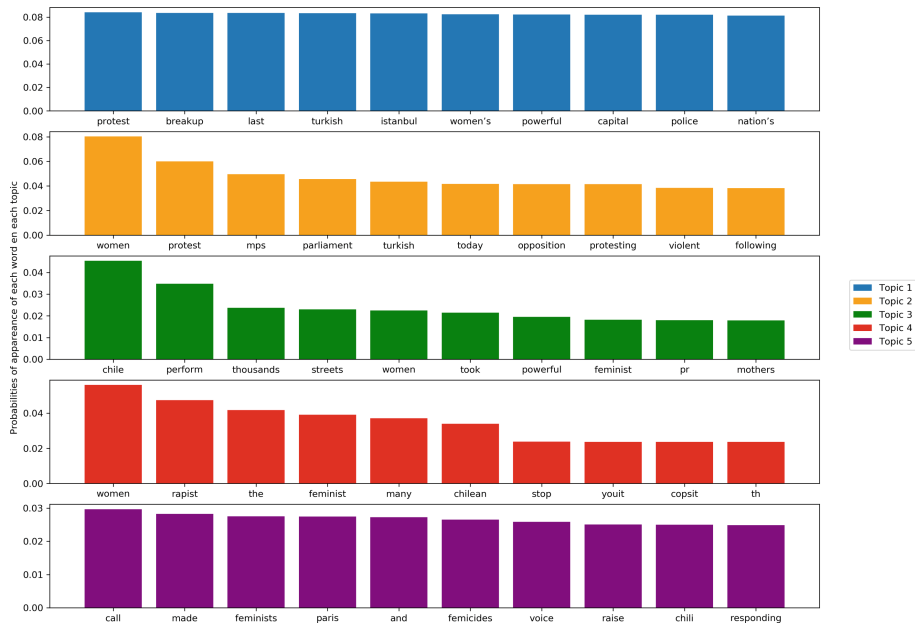
Fig. 8. Five topics obtained with Bi-term Topic Model for Spanish Tweets. The words are ordered by their probability of appearance in a given topic.

## 4 Discussion

As can be seen in the previous section, we performed several analysis to the collected data. As can be seen in the descriptive analysis, the phenomenon was shared and retweeted thousand of times, showing us that the phenomenon became widespread in matter of days. The total amount of languages that we found in the Twitter corpus show that the performance was also replicated in several countries and cities. Also, it is noticeable that the majority of the things that were said about the movement were mainly positive (92%). In relation to the results obtained by both of the Topic Models, we observed that both of these algorithms were able to capture the real-world events that occurred in different parts of the world. In the Spanish corpus, we obtained as a result the events that occurred in Chile during the first week after the spring of the movement (Performance in schools and in Estadio Nacional), while in the corpus in English, we obtained as a result the events that occurred in Turkey, both in the streets of Istanbul and in the parliament.



**Fig. 9.** Five topics obtained with LDA for English Tweets. The words are ordered by their probability of appearance in a given topic.



**Fig. 10.** Five topics obtained with Bi-term Topic Model for English Tweets. The words are ordered by their probability of appearance in a given topic.

## 5 Conclusions

In this work we analyzed over half a million tweets written in various languages. It shows the widespread phenomenon of the performance made by the feminist collective Las Tesis. It shows, how this performance affected and influenced many feminist organizations in the world. The performance was replicated over 10 countries, and the song was translated in many languages. In order to analyze the discussion that this performance engaged in all the world we used to algorithms to create automatically different topics. We used LDA and BTM, in both Spanish and English, to establish what the users in Twitter were speaking about. We see that BTM creates more cohesive topics, since BTM has been shown to work better in shorter texts. As future work, we pretend to work together with Sentiment Analysis to create topics for positive and negative tweets. We also will work on Machine Learning models in order to automatically classify those tweets according to their sentiment, thus not relying on sentiment dictionaries.

## References

1. Bansal, N., Koudas, N.: BlogScope: a system for online analysis of high volume text streams. In: Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 1410–1413. ACM Press, New York (2007)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003). <https://doi.org/10.1162/jmlr.2003.3.4-5.993>
3. Chang, J., Boyd-Graber, J., Blei, D.M.: Connections between the lines: augmenting social networks with text. In: KDD 2009: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 169–178 (2009)
4. Dumais, S.T.: Latent semantic analysis. *Ann. Rev. Inf. Sci. Technol.* **38**, 188–230 (2005). <https://doi.org/10.1002/aris.1440380105>
5. Esuli, A., Sebastiani, F.: SentiWordNet: a publicly available lexical resource for opinion mining. In: Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), Genoa, pp. 417–422 (2006)
6. Gruhl, D., Chavet, L., Gibson, D., Meyer, J., Pattanayak, P.: How to build a WebFountain: an architecture for very large-scale text analytics. *IBM Syst. J.* **43**(1), 64–77 (2004)
7. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999) (1999)
8. Huang, J., Thornton, K.M., Efthimiadis, E.N.: Conversational tagging in Twitter. In: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT 2010), pp. 173–178. Association for Computing Machinery, New York (2010). <https://doi.org/10.1145/1810617.1810647>
9. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: tweets as electronic word of mouth. *J. Am. Soc. Inform. Sci. Technol.* **60**(11), 2169–2188 (2009)
10. Liu, Y., Niculescu-Mizil, A., Gryc, W.: Topic-link LDA: joint models of topic and author community. In: ICML 2009: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 665–672. ACM (2009)

11. McCallum, A., Wang, X., Mohanty, N.: Joint group and topic discovery from relations and text. In: Airoldi, E., Blei, D.M., Fienberg, S.E., Goldenberg, A., Xing, E.P., Zheng, A.X. (eds.) ICML 2006. LNCS, vol. 4503, pp. 28–44. Springer, Heidelberg (2007). [https://doi.org/10.1007/978-3-540-73133-7\\_3](https://doi.org/10.1007/978-3-540-73133-7_3)
12. Mendoza, M., Poblete, B., Valderrama, I.: Nowcasting earthquake damages with Twitter. *EPJ Data Sci.* **8**(1), 1–23 (2019). <https://doi.org/10.1140/epjds/s13688-019-0181-0>
13. O’Dea, B., Wan, S., Batterham, P.J., Calear, A.L., Paris, C., Christensen, H.: Detecting suicidality on Twitter. *Internet Interv.* **2**(2), 183–188 (2015). ISSN 2214-7829. <https://doi.org/10.1016/j.invent.2015.03.005>
14. Savage, M., Burrows, R.: The coming crisis in empirical sociology. *Sociology* **41**(5), 885–899 (2007)
15. Voorveld, H.: Brand communication in social media: a research agenda. *J. Advert.*, 1–13 (2019). <https://doi.org/10.1080/00913367.2019.1588808>
16. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web, pp. 1445–1456 (2013). <https://doi.org/10.1145/2488388.2488514>
17. Yu, D., Xu, D., Wang, D., Ni, Z.: Hierarchical topic modeling of Twitter data for online analytical processing. *IEEE Access* **7**, 12373–12385 (2019). <https://doi.org/10.1109/ACCESS.2019.2891902>
18. Zhang, H., Giles, C.L., Foley, H.C., Yen, J.: Probabilistic community discovery using hierarchical latent Gaussian mixture model. In: AAAI 2007: Proceedings of the 22nd National Conference on Artificial Intelligence, pp. 663–668 (2007)



# User-Oriented Quality Estimation of Social News Systems and Its Content

## Gender-Dependent Assessment of Reddit

Katrin Scheibe<sup>(✉)</sup> and Franziska Zimmer

Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany  
{katrin.scheibe, franziska.zimmer}@hhu.de

**Abstract.** Reddit is a social news service and information platform where users can discuss different topics in sub-communities, so called “subreddits.” In this study, the perceived information system quality and the perceived content quality of the information service Reddit, the information seeking behavior as well as the motives of Reddit’s users following the Uses and Gratifications Theory are analyzed differentiated by the perception of male and female users. To this end, a survey with 495 Reddit users was conducted. Results show that users’ motives to apply Reddit are mostly entertainment as well as information. All users agree that Reddit is enjoyable, useful, and easy to use, whereby no major difference in the perception by male and female users can be observed. All in all, the content is perceived as up-to-date and can be easily read and easily understood or comprehended. Most users are browsing through Reddit to find information, whereas male users are using the advanced search option more often than female users.

**Keywords:** Reddit · Information system quality · Social news service · Content · Gender

## 1 Introduction

Information systems are developed and designed to enable their users to access the needed information, which also facilitates the process of information seeking [16]. Studying the user-oriented estimation of a system’s quality is therefore necessary to understand which expectations and needs are fulfilled by seeking information and the use of a service [15]. Moreover, it gives insights about improving the quality of information services as well as managing and designing them [1].

A widespread information system, also known as a social news aggregator, popular among adolescents and young adults, is Reddit (Fig. 1). It was launched in 2005 and reports over three billion page views per month [5]. Looking at Alexa’s [2] global ranking of all websites worldwide, it is on the 14<sup>th</sup> position and therefore the most popular social news service, next to e.g., Digg and HackerNews.

Following the definition by Weninger et al. [29:579], social news websites are services “[...] in which (1) users generate or submit links to content, (2) submissions are voted on and ranked according to their vote totals, (3) users comment on the submitted

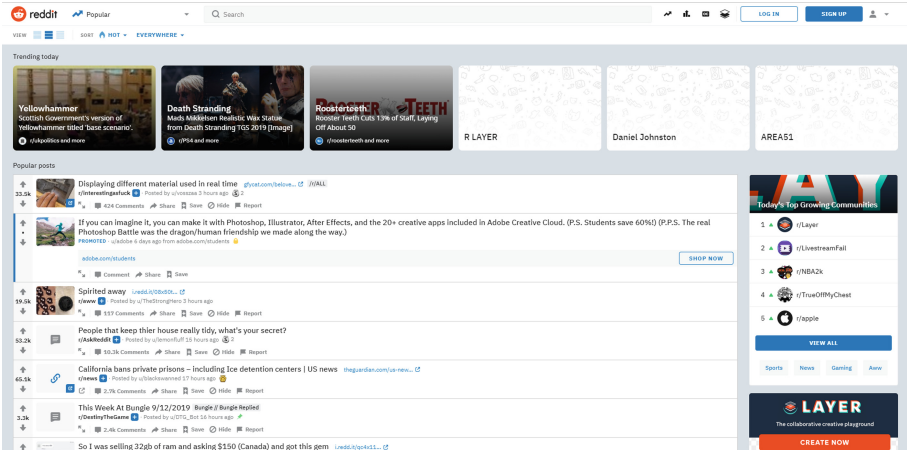


Fig. 1. The frontpage of Reddit.

content, and (4) comments are voted on and ranked according to their vote totals.” Also, users are interacting anonymously on Reddit and are able to post their own content in form of texts, images, or videos. But, the primary focus of Reddit is on its user-generated content and the information exchange of external sources [25]. It provides a platform for communication about internet-based information where many different topics are discussed in sub-communities, so called “subreddits” [26]. An example would be the subreddit *r/funny*, where users post humorous and fun content, another subreddit is *r/worldnews* where current events and news are shared.

Zimmer et al. [31] investigated the information service quality and the content quality as well as information service acceptance by the users of Reddit following the Information Service Evaluation Model. The Information Service Evaluation (ISE) Model by Schuman and Stock [22] displays a comprehensive research framework that unifies e.g., different models and techniques to investigate the quality of an information system and thereby the (information) behavior of its users. It also considers the aspect of acceptance (e.g., adoption of the system), the environment (e.g., information marketing and similar services) as well as time (development of the system over time). Information systems are made to satisfy human information needs. While applying a service, people are expressing information behavior when following their information needs. Wilson [30:49] defines information behavior as “[...] the totality of human behavior in relation to sources and channels of information, including both active and passive information seeking, and information use.” According to Schuman and Stock [22:2] the concept of information behavior includes “the behavior [of information production (e.g., user-generated content in social media) and the behavior of information seeking (e.g., browsing through web sites or applying search engines).”

In this evaluation of Reddit as an information system the focus will be set on the user perception of the system’s quality and the system’s content, as Kusunoki and Sarcevic [16:860] outline the importance of the user’s perspective. Therefore, we limit facets of the ISE model to the perceived information system quality following the Technology

Acceptance Model (TAM) [7], the perceived content quality based on different aspects defined by Parker, Moleshe et al. [20] as well as the information seeking behavior of users. Furthermore, it is necessary to study what motivates users of Reddit, and especially different genders, to apply this particular social news and information service, as Bogers and Wernersen [4] found that the social aspect of Reddit is not important, but the informational value of the service is.

In 1974, the researchers Katz et al. [13] outlined findings about uses and gratifications research, which resulted in the Uses and Gratifications Theory (U&GT). It is a popular theory in media and communication studies to explain why people are using certain media. According to the U&GT, media consumption is goal-directed and should result in the satisfaction of a person's needs. The audience, or the users, are searching for gratifications while being exposed to media. It is always guided by expectations and depends on a person's social and psychological background and which media is chosen. The audience decides actively whether to apply a service or not [3]. A total of 35 different needs for media consumption were identified by Katz et al. [14]. Whereof later four central motives were summarized by McQuail [18], which are information, entertainment, social interaction, and self-actualization. In line with Shao [23] one can also speak of self-presentation (with regard to self-actualization) for the activity of producing content on a social media service.

According to different research articles [4, 8, 9], Reddit is applied by more male than female users. To some extent men and women use social media and the internet in general for different purposes, e.g. men might use it more for gaming and entertainment, whereas women for communicating and connecting with people [12]. In this research article, the perceived information system quality and the perceived content quality of the information service Reddit, the information seeking behavior of Reddit's users and the motives of Reddit's users following the Uses and Gratifications Theory are analyzed with particular attention to the different opinions of female and male users. The principal question here is whether the perceived service and content quality or the motivation to use Reddit differ between genders and if those are the reasons why Reddit is applied more by male users. Furthermore, gender research on Reddit is limited and our findings may serve as recording for ongoing gender studies. Based on these considerations this study aims at answering the following research questions (RQs):

**RQ1:** What are the motives of male and female users to use Reddit?

**RQ2:** How do male and female users rate the information service quality of Reddit?

**RQ3:** How do male and female users rate the content quality of Reddit?

**RQ4:** How do male and female users seek for information on Reddit?

## 2 Related Work

When looking at the aspect why different genders use or social networking sites (SNSs), a few differences can be observed. Overall, men seem to use social networking sites to form new relationships, women use them to help keep existing ones [19]. A study determined that women are more likely to apply SNSs to compare themselves with other users and to search for information. In contrast, men seem to look at profiles of others

in order to find friends [11]. When it comes to the production of content, female users tend to share personal issues (for example family matters) whereas men like to discuss public events like politics and sports [28], or technology and money [21]. In this context, Reddit should be named, as it “essentially started out as very techy-and nerd-oriented” [24:6], which could explain the majority of the users being male. Nonetheless, Reddit is since being enjoyed by both genders. Even though Reddit is defined as being a SNSs, only few people use it in this traditional sense. Reddit is almost never applied to build or sustain long-term relationships [4, 24]. If the users want to stay in contact, they usually shift the conversations to Facebook or other SNSs. As Reddit is a relatively anonymous SNS, the users are mindful in how they present themselves if their real identity could potentially be traced back. This form of anonymity also potentially gives a platform to the culture of careless words [24]. But, Reddit is valued for the information as well as the quality of it. Users also like the possibility to customize Reddit. They are able to actively shape the placement and reception of posts in their favorite subreddits of interest by comments and votes [4]. A number of studies examined the content of Reddit. For example, Stoddard [26] found that higher quality articles seem to be the most popular forms of content. To determine which users post such content, a study observed that the users, regardless of whether they are experienced or inexperienced with various levels of reputation, tend to post any kind of content, being it professional articles or conversational posts [17]. In this context, it was also observed that the earlier a post is voted on, the more likely its popularity will be affected [26]. This phenomenon also extends to the top comments – the early comments receive the most replies [29]. Also, the number of downvotes increases faster than of upvotes [27]. Furthermore, half of the valuable content on Reddit seems to be ignored on the first submission. This potential thread could be solved by a combination of social norms, repeated interaction, and reputation mechanisms [10]. Even though Reddit is seen favorably for its content, the design of the website is perceived rather negatively, as the interface, navigation, user hostile search function as well as the search results are not seen as being positive. Nonetheless, a bonus factor is the friendly community which is highly valued by Reddit’s user base [31].

To sum up, gender research on social media is an emerging topic, but to date, there is no study that examined the perception and use of Reddit by male and female users. This study should serve as a first contribution to this research field.

### 3 Methods

To answer the research questions (RQ1–4), an online questionnaire was developed. The survey was constructed on Umfrageonline.com and took place between May 29, 2017 and July 7, 2017. It was shared on different social media platforms like Facebook survey groups and on different subreddits. The survey was answered by all participants on a voluntary basis with no compensation. All participants had to state their Reddit usage status (‘I use Reddit’, ‘I do not use Reddit anymore’, ‘I never used Reddit’.) Overall, the survey was answered by 672 participants, of which 599 are active Reddit users, 58 never used the service and 15 are not using it anymore. Only the answers given by active users, meaning they visit the site regularly, were used for this investigation. 495 of those active users completed the survey.



At the end of the survey, the attendees were asked about demographic aspects (age, gender, country of origin, highest educational level.) The majority of the questions contained pre-formulated answers, for example regarding the question, “how do you search on Reddit?.” The answers given were “only by browsing,” “via search query box,” and “using advanced search.”

To answer RQ1, questions modeled after the Uses and Gratifications Theory according to Katz et al. [13] were used. The participants could select via a multiple choice question the four dimensions: entertainment, information, socializing, and self-presentation.

In line with the Technology Acceptance Model (TAM) proposed by Davis [7], the second research question (RQ2) on how the different genders perceive the information service quality of Reddit can be answered. The aspects that were asked about included: how enjoyable [6], useful, trustable [7], and easy to use Reddit is regarded as. Here, the participants could rate each aspect on a five point Likert scale (1 meaning “strongly disagree” to 5 meaning “strongly agree”).

To answer RQ3, how the different genders perceive the content on Reddit, again, a five point Likert scale (1 meaning “strongly disagree” to 5 meaning “strongly agree”) was used. The content could be rated by each category: it is up-to-date; true; credible; unbiased, unprejudiced and impartial; can be easily read; has a formal structure; can be easily understood or comprehended. The categories for this were derived from Parker et al. [20]. As the quality of content is hard to quantify, users should be asked about aspects such as freshness of content, its believability, objectivity, readability, or understandability.

For research question four (RQ4), how female and male users search on Reddit for information, a multiple choice question was modeled. As Reddit offers the users the possibility to utilize advanced search options, the participants could select the answers “only by browsing (clicking through subreddits)”, “via a search query box”, and “using the advanced search.”

As RQ1, RQ2, RQ3 and RQ4 are answered by one survey question each, Cronbach’s Alpha was not calculated for validity of the survey. The data analysis was conducted with IBM SPSS 25. To answer the stated research questions, several statistical tests were applied. For general overview of the sample, descriptive statistics, the Pearson  $\chi^2$  were calculated. In order to estimate whether there are statistically significant differences between male and female users, the non-parametric Mann Whitney U test was conducted (since the answers marked on the Likert scales were handled as ordinal data).

## 4 Results

A total of 495 Reddit users participated in the survey, whereof 59.80% are male and 40.20% are female participants. This quite balanced distribution in our sample gives us a good basis for calculating gender-dependent differences. Overall, the median age of the participants is 23. For female participants, the median age is 23 and the mean age 24.93. Whereof for male participants, the median age is 22 and the mean age 23.43. The female participants were slightly older. Furthermore, most of the participants (around 40%) are from the United States of America.

**Table 1.** Motives of users to apply Reddit differentiated by gender.

| Motives           | All users<br>(N = 495) | Male users<br>(N = 296) | Female users<br>(N = 199) | Sig. |
|-------------------|------------------------|-------------------------|---------------------------|------|
| Entertainment     | 96.00%                 | 95.90%                  | 96.00%                    | .985 |
| Information       | 86.70%                 | 87.50%                  | 85.40%                    | .506 |
| Socializing       | 21.00%                 | 20.30%                  | 22.10%                    | .022 |
| Self-Presentation | 5.10%                  | 6.10%                   | 3.50%                     | .057 |

Answering the first research question (RQ1), what motivates users of Reddit to apply the information system (differentiated by gender), 96.00% are using it for entertainment purposes (Table 1). There is nearly no difference between male (95.90%) and female (96.00%) users regarding this aspect. 86.70% of all participants agreed that their motivation to apply Reddit is to get information. Here, male users (87.50%) are a little bit more into getting information on Reddit than female users (85.40%). Exactly 21.00% use Reddit for socializing and getting in contact with other people. More female (22.10%) than male (20.30%) participants named socializing as their motive. And, only 5.10% named self-presentation as to why they apply Reddit (6.10% male users and 3.50% female users). Therefore, male users are a little bit more into getting informed on Reddit as well as to present themselves and a few more female users stated that they are motivated to use Reddit for socializing. The main reason to use Reddit is the aspect of entertainment followed by information.

**Table 2.** How different genders perceive the service quality of Reddit.

| Service quality | All users      |     | Male users     |     | Female users   |     | Sig. |
|-----------------|----------------|-----|----------------|-----|----------------|-----|------|
|                 | Median         | IQR | Median         | IQR | Median         | IQR |      |
| Enjoyable       | 4.00 (N = 494) | 1   | 5.00 (N = 295) | 1   | 4.00 (N = 199) | 1   | .778 |
| Useful          | 4.00 (N = 492) | 1   | 4.00 (N = 293) | 1   | 4.00 (N = 199) | 2   | .206 |
| Trustable       | 3.00 (N = 487) | 2   | 3.00 (N = 292) | 2   | 3.00 (N = 195) | 0   | .832 |
| Easy to use     | 4.00 (N = 495) | 2   | 4.00 (N = 296) | 2   | 4.00 (N = 199) | 2   | .802 |

For the perceived service quality differentiated by gender (RQ2), participants should rate statements about Reddit being enjoyable, useful, trustable, or easy to use (Table 2). For the statement that Reddit is enjoyable, looking at male users (median: 5.00) and female users (median: 4.00) they both agreed, while the male ones rated it slightly better (median: +1.00). Therefore, most of the male respondents strongly agree that Reddit is an enjoyable service. Looking at the usefulness of Reddit, male (median: 4.00) and female (median: 4.00) users both agreed on this aspect, while male users rated it slightly better, as the interquartile range (IQR) for female users is 2 and for male users 1.

**Table 3.** How different genders perceive the content quality of Reddit.

| Service quality                             | All users      |     | Male users     |     | Female users   |     | Sig. |
|---|----------------|-----|----------------|-----|----------------|-----|------|
|   | Median         | IQR | Median         | IQR | Median         | IQR |      |
| Up-to-date                                  | 4.00 (N = 486) | 1   | 4.00 (N = 291) | 1   | 4.00 (N = 195) | 1   | .747 |
| True  | 3.00 (N = 479) | 1   | 3.00 (N = 288) | 1   | 3.00 (N = 191) | 1   | .179 |
| Credible                                    | 3.00 (N = 484) | 1   | 3.00 (N = 290) | 1   | 3.00 (N = 194) | 1   | .526 |
| Unbiased,<br>unprejudiced,<br>and impartial | 2.00 (N = 488) | 1   | 2.00 (N = 293) | 2   | 2.00 (N = 195) | 1   | .680 |
| Easily read                                 | 4.00 (N = 494) | 1   | 4.00 (N = 295) | 2   | 4.00 (N = 199) | 1   | .114 |
| Has formal<br>structure                     | 3.00 (N = 470) | 2   | 3.00 (N = 287) | 1   | 3.00 (N = 192) | 2   | .095 |
| Easily<br>understood or<br>comprehended     | 4.00 (N = 491) | 1   | 4.00 (N = 292) | 1   | 4.00 (N = 199) | 2   | .156 |

The perception of Reddit as being trustable has a median of 3.00 (neutral) for both genders. Again, female and male users rate it mostly the same, whereas the IQR for female users is 0 and for male users is 2. According to all users, they agree (median: 4.00) on the statement that Reddit is an easy to use information system. Here, for both genders the median is 4 and the IQR is 2.

How do different genders perceive the content quality of Reddit is the third research question (RQ3) of this study. The results are shown in Table 3. Looking at the answer of all users, they agree (median: 4.00) that the content on Reddit is up-to-date. Female users (median: 4.00) as well as male users (median: 4.00) both rate the contents' freshness with a median of 4.00. Considering the statement that the content on Reddit is true, all users have a neutral point of view on this aspect (median: 3.00). Again, there is no difference between male users (median: 3.00; IQR: 1) and female users (median: 3.00; IQR: 1). The credibility factor (median: 3.00) was rated the same as the truth factor by all users. However, male users (median: 3.00; IQR: 1) perceive the credibility of the content exactly the same as female users (median: 3.00; IQR: 1). Further results show that all users (median: 2.00; IQR: 1) do not perceive the content of Reddit as unbiased, unprejudiced, and impartial. Here, both genders view it nearly the same, but male users (median: 2.00; IQR: 2) a little bit less negative than female users (median: 2.00; IQR: 1). For the next statement that the content on Reddit can be easily read, the users overall agree with a median of 4.00. Female users agree more (median: 4.00; IQR: 1) than male users (median: 4.00; IQR: 2). Users of Reddit have a neutral opinion on the formal structure of the content (median: 3.00; IQR: 2). For this, male users (median: 3.00; IQR: 1) agree more than female users (median: 3.00; IQR: 2), but it is only a minor difference. There is agreement with the statement that the content can be easily understood or comprehended (median: 4.00; IQR: 1). Male users agree a little bit more

with a median of 4.00 and an IQR of 1 whereas female users agree with a median of 4.00 and an IQR of 2.

All in all, the users agree that the content is up-to-date, can be easily read, and easily understood or comprehended. The statements that the content is true, credible, and has formal structure have been rated as neutral. Only disagreement was given for the statement that the content is unbiased, unprejudiced, and impartial.

**Table 4.** The information seeking behavior of different genders on Reddit.

| Seeking behavior | All users<br>(N = 495) | Male users<br>(N = 296) | Female users<br>(N = 199) | Sig. |
|------------------|------------------------|-------------------------|---------------------------|------|
| By browsing      | 73.70%                 | 73.00%                  | 74.90%                    | .637 |
| Search query box | 63.40%                 | 62.20%                  | 65.30%                    | .474 |
| Advanced search  | 26.10%                 | 28.40%                  | 22.60%                    | .152 |

Table 4 shows how users of Reddit are seeking for information on the service (RQ4). Most users are simply clicking through the web pages, posts, and subreddits on Reddit by browsing (73.70%), whereby 73.00% of the male users and 74.90% of the female users apply this method. More female users (65.30%) than male users (62.20%) are using the search query box for seeking information. Overall, 63.40% of the participants use it. The advanced search is used by more male users (28.40%) than female users (22.60%).

## 5 Discussion

By applying a survey with around 500 participants, we shed light on the gender-dependent differences in usage of and user's motives to apply Reddit. In addition, it was investigated how the service quality as well as the content quality is perceived. Another aspect of this research was the question about how the users apply the search functions of Reddit and if the genders prefer different functionalities.

If the motivational aspects according to the Uses and Gratifications Theory of this study are concerned, slight differences can be observed. In this study, male users apply Reddit more often than female users to find information. They also use the service more to present themselves than female users. Female users in contrast like to use Reddit to socialize with others. Indeed, more female users than male users are using Reddit for socializing as Joiner et al. [12] stated about general internet usage. Overall, the most important reason for all users is Reddit's entertainment factor as well as its informative content.

Taking a look at the perceived service quality and the aspects if Reddit is enjoyable, useful, trustable, and easy to use, both genders seem to agree on their perception of these dimensions. They perceive Reddit as being enjoyable. Male users rate Reddit as a little bit more useful than female users. Both genders seem to find the service easy to use. But, female as well as male users do not fully seem to trust Reddit.

Moving on to the perceived content quality of the posts on Reddit, both genders agree that the content is up-to-date, can be easily read and understood. If the truthfulness and

credibility of the content is concerned, male and female users rate those statements as neutral. The same applies to the content structure. One point stood out: both genders do not see the content as being unbiased, unprejudiced, and impartial.

Last but not least, the information seeking behavior was observed. Reddit offers its users advanced search options, which is only utilized by around 30% of the male users and 23% of the female users. Most of the systems' users like to only browse the web pages, posts, and subreddits. A few more female users than male users use the simple search query box.

Overall, Reddit is enjoyed by both genders. Its application does not seem to vary among the genders, as only a few differences could be observed. This research hopefully shed light on the usage of one of the internet's most favored websites and its utilization by men and women. It appears that once the service is being applied, there are only few significant gender-dependent differences.

When it comes to a general conclusion, users (it does not matter whether female or male) of the social news system Reddit seem to prefer the service because of the informative, but easy to read and entertaining content. Moreover, the simple and unpretentious design of Reddit (Fig. 1) makes it easy to use. Social news systems benefit from their user-generated content and user base.

Some limitations of this work have to be mentioned. First, the questionnaire was answered by 495 participants, which is a small fraction compared to Reddits popularity and its billion monthly visits. The results may display a larger difference in the perception of different genders if the sample was bigger. It could be possible to detect more gender-related insights by interviewing former users and non-users of this service (e.g., why are female internet users less interested in applying Reddit?).

Further research should focus on the aspect of anonymity on social networks. It is striking to see that male and female users seem to apply Reddit nearly the same way and have similar motives. The question arises if this is due to the nature of the service itself and its content, or if people tend to behave the same on social networks if they are nameless. As one Reddit user puts it "you don't have to worry about being tagged for who you are. It's more about what you say" [24:11]. Furthermore, it would be helpful and interesting to conduct interviews in order to collect qualitative data and describe more detailed results. It would also be interesting to study the service and content quality of similar information platforms, like the social news system HackerNews or Digg to compare the perceived quality of those services.

## References





1. Aladwani, A.M., Palvia, P.C.: Developing and validating an instrument for measuring user-perceived web quality. *Inf. Manag.* **39**(6), 467–476 (2002)
2. Alexa. <https://www.alexa.com/siteinfo/reddit.com>. Accessed 23 Aug 2019
3. Blumler, J.G., Katz, E.: *The Uses of Mass Communication*. Sage, Newbury Park (1974)
4. Bogers, T., Wernersen, R.: How "social" are social news sites? exploring the motivations for using reddit.com. In: *iConference 2014 Proceedings*, pp. 329–344. iSchools, Grandville (2014)
5. Carr, D.: Left alone by Its owner, reddit soars (2012). <https://www.nytimes.com/2012/09/03/business/media/reddit-thrives-after-advance-publications-let-it-sink-or-swim.html>

6. Chesney, T.: An acceptance model for useful and fun information systems. *Hum. Technol.* **2**(2), 225–235 (2006)
7. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **13**(3), 313–340 (1989)
8. Dou, W., Cho, I., ElTayeb, O., Choo, J., Wang, X., Ribarsky, W.: DemographicVis: analyzing demographic information based on user generated content. In: 2015 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 57–64. IEEE, New Jersey (2015)
9. Duggan, M., Smith, A.: 6% of online adults are reddit users. *Pew Internet Am. Life Proj.* **3**, 1–10 (2013)
10. Gilbert, E.: Widespread underprovision on reddit. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, pp. 803–808. ACM, New York (2013)
11. Haferkamp, N., Eimler, S.C., Papadakis, A.-M., Kruck, J.V.: Men are from mars, women are from venus? examining gender difference in self-presentation on social networking sites. *Cyberpsychol. Behav. Soc. Netw.* **15**(2), 91–98 (2011)
12. Joiner, R., et al.: Gender, internet identification, and internet anxiety: correlates of internet use. *CyberPsychol. Behav.* **8**(4), 371–378 (2005)
13. Katz, E., Blumler, J., Gurevitch, M.: Utilization of mass communication by the individual. In: Blumler, J., Katz, E. (eds.) *The Uses of Mass Communications: Current Perspectives on Gratifications Research*, pp. 19–32. Sage, Beverly Hills (1974)
14. Katz, E., Gurevitch, M., Haas, H.: On the use of the mass media for important things. *Am. Sociol. Rev.* **38**, 164–181 (1973)
15. Kettinger, W.J., Lee, C.C.: Perceived service quality and user satisfaction with the information services function. *Decis. Sci.* **25**(5/6), 737–766 (1994)
16. Kusunoki, D.S., Sarcevic, A.: A participatory evaluation design framework. In: iConference 2013 Proceedings, pp. 860–864. iSchools, Grandville (2013)
17. Leavitt, A., Clark, J.A.: Upvoting hurricane sandy: event-based news production processes on a social news site. In: Proceedings of the 32nd Annual ACM Conference on Human factors in computing systems, pp. 1495–1504. ACM, New York (2014)
18. McQuail, D.: *Mass Communication Theory*, 1st edn. Sage, London (1983)
19. Muscanell, N.L., Guadagno, R.E.: Make new friends or keep the old: gender and personality differences in social networking use. *Comput. Hum. Behav.* **28**(1), 107–112 (2012)
20. Parker, M.B., Moleshe, V., de la Harpe, R., Wills, G.B.: An evaluation of information quality frameworks for the world wide web. In: 8<sup>th</sup> Annual Conference on WWW Applications, pp. 1–11 (2006)
21. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 199–205. The AAI Press, Menlo Park (2006)
22. Schumann, L., Stock, W.G.: The information service evaluation (ISE) model. *Webology* **11**(1), 1–20 (2014)
23. Shao, G.: Understanding the appeal of user-generated media: a uses and gratification perspective. *Internet Res.* **19**(1), 7–25 (2009)
24. Shelton, M., Lo., K., Nardi, B.: Online media forums as separate social lives: a qualitative study of disclosure within and beyond reddit. In: iConference 2015 Proceedings, pp. 1–12. iSchools, Grandville (2015)
25. Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., Strohmaier, M.: Evolution of reddit: from the front page of the internet to a self-referential community? In: Proceedings of the 23rd International Conference on World Wide Web, pp. 517–522. ACM, New York (2014)
26. Stoddard, G.: Popularity and quality in social news aggregators: a study of reddit and hacker news. In: Proceedings of the 24<sup>th</sup> International Conference on World Wide Web, pp. 815–818. ACM, New York (2015)

27. Van Miegham, P.: Human psychology of common appraisal: the reddit score. *IEEE Trans. Multimedia* **13**(6), 1404–1406 (2011)
28. Wang, Y.-C., Burke, M., Kraut, R.: Gender, topic, and audience response: an analysis of user-generated content on Facebook. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 31–34. ACM, New York (2013)
29. Weninger, T., Zhu, X.A., Han, J.: An exploration of discussion threads in social news sites: a case study of the reddit community. In: *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 579–583. ACM, New York (2013)
30. Wilson, T.D.: Human information behavior. *Informing. Science* **3**(2), 49–56 (2000)
31. Zimmer, F., et al.: An evaluation of the social news aggregator reddit. In: *Proceedings of the 5<sup>th</sup> European Conference on Social Media*, pp. 364–373. Academic Conferences and Publishing, Reading (2018)



# Defining Network Borders on Instagram: The Case of Russian-Speaking Bloggers with Migration Background

Anna Smoliarova<sup>(✉)</sup> , Konstantin Platonov , Ekaterina Sharkova ,  
and Tamara Gromova 

St. Petersburg State University, Saint Petersburg, Russia  
a.smolyarova@spbu.ru

**Abstract.** Mass self-communication [12] has led to a continuous process of de-institutionalization of media systems; today, myriads of bloggers are struggling to receive audiences' attention. Networked user communities contribute to the creation of attitudes and transfer of information and ideas [6]. The two major approaches to reconstructing the discussion structure are the 'issue mapping' and the 'actor mapping'. 'Actor mapping' focuses on a dataset of bloggers that is checked for interconnectedness and requires a development of a pre-defined list of bloggers; however, such lists are easy to create only in case of top political bloggers or journalists. This study aims to test a methodological approach of 'actor mapping' on Instagram; more precisely, we aim at testing a method for creating a corpus of relevant accounts on Instagram. The methodological design we have developed aims to create a database of Instagram blogs about migration in Russian language beyond an issue-triggered discussion. The paper discusses also the restrictions of Instagram as platform and contributes to the general knowledge how Instagram bloggers use hashtags.

**Keywords:** Community engagement · Design and evaluation methodologies for social computing and social media · Language and culture in social computing and social media · Online special interest communities · Social identity and presence · Social network analysis · User generated content

## 1 Introduction

Mass media monopoly in publishing and disseminating information was destroyed after emerging of digital arenas – blog platforms and social network services. These platforms offered new possibilities to share human experience with other humans. As Manuel Castells stated in the middle of 2000-s, “I call it mass self-communication because it reaches tens of millions... It is self-communication because it is self-directed in the elaboration and sending of the message, self-selected in the reception of the message, and self-defined in terms of the formation of the communication space” [12]. Creation of user-generated content is seen by philosophers as a social act of “performative self-exposure and ... a performative exposure of taste and consumption” [27]. The ways how



humans exposure themselves to other humans in the digital world are formed, reproduced and developed under the influence of platform affordances [9], thus, gaining scholars' attention to the comparative studies of users' behavior on different platforms [14]. Still, to prove hypotheses about platforms' distinctions (for ex., whether a statistically significant distinction between users running Russian-language blogs about migration on Instagram or Telegram does exist) one might need tools to define the borders of a network that constrain all the thematically related blogs on at least one platform. In this paper we discuss methodological decisions and their restrictions, testing the algorithms on a case of Russian-speaking Instagram bloggers with migration background.

The structure of this paper is organized as follows. The next section provides a literature review on hashtags as a main tool for parsing the corpora of blogs on Twitter, peculiarities of hashtagging practices on Instagram and introduces the case study. We then present our methodology (Sect. 3) and the findings of our study (Sect. 4). We conclude with a short summary of our results.

## 2 Theoretical Framework

### 2.1 Development of Hashtagging: From Chats in the Nineties to Social Media Platforms

The usage of a “hash” symbol has been firstly described in the mid-Nineties when it was sued as a marker for topic-specific groups in Internet Relay Chat [20]. Since this very beginning, hashtags fulfilled “a function of ordering and systematizing” [20]. The conventional use of hashtagging became more widespread, and in 2009 Twitter conventionalized this grassroots practice with an official status. The microblog service hyperlinked hashtagged items and made them searchable. In a year a special service, Trending Topics, was launched to boost visibility and applicability of hashtags. Trending Topics, and later Tweetdeck, reflected the dynamics of the platform that served as the most operative channel to disseminate breaking news and even provide users with updates in a case of emergency such as floods or riots. “The value of hashtags as a mechanism for coordinating news discussion and information curation” [6] is even higher due to the development of special functionality that allows any user to find and follow all tweets containing particular hashtags. They played a crucial role for political uprisings and social movements connecting users who “were directly or indirectly involved in the events” [26]. In comparison with Twitter, Instagram as a platform is less news-oriented and its design strategies fit more to the development of a stable community [31]. Taking into consideration the difference between these two platforms, we will describe peculiarities of hashtagging practices in the next paragraph.

### 2.2 Instagram Hashtags: Creating a Folksonomy

Instagram only partly follows Twitter in terms of communicative structure. As well as Twitter, Instagram platform enables two types of users' networks [6]. First of them, follower – followee, is based on a long-term interest and forms relatively stable relationships. The second one arises around hashtags shared by several users. Even though

Instagram users also adopted hashtags for political and election processes [19], participatory hashtagging is rather depoliticized in comparison with Twitter [31]. While on Twitter hashtags “emerge almost instantaneously as news breaks” [6], forming ad hoc publics, “tagging on Instagram leads to the generation of a folksonomy” [23].

We understand folksonomy as a collaborative, collective, and social organization of information created by users who describe it explicitly with special metadata, as suggested by Angius et al. [2]. As opposite to a taxonomic system, a folksonomic system makes information more findable [13, 40]. Daer et al. underline that usage of a hashtag reveals that a marked post was “intended to be found and read by people searching for that specific term” [13]. Instagram users mostly tag their posts with content-related hashtags or thematizing context-marker hashtags [17, 41]. Not only Instagram users construct a folksonomy from “the total quantity of all assigned tags” [34]. Flickr has been researched as the first online platform with folksonomical tagging behavior [3, 30].

Contrary to a significant research of ad hoc publics constructed around hashtags, the use of hashtags in retrieval of information is underresearched [8]. Technical affordances of a hashtag and its potential in making an input in a social media platform more findable reorganize the online communication into a form of a ‘searchable talk’ [42].

The differences between the ways how users interact with hashtag practices on Instagram and Twitter might be explained with several reasons. Instagram started as a platform for sharing photos, and its functionality includes archiving and systematizing photo material. Secondly, Instagram is actively used as a platform for commercial marketing [11, 35]. In comparison with Trending topics, Instagram sorts users’ posts by hashtags and counts the post’s popularity within each hashtag labelling this post. Therefore, the choice of hashtags contributes not only to the visibility of the posts but also to the content monetization. Commercial accounts also rely on payed services that offer automated posting and adding hashtags, as well as suggest hashtags which usage will target commercial posts to a broaden audience. These methods of social media management are prohibited by Instagram, still, they are widely used.

### 2.3 Case Study: Global Russians

The paper on hand contributes to the perspectives of social network analysis by migration scholars. Even though transnational migration has been seen as potential environment for globalizing communication, SNA began to be usefully applied in migration studies only when it has already grown across a range of disciplines [37]. Scholars focus mostly on correlations between social media usage and bridging and bonding capital [4, 10, 15].

Studies of the structure of the networked communities by social network analysis interpret interconnectedness of communicators in different ways [22]. The two major approaches to reconstructing the discussion structure are the ‘issue mapping’ and the ‘actor mapping’. ‘Issue mapping’ [1, 25] traces conversations of social media users around a particular topic, tracing how ad hoc publics [7] or affective publics [32] emerge. The mechanism that brings these publics to (sometimes very short) life is usage of hashtags connected to the breaking news [6]. ‘Actor mapping’ focuses on a dataset of bloggers that is checked for interconnectedness. This method requires a development of a pre-defined list of bloggers; however, such lists are easy to create only in case of top

political bloggers or journalists [36]. Studying the variety of stand-alone blogs or blogs on social media platforms remains challenging for the media researchers [cf. 5, 18].

As a case study, we have chosen an undefined range of Instagram users running their blogs about migration experience and life in a new country and posting content on Russian. The geography of Russian-speaking communities outside Russia is wider than ever, while the total population is comparable in population with the population of the Russian Federation [33, 38]. The heterogenous Russian-speaking population abroad represents “the whole repertoire of migrant groups and identities” [33] embedded in different national and social contexts.

Diaspora news media are replaced today with individual blogs that provide useful information, disseminate experience and knowledge among migrants, allow for accumulating social capital, and provide psychological support. But, till today, the interconnectedness of the Russian ‘e-diaspora’ [16] has been studied only in terms of diasporic media and organizational websites [24, 29]. We address this gap by creating a corpus of blogs on Instagram run by the Russian-speaking users with migration background; in this task, understanding the borders of the community is the main challenge.

### 3 Methodology

Developing a research design for defining the borders of the community of Instagram blogs about migration in Russian language runs into several restrictions. It is impossible to use a search engine to identify Instagram blogs as a method developed in previous research on thematic blogs [39]. There are neither Russian-language blogs about migration included in the lists of top Instagram users nor special hashtags in the lists of top popular hashtags on Instagram. In our study, we partly follow the research design developed by Etling et al. [18] to identify a network of approximately 35,000 blogs in Arabic language. To create this database from a corpus of Arabic-, French-, and English-language blog data, expert interviews and several Arab-oriented content aggregators were used, as well as ‘snowballing’ techniques for identifying additional bloggers. We also used the idea of hashtag as a folksonomy unit and crawling tool [20, 21] at the initial step to get access to the field.

The methodological design we have developed aims to create a database of Instagram blogs about migration in Russian language beyond an issue-triggered discussion. Contrary to the previous research our design doesn’t require expert interviews or existence of blogger aggregators and is more suitable to a distant monitoring of a high-choice social media environment. In this study we tested the following algorithm. First, we collected general hashtags that described migration experience in Russian. Second, we extracted a list of users who used these hashtags via SocialKit. Third, we downloaded posts created by random sample of users and collected more hashtags used by bloggers to categorize their posts. A new list of users has been extracted based on these additional hashtags until no more users were found. In the Findings section we demonstrate the testing process in detail.

First testing of the algorithm has been run in April 2019, the data presented in the next section were downloaded in January 2020. According to ethical assumptions, we parsed the data only about public (non-private, open) accounts.

Suggesting the algorithm described above, we assumed that this methodological design allows to reach a saturation of data that can be described as a network border. Saturation in this case means that the set of hashtags allows to detect every blog about migration experience or another topic. The network border is reached when adding any new hashtag connected to the topic doesn't result in finding new users that were not earlier included in the list of bloggers. In the next section we will discuss the implementation of the algorithm, restrictions of hashtagging search in the Instagram studies.

## 4 Findings

### 4.1 General Evaluation of Network Size

As stated in the Methodology section, during the first step we defined the most general hashtags that could be used by a Russian-speaking Instagram user willing to blog about their migration experience in a host country (Table 1).

**Table 1.** General hashtags about migration experience.

| Hashtag            | English translation | Number of posts |
|--------------------|---------------------|-----------------|
| #жизньзаграницей   | life abroad (1)     | 131357          |
| #иммиграция        | immigration         | 93197           |
| #эмиграция         | emigration          | 73857           |
| #русскиезаграницей | Russians abroad     | 47632           |
| #пмж               | permanent residency | 41508           |
| #внж               | residency           | 41026           |
| #миграция          | migration           | 25533           |
| #жизньзарубежом    | life abroad (2)     | 13967           |

Using the tool “search for hashtags” we constructed tag clouds for each of 8 general hashtags. Additional hashtags belong to five groups:

- 1) misspelled versions of a hashtag
- 2) hashtag + emoji as one word
- 3) thematical extensions (for ex. “life abroad with children” or “immigration as it is”)
- 4) geographical extensions (for ex. “immigration to France” (any other country) or “permanent residence in Germany”)
- 5) hashtag that is a composite of the hashtag from the sample and an extension, but semantically doesn't represent migration experience. In Russian language the hashtag with meaning “life abroad (1)” can be a part of a hashtag with meaning “life beyond stereotypes” etc. These hashtags were excluded from the dataset.

Another tool that we used to broaden the set of general hashtags was the Instagram hashtag search that suggests other possible hashtags while giving out search results

**Table 2.** Geography of Russian-speaking Instagram (selected countries)

| N  | Country        | #Immigrationin... | #Russiansin... | #Lifein... |
|----|----------------|-------------------|----------------|------------|
| 1  | Australia      | 4557              | 9701           | 7054       |
| 2  | Austria        | 0                 | 7573           | 3292       |
| 3  | Belgium        | 0                 | 5983           | 1785       |
| 4  | Bulgaria       | 0                 | 5437           | 8342       |
| 5  | Brazil         | 0                 | 2810           | 26724      |
| 6  | Canada         | 22313             | 14199          | 40319      |
| 7  | China          | 0                 | 21736          | 40390      |
| 8  | Cyprus         | 0                 | 16345          | 22545      |
| 9  | Czech Republic | 1147              | 17799          | 16092      |
| 10 | Egypt          | 0                 | 2282           | 6616       |
| 11 | England        | 566               | 54662          | 25439      |
| 12 | Europe         | 6403              | 35111          | 57440      |
| 13 | Finland        | 0                 | 4387           | 3975       |
| 14 | France         | 347               | 39874          | 31878      |
| 15 | Germany        | 1682              | 200118         | 109210     |
| 17 | Ireland        | 132               | 1130           | 2255       |
| 18 | Israel         | 131               | 9620           | 16587      |
| 19 | Italy          | 1872              | 56614          | 44346      |
| 20 | Japan          | 0                 | 4232           | 9135       |
| 21 | Korea          | 0                 | 11419          | 21501      |
| 22 | Montenegro     | 194               | 3007           | 2933       |
| 23 | Netherlands    | 115               | 5045           | 5853       |
| 24 | New Zealand    | 1036              | 869            | 1182       |
| 25 | Poland         | 1271              | 3649           | 28532      |
| 26 | Slovakia       | 242               | 118            | 516        |
| 27 | Slovenia       | 700               | 892            | 1043       |
| 28 | Spain          | 4254              | 52259          | 47963      |
| 29 | Thailand       | 462               | 8601           | 13645      |
| 30 | Turkey         | 211               | 163954         | 16053      |
| 31 | UAE            | 0                 | 18502          | 7572       |
| 32 | Uruguay        | 533               | 520            | 10         |
| 33 | US             | 40726             | 272909         | 188879     |
| 34 | Vietnam        | 0                 | 2367           | 130193     |

on the main page. With this tool we received two general hashtags with the meaning #Russiansabroad (#русскиезаграницей – 47623 posts, #русскиезарубежом – 8 158 posts) and three hashtags with geographical extensions: #Russiansin(country), #bloggersin(country), #lifein(country). Although the hashtag #immigration is very popular as a general one, when a geographical extension is added, the hashtags #Russiansin(country) and #lifein(country) turn to be used much more often than a combination of #immigration hashtag and a particular country (Table 2).

Among other countries US, Canada, Europe, Australia and Spain are the most often mentioned countries if the post has been marked with the #immigrationin(country) hashtag. The gaps between three leaders are explicit: US is mentioned two times more often than Canada, while Canada became a geographical extension of the #immigration hashtag even three times more often than Europe. US remains on the leading position among posts with the hashtag #Russiansin(country), still, all other position are replaced and taken by Germany, Turkey, Italy and UK(England). The list of top-5 countries in the hashtag #Lifein(country) unites leaders from these two lists and includes US, Vietnam, Germany, Europe and Spain.

## 4.2 Collecting Users by a Hashtag: Restrictions

Due to the restrictions of the Instagram API we had to rely on data gathered via a social media management program – SocialKit, that was developed for working with Russian-speaking audiences.

While Instagram and SocialKit automatically provide data about the number of posts published up to date with a particular hashtag, finding out how many users used this hashtag is more difficult task, not to mention getting the full list of the users. Still, SocialKit gathers lists of users by hashtags with their metadata – the number of followers, the number of accounts they follow and the number of publications. The service also offers a possibility to filter the users by their activity, gender, business/private status as well as localization. Within our study we applied the same filter for each list of users downloaded with a hashtag from the dataset. It included following strings: to exclude private accounts, accounts without an avatar, accounts without Russian letters in the profile description; to include all users regardless of their gender and to include only users with more than 100 followers. Lifestyle industry defines Instagram users as nano-influencers since they have more than 1 000 followers [Maheshwari], hence, we decided to narrow our dataset only after the level of 100 followers because it is the lowest border of an average Instagram account (estimated as 100–300 followers).

To estimate the quality of the samples downloaded via a hashtag search in Social Kit, we choose randomly one hashtag, #RussiansinBelgium (5983 posts). We might expect that this hashtag will return us a search result with accounts defined by their owners as migrants and the geographical extension of a hashtag will result in a precise sample of users living in Belgium. Instagram accounts were evaluating manually by two categories: semantical correspondence (a blog run in Russian about experience in the new country) and geographical accuracy (whether the user is located in Belgium). SocialKit search inquiry resulted in a list of 255 users with the audience from 102 to 253,525 followers. Among these 255 users we registered 4 accounts serving as a translocal community and connecting blogs across the world; 55 accounts of blogs about migration

experience (not only about Belgium) and 50 accounts with some relation to the migration topicality (business accounts of Russian speaking users located in Belgium or blogs of Russian-speaking migrants without an explicit focus on their migration experience). 124 accounts, or 57% of the search result, failed to prove any connection to Belgium or migration topicality. This part of the list of users parsed via a hashtag #RussiansinBelgium included accounts of astrologists and fortune-tellers, fitness marathons, private blogs from different cities in Russia, business accounts selling Korean cosmetics etc. This result might be explained with the commercial nature of Instagram blogs that triggers business users of this platform to use hashtags in a constant rush for new clients.

To evaluate the scale of this platform restriction, a non-commercial parsing tool has been created. Its algorithm includes a synthetic search of a small number of accounts for one hashtag #lifeabroad (100 users) and a download of 100 last published posts. Within this dataset the parser found around 2900 hashtags, 93% of hashtags were evaluated as unique. Thus, the algorithm we elaborated at the initial step meets two critical restrictions: the level of noise (an unexpected share of accounts that use hashtags without any connections with their folksonomic nature) and the number of unique hashtags bloggers use to raise visibility and to trick the Instagram algorithms.

While our first intension in constructing the methodological design was to broaden the parsing tool to reach all the potential hashtags, the peculiarities of the platform lead to the significant level of the dataset contamination that exert a substantial influence upon the quality of the algorithm's results.

### 4.3 Defining an Instagram Blog with a Set of Hashtags

As we have shown in the previous paragraph, the findings suggest that the saturation of data understood as a case when the set of hashtags allows to detect every blog about migration experience topic is not accessible with an achievable reliability. Due to the dataset's contamination such data will without exception include a significant share of users that don't belong to the semantical field the study is focused on.

To solve the problem with dataset's contamination we tested the potential of merging users' lists parsed by several hashtags on the case of Spain. The list of hashtags included following ones: #bloggersSpain; #lifeinSpain, #lifeinSpainES; #immigrationtoSpain; #migrationtoSpain; #emigrationtoSpain and #lifeinSpain with several additions or misspelled versions. We narrowed the list of users to those who used at least three hashtags simultaneously. Within this small sample the level of contamination turned to be considerably low: only 3% of the accounts were absolutely not connected to the migration issues or Spain, and 8% of the accounts might be described as questionable (for ex. one belongs to a user from Belgium, three accounts were run by an migration agency without any focus on Spain in particular). Still, one might assume that this accuracy excludes some users who also post about their migration experience. Random manual checking of users with two hashtags on the list has shown that they also used general hashtags to mark their posts: #immigration or #lifeabroad.

We applied this algorithm to a dataset of 2900 hashtags parsed with the hashtag #lifeabroad. During the test download 67 users with hashtags were collected. We coded manually all the users based on their accounts' descriptions on Instagram and detected 38 accounts run by users with migration background (57% of the testing sample). The share

of unique hashtags remains significant, still, the hashtags relevant for the topic “migration” became more salient, for ex., #studyabroad, #Russiansin(country), #lifein(country). We have not detected any user that was coded as irrelevant and simultaneously possesses a combination of three topic-relevant hashtags in the testing sample.

## 5 Conclusion

We hypothesized that the network border can be reached with a list of hashtags that describes all possible blogs as full as possible. This approach preserves its functionality within discourse studies and contributes to a mapping of knowledge categories within an Instagram folksonomy. Hence, while aiming to map the blogosphere for further content analysis, it is needed to focus on the lists of users, and in this process expansion of the list of hashtags leads to the growth of the dataset’s contamination level. Consequently, more efforts are required to constant checking of the dataset whether the level of its contamination remains as lower as possible. Contrary to expectations, we conclude that the combination of general hashtags with localized hashtags is necessary and sufficient for saturating the data. Instead of iterations with hashtags’ search, the users’ lists should be iteratively scrubbed to minimize the level of dataset’s contamination before further analysis.

**Acknowledgements.** The research has been supported in full by Russian Presidential Grant for Young PhD Scientists, research grant MK-1448.2020.6.

## References

1. Alinejad, D., Candidatu, L., Mevsimler, M., Minchilli, C., Ponzanese, S., Van Der Vlist, F.N.: Diaspora and mapping methodologies: tracing transnational digital connections with ‘mattering maps’. *Glob. Netw.* **19**(1), 21–43 (2019)
2. Angius, A., Concas, G., Manca, D., Pani, F.E., Sanna, G.: Classification and indexing of web content based on a model of semantic social bookmarking. In: *Proceedings of the 6th International Conference on Knowledge Management and Information Sharing, KMIS 2014, Rome, Italy, 21–24 October 2014*. ISBN 978-989-758-050-5
3. Beaudoin, J.: Folksonomies: Flickr image tagging: patterns made visible. *Bulletin of the American Society for Information Science and Technology* **34**(1), 7–11 (2007)
4. Binder, J.F., Sutcliffe, A.G.: The best of both worlds? online ties and the alternating use of social network sites in the context of migration. *Societies* **4**(4), 753–769 (2014)
5. Bruns, A., Burgess, J., Highfield, T., Kirchoff, L., Nicolai, T.: Mapping the Australian networked public sphere. *Soc. Sci. Comput. Rev.* **29**(3), 277–287 (2011)
6. Bruns, A., Burgess, J.: Researching News Discussion on Twitter. *Journalism Stud.* **13**(5–6), 801–814 (2012). <https://doi.org/10.1080/1461670x.2012.664428>
7. Bruns, A., Burgess, J.: Twitter hashtags from ad hoc to calculated publics. In Rambukkana, N. (ed) *Hashtag Publics: The Power and Politics of Discursive Networks* (2015), pp. 13–28. Peter Lang (2015)
8. Buarki, H., Alkhateeb, B.: Use of hashtags to retrieve information on the web. *Electron. Libr.* **36**(2), 286–304 (2018). <https://doi.org/10.1108/el-01-2017-0011>



9. Bucher, T., Helmond, A.: The affordances of social media platforms. In Burgess J, Marwick A, Poell T (eds) *The SAGE Handbook of Social Media*, pp. 223–253. SAGE (2017)
10. Bucholtz, I.: Bridging bonds: Latvian migrants' interpersonal ties on social networking sites. *Media Cult. Soc.* **41**(1), 104–119 (2019)
11. Casaló, L.V., Flavián, C., Ibáñez-Sánchez, S.: Understanding consumer interaction on instagram: the role of satisfaction, hedonism, and content characteristics. *Cyberpsychology Behav. Soc. Netw.* **20**(6), 369–375 (2017)
12. Castells, M.: Communication, power and counter-power in the network society. *Int. J. Commun.* **1**(1), 29 (2007)
13. Daer, A.R., Hoffman, R., Goodman, S.: Rhetorical functions of hashtag forms across social media applications. In: *Proceedings of the 32nd ACM International Conference on the Design of Communication*, pp. 1–3 (2014)
14. DeVito, M.A., Birnholtz, J., Hancock, J.T.: Platforms, people, and perception: Using affordances to understand self-presentation on social media. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 740–754 (2017)
15. Dekker, R., Engbersen, G., Faber, M.: The use of online media in migration networks. *Popul. Space Place* **22**(6), 539–551 (2016)
16. Diminescu, D., Matthieu, R., Mehdi, B., Jacom, M.: Digital diasporas atlas exploration and cartography of diasporas in digital networks. In: *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 657–658 (2011)
17. Dorsch, I.: Content description on a mobile image sharing service: hashtags on instagram. *J. Inf. Sci. Theory Pract.* **6**(2), 46–61 (2018)
18. Etling, B., Kelly, J., Faris, R., Palfrey, J.: Mapping the Arabic blogosphere: politics and dissent online. *New Media Soc.* **12**(8), 1225–1243 (2010)
19. Gladchenko, I.A.: ThePresidentPeoples' Servant: political mobilization in online communications and social networks during the 2019 presidential election in Ukraine. *J. Polit. Stud.* **3**, 94–106 (2019)
20. Heyd, T., Puschmann, C.: Hashtagging and functional shift: adaptation and appropriation of the. *J. Pragmatics* **116**, 51–63 (2017). <https://doi.org/10.1016/j.pragma.2016.12.004>
21. Highfield, T., Leaver, T.: A methodology for mapping Instagram hashtags. *First Monday* **20**(1), 1–11 (2015)
22. Herring, S.C., et al.: Conversations in the blogosphere: an analysis “from the bottom up”. In: *IEEE Proceedings of the 38th Annual Hawaii International Conference on System Sciences (2005)*. <https://doi.org/10.1109/hicss.2005.167>
23. Ibba, S., Orrù, M., Pani, FE., Porru, S.: Hashtag of instagram: from Folksonomy to complex network. In: *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. 7th International Conference on Knowledge Engineering and Ontology Development. Lisbon, Portugal, 12.11.2015–14.11.2015*, pp. 279–284. SCITEPRESS - Science and Technology Publications (2015)
24. Karatzogianni, A., et al.: Intercultural conflict and dialogue in the transnational digital public sphere: findings from the Mig@ Net Research Project (2010–2013). In: Karatzogianni, A., Nguyen, D., Serafinelli, E. (eds.) *The Digital Transformation of the Public Sphere*, pp. 235–257. Palgrave Macmillan, London (2016)
25. Kok, S., Rogers, R.: Rethinking migration in the digital age: transglobalization and the Somali diaspora. *Glob. Netw.* **17**(1), 23–46 (2017)
26. Lee, C., Chau, D.: Language as pride, love, and hate: archiving emotions through multilingual instagram hashtags. *Discourse Context Media* **22**, 21–29 (2018). <https://doi.org/10.1016/j.dcm.2017.06.002>

27. Macek, J.: More than a desire for text: online participation and the social curation of content. *Convergence* **19**(3), 295–302 (2013)
28. Maheshwari, S.: Are You Ready for the Nanoinfluencers? *New York Times*, 11.11.2018. <https://www.nytimes.com/2018/11/11/business/media/nanoinfluencers-instagram-influencers.html>
29. Morgunova, O., Zinnurov, R.T.: Connected by digital imagination: discourses of belonging and the community building of Russophone migrants in the USA and Great Britain 1. In: Mustajoki, A., Protassova, E., Yelenevskaya, M. (eds) *The Soft Power of the Russian Language*, pp. 210–220. Routledge (2019)
30. Nov, O, Naaman, M., Ye, C.: What drives content tagging: the case of photos on flickr. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1097–1100. ACM, New York (2008)
31. Oh, C., Lee, T., Kim, Y., Park, S.H., Suh, B.: Understanding participatory hashtag practices on instagram. In: Kaye, J., Druin, A., Lampe, C., Morris, D., Hourcade, J.P. (eds.): *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*. the 2016 CHI Conference Extended Abstracts. Santa Clara, California, USA, 07.05.2016–12.05.2016, pp. 1280–1287. ACM Press, New York (2016)
32. Papacharissi, Z.: Affective publics and structures of storytelling: sentiment, events and mediality. *Inf. Commun. Soc.* **19**(3), 307–324 (2016)
33. Pechurina, A.: Post-Soviet Russian-speaking migration to the UK: the discourses of visibility and accountability. In: Nikolko, M., Carment, D. (eds.) *Post-Soviet Migration and Diasporas*, pp. 29–45. Palgrave Macmillan, Cham (2017)
34. Peters, I.: *Folksonomies: Indexing and retrieval in Web 2.0*. Berlin: De Gruyter Saur (2009)
35. Phua, J., Jin, S.V., Kim, J.J.: Gratifications of using Facebook, Twitter, Instagram, or Snapchat to follow brands: the moderating effect of social comparison, trust, tie strength, and network homophily on brand identification, brand engagement, brand commitment, and membership intention. *Telematics Inform.* **34**(1), 412–424 (2017)
36. Reese, S.D., Rutigliano, L., Hyun, K., Jeong, J.: Mapping the blogosphere: professional and citizen-based media in the global news arena. *Journalism* **8**(3), 235–261 (2007)
37. Ryan, L., D'Angelo, A.: Changing times: migrants' social network analysis and the challenges of longitudinal research. *Soc. Netw.* **53**, 148–158 (2018)
38. Ryazanova-Clarke, L.: Russian with an accent: globalization and the post-Soviet imaginary. In: Ryazanova-Clarke, L. (ed.) *The Russian Language Outside the Nation* (2006), pp. 249–280. Edinburgh University Press (2014)
39. Sharman, A.: Mapping the climate sceptical blogosphere. *Glob. Environ. Change* **26**, 159–170 (2014)
40. Vander Wal, T.: Folksonomy coinage and definition (2007). <http://www.vanderwal.net/folksonomy.html>
41. Veszelszki, Á.: #time, #truth, #tradition: an imagetext relationship on Instagram: Photo and hashtag. In: Benedek, A., Veszelszki, A. (eds.) *In the Beginning was the Image: The Omnipresence of Pictures: Time, Truth, Tradition*, pp. 93–113. Peter Lang, New York (2016)
42. Zappavigna, M., Martin, J.R.: Communing affiliation: social tagging as a resource for aligning around values in social media. *Discourse Context Media* **22**, 4–12 (2018). <https://doi.org/10.1016/j.dcm.2017.08.001>



# Effects of Linguistic Proficiency and Conversation Topic on Listener's Gaze in Triadic Conversation

Ichiro Umata<sup>1</sup>(✉), Koki Ijuin<sup>2</sup>, Tsuneo Kato<sup>3</sup>, and Seiichi Yamamoto<sup>3</sup>

<sup>1</sup> KDDI Research, Inc., Garden Air Tower, 3-10-10, Iidabashi,  
Chiyoda-ku, Tokyo 102-8460, Japan  
ic-umata@kddi-research.jp

<sup>2</sup> The National Institute of Advanced Industrial Science and Technology,  
2-3-26, Aomi, Koto-ku, Tokyo 135-0064, Japan  
koki-ijuin@aist.go.jp

<sup>3</sup> Department of Information Systems Design, Doshisha University,  
Kyotanabe-shi, Kyoto 610-0321, Japan  
{tsukato,seyamamo}@mail.doshisha.ac.jp

**Abstract.** Gaze is reported to have important functions in communication, such as expressing emotional states, exercising social control, highlighting the informational structure of speech, and coordinating floor-apportionment. For these reasons, studying these communicative functions is expected to contribute to HCI systems by identifying communication characteristics and the role of each participant. This study analyzes how the communicative functions of utterances affect the listener's gazing activities from the viewpoint of grounding, based on a triadic corpus with newly labeled grounding tags. The results showed that the duration of a listener's gaze is longer in second language (L2) conversations, in goal-oriented conversations, and during utterances presenting new information. These results suggest that linguistic proficiency, conversation topic, and grounding factors all affect a listener's gazing activities, providing us with some information that could assist in the design of HCI, HRI, and CSCW systems that better reflect the interaction contexts and linguistic proficiency of users.

**Keywords:** Gaze · Communication · Grounding · Topic · Linguistic proficiency

## 1 Introduction

Nonverbal cues play important roles in everyday communication, and studies have examined their importance not only from the affectional and attitudinal aspects of communication (Mehrabian and Ferris 1967; Mehrabian and Wiener 1967), but also in coordination of communication and “grounding”, i.e. constructing a shared understanding of the communication context (Clark and Brennan 1991; Clark 1996; Clark and Krych 2004), suggesting that the potential exists to expand and augment HCI (human-computer interaction) systems.

Among nonverbal modalities in communication, gazing activities have been considered fundamental, attracting considerable attention from researchers working in the area of multi-modal communication. Studies have reported that gaze has important communicative functions including expressing emotional states, exercising social control, highlighting the informational structure of speech, and organizing the speech floor (Argyle et al. 1968; Duncan 1972; Holler and Kendrick 2015; Kendon 1967). From the viewpoint of interaction organization in communication, studies have reported that gaze can be a cue for speech floor coordination not only in dyadic (Kendon 1967), but also in multi-party conversations (Kalma 1992; Learner 2003). Although the findings of other studies are not necessarily entirely consistent with those of the studies mentioned above (Beattie 1978; Rutter et al. 1978), this can probably be attributed to the multi-functional nature of gaze in communication (Kleinke 1986), while recent studies have confirmed the speech floor coordination function of gaze for dyadic (Ho et al. 2015) and multi-party conversations (Jokinen et al. 2013; Ishii et al. 2016; Versteeg et al. 2001; Ijuin et al. 2018). Another study indicated that gaze can be a collaborative signal that serves as a cue to coordinate the insertion of responses (Bavelas et al. 2002). Furthermore, another study reported that even uninvolved observers of dyadic interactions followed the interactants' speaking turns with their gaze (Hirvenkari et al. 2013).

Inspired by and based on studies that examined the functions of social gaze, system studies that incorporate gaze modalities have been proposed in the HCI and CSCW (computer-supported cooperative work) fields. Such studies have covered not only conversational agents (Cassel et al. 1994; Versteeg et al. 2001; Garau et al. 2001; Heylen et al. 2005; Rehm et al. 2005) but also robots and devices with simulated gaze expression (Sidner et al. 2004; Bennewitz et al. 2005; Kuno et al. 2007; Foster et al. 2012; Lala et al. 2019; Jaber et al. 2019; McMillan et al. 2019).

Although HCI, HRI (human-robot interaction) and CSCW systems have to some extent been able to take gazing cues into account and integrate gaze functions, they have been less successful in incorporating linguistic proficiency that may affect gazing activities. A remote work study in the HCI field argued that video transmission of facial information and gesture helped non-native pairs to negotiate a common ground, whereas this did not provide significant help for native pairs (Veinott et al. 1999). An analysis of second language conversation reported that eye gazes and facial expressions play an important role in monitoring both partners' understanding in the repair process (i.e. a modification to the content or presentation of the current proposition under consideration (Schegloff et al. 1977; Traum 1994)) where participants with different levels of linguistic proficiency are involved (Hosoda 2006).

Some quantitative studies have also examined the effect of linguistic proficiency on the speech floor coordination function of gaze. Analyses of the duration of the listener's gaze during utterances have shown that when other participants are looking at the speaker in a second language (L2) conversation, the duration is significantly longer than in a first language (L1) conversation (Yamamoto et al. 2013; Umata et al. 2013; Yamamoto et al. 2015). These studies, however, have not considered the communicative context effects. Kleinke pointed out that the conditions of a conversational setup may affect the relative importance of the multiple functions of gaze in communication (Kleinke et al. 1986). Holler and Kendrick analyzed three-party conversations among native English speakers,

showing that unaddressed participants were able to anticipate next turns in question-response sequences involving just two of the participants (Holler and Kendrick 2015). There are also studies that have shown the effects of interaction contexts on gazing behavior in social interactions (Rossana 2013; Kendrick and Holler 2017; Rossana et al. 2009; Stivers and Rossana 2010). The role of gaze in communication is affected by the context, and it is important to analyze the function of gaze during utterances while taking their communicative function into consideration.

The current study examined the effects of linguistic proficiency on the listener's gaze in triadic communication considering the communicative function of utterances from the viewpoint of grounding. Each utterance was categorized according to the grounding acts in the dialogue, and the gazing activities of the listeners were compared between native and the second language conversations. We anticipated that conversation topics could also affect a listener's gazing activities, and included the topic factor in our analysis. The results suggest that both language proficiency and topic factors independently affect the duration of a listener's gaze in utterances in cases where the speaker provides some new pieces of information, but not in utterances where they just acknowledge the previous speaker's utterance.

## 2 Corpus

Our analysis is based on a multimodal triadic interaction corpus with eye-gaze data collected and analyzed in previous studies (Yamamoto et al. 2015; Ijuin et al. 2018; Umata et al. 2018).

The corpus consists of triadic conversations in a mother tongue (L1) and those in a second language (L2) made by the same interlocutors in the same group (for details, refer to Yamamoto et al. 2015). For the current study, all utterances were newly labeled with grounding act tags (details are provided below in this section), and all the conversation data were subjected to analysis in this study. A total of 60 subjects (23 females and 37 males: 20 groups) between the ages of 18 and 24 participated in data collection, and each conversational group consisted of three participants. All participants were native Japanese speakers.

Their seats were placed about 1.5 m apart from each other in a triangular formation around a round table (see Fig. 1 and Fig. 2). The corpus covers two conversation types to examine whether such differences in types affect their interaction behaviors.

The first type is free-flowing, natural chatting that ranges over various topics such as hobbies, weekend plans, studies, and travels. The other type is goal-oriented, in which participants collaboratively decided what to take with them on trips to uninhabited islands or mountains. All the participants would be under pressure to contribute to the conversation to reach an agreement in the goal-oriented conversations, whereas such pressure would not be so strong in free-flowing conversations where reaching an agreement was not obligatory.

We expected that conversational flow would be more predictable in the goal-oriented conversations where the vocabulary was more limited and the domain of the discourse was defined more narrowly by the task than in the free-flowing conversations.

The order of the conversation types was arranged randomly to counterbalance any order effect. The order of the languages used in the conversations was also arranged

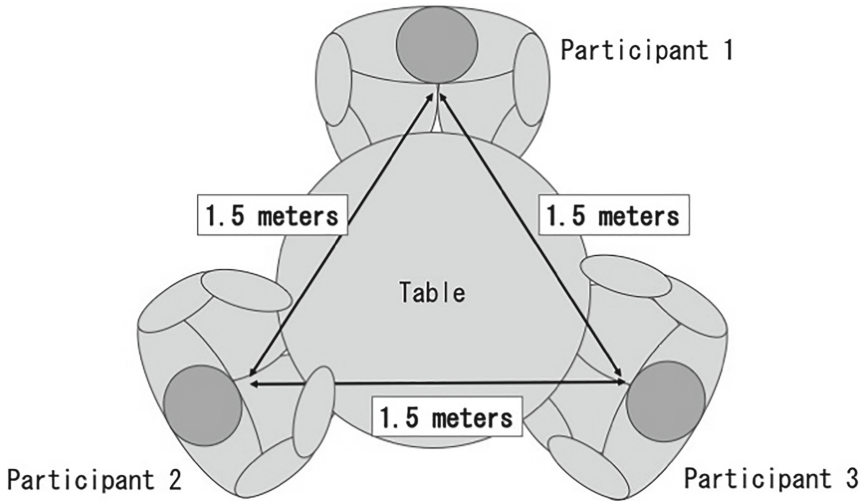


Fig. 1. Seating positions of the three participants.



Fig. 2. Seating positions of the three participants.

randomly. Each group had approximately six-minute conversations of the two types in both Japanese and English. We collected multimodal data from 80 three-party conversations in L1 (Japanese) and in L2 (English) languages (20 free-flowing in Japanese, 20 free-flowing in English, 20 goal-oriented in Japanese, and 20 goal-oriented in English). Twenty groups engaged in all four conversation types. All the participants except those in the first three groups answered a questionnaire evaluating their conversation after each conversation condition. This material is to be analyzed in other studies (see Umata et al. 2013).

Their eye gazes and voices were recorded via three sets of NAC EMR-9 head-mounted eye trackers and headsets with microphones. The viewing angle of the EMR-9 was 62° and the sampling rate was 60 frames per second. We used the EUDICO Linguistic Annotator (ELAN) developed by the Max Planck Institute as the tool for gaze and utterance annotation (ELAN) (see Fig. 3). Each utterance is segmented from speech at inserted pauses of more than 500 ms, and the corpus was manually annotated in term of the time spans for utterances, backchannel, laughing, and eye movements. The corpus already had the grounding act tags according to the categories established by Traum (1994) for 20 groups engaging in goal-oriented conversations. For the current study, we trained a university student to perform annotation according to the categories for 20 groups engaging in free-flowing conversations. She annotated the tags using ELAN with video, gaze, and utterance transcription data in the same manner as in the previous study (Umata et al. 2019). Table 1 shows the grounding act tags and their descriptions, and Fig. 1 shows the frequency of grounding acts in L1 and L2 conversation.

**Table 1.** Traum’s grounding acts

| Grounding act                              | Description  |
|--|--|
| Initiate ( <i>init</i> )                   | The initial presentation of a proposition  |
| Continue ( <i>cont</i> )                   | A continuation of a previous act performed by the same speaker                               |
| Repair                                     | A modification to the content or presentation of the current proposition under consideration |
| Request-Repair ( <i>reqRepair</i> )        | A request that the other participant perform a Repair  |
| Acknowledge ( <i>ack</i> )                 | Evidence that a previous utterance has been understood                                       |
| Request-Acknowledge ( <i>reqAck</i> )      | A request that the other participant perform an Acknowledge                                  |
| Cancel                                     | An abandonment of the proposition under consideration  |
| Acknowledge - Initiate ( <i>ack init</i> ) | “ack” and “init” occurring at the same time in one utterance unit                            |

### 3 Analyses of Gazes in Utterances

We analyzed the gazing activities of listeners in triadic conversation taking the factors of linguistic proficiency, topic and grounding into account. Previous studies of the listener's gaze during utterances have shown that when other participants are looking at the speaker in a second language (L2) conversation, gaze is significantly longer than in a first language (L1) conversation (Yamamoto et al. 2013; Umata et al. 2013; Yamamoto et al. 2015), suggesting that listeners use visual information to compensate for their lack in linguistic proficiency in an L2 conversation. We assumed that the linguistic proficiency factor would affect the listener's gazing activity. We also assumed that listeners would rely more heavily on visual information in a collaborative task where the requirement for communication organization is strong. The grounding act factor was also expected to affect the gazing activity of listeners; i.e. they would have greater reliance on visual information during an utterance in which new information is presented. Our hypotheses are listed as follows:

- H1:** The linguistic proficiency factor would affect the duration of a listener's gaze: The listeners would gaze at the speaker for longer in second language conversations where they compensate for their lack of linguistic proficiency with gazing cues.
- H2:** The topic factor would affect the duration of a listener's gaze: The listeners would gaze at the speaker for longer in goal-oriented conversation where the requirement for communication organization is stronger when an agreement has to be reached.
- H3:** The grounding act factor would affect the duration of a listener's gaze: The listeners would gaze at the speaker for longer during utterances presenting new information (namely, *init*, *cont* and *ack init*) than in utterances just acknowledging the previous utterance

We compared the duration of each listener's gaze during four major categories of grounding acts (i.e., *init*, *ack init*, *cont*, *ack*) between L1 and L2 conversations. We used the average of the listener's gazing ratio to analyze how long the speaker was gazed at by other participants [4]. The average of listener's gazing ratios was defined as:

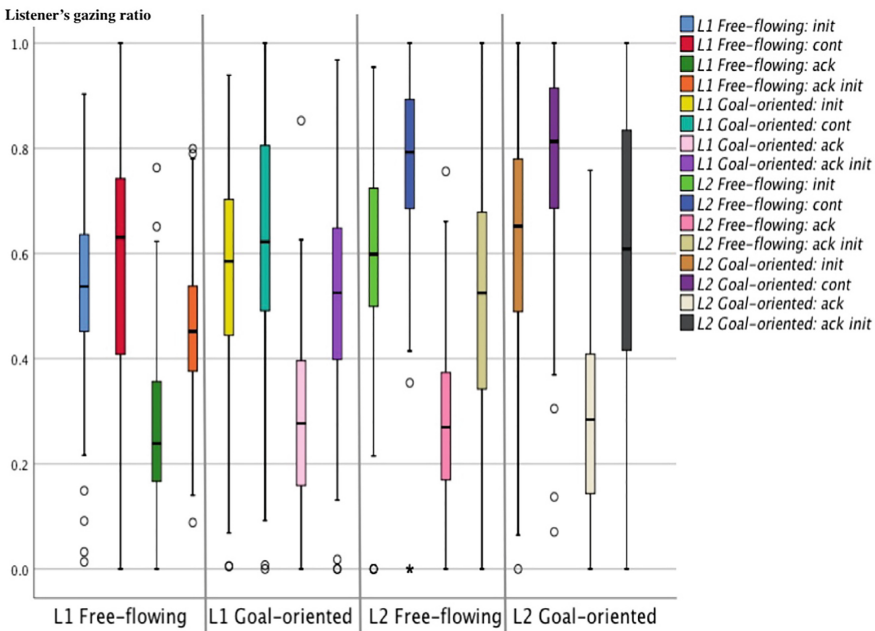
$$\text{Average of listener's gazing ratio} = \frac{1}{n} \sum_{i=1}^n \frac{DLG_j(i)}{D(i)}$$

Here,  $D(i)$  is the duration of the  $i$ th utterance and  $DLG_j(i)$  is the total gaze duration of the  $j$ th participant ( $j = 1, 2, 3$ ) in each group gazing at the speaker in the  $i$ th utterance.

We expected that the topic factor would affect the duration of the listener's gaze: the listeners would gaze at the speaker for longer in a goal-oriented conversation where they collaboratively decided what to take with them on a trip to a deserted island or to the mountains. We also expected that the linguistic proficiency factor would affect the duration of the listener's gaze: the listeners would gaze at the speaker for longer in second language conversations where they compensate for their lack of linguistic proficiency with gazing cues, especially in speech turn organization. We conducted an analysis of variance (ANOVA) with language difference, topic difference, and grounding act as within-subject factors. The results revealed significant main effects of language



( $F_{(1, 113)} = 45.875, p < .001$ ), topic ( $F_{(1, 113)} = 16.416, p < .001$ ), and grounding act ( $F_{(2.589, 292.612)} = 204.8, p < .01$ ), and the multiple comparison analysis showed the differences among four major grounding acts were all significant ( $p < .001$ ). Also, we observed significant first-order interaction between language and grounding acts ( $F_{(2.702, 305.327)} = 24.551, p < .001$ ), and between topic and grounding act ( $F_{(3, 339)} = 4.516, p < .005$ ). Sub-effect tests showed significant simple main effects of language in grounding act “*init*” ( $F_{(1, 113)} = 15.81, p < .001$ ), “*cont*” ( $F_{(1, 113)} = 78.20, p < .001$ ), and “*ack-init*” ( $F_{(1, 113)} = 12.20, p < .01$ ), and topic in grounding act “*cont*” ( $F_{(1, 113)} = 6.22, p < .05$ ) “*ack-init*” ( $F_{(1, 113)} = 12.20, p < .01$ ), and a marginally significant simple main effect of topic in grounding act “*init*” ( $F_{(1, 113)} = 3.18, p < .1$ ), but no significant simple main effect of either language nor topic in grounding act “*ack*”. The distribution of listeners’ gazing ratios (LGRs) is shown in the figure below.



**Fig. 3.** The distribution of listeners’ gazing ratios (LGRs)

As shown in Fig. 3, listeners gazed at the speaker for longer in L2 conversations than in L1 conversations. This was also the case in goal-oriented conversations compared to free-flowing conversations. Moreover, listeners gazed at the speaker for longer during *init*, *cont* and *ack init* utterances.

### 4 Discussion

We compared the duration of the listener’s gaze in triadic conversations to examine the effects of linguistic proficiency, topic and grounding on gazing activities. The results of

ANOVA revealed significant main effects of language difference, topic and grounding, supporting our hypotheses H1, H2, and H3: the duration of a listener's gaze is longer in L2 conversations, in goal-oriented conversations, and in *init*, *cont* and *ack init* utterances. The grounding factor had the greatest effect, followed by that of language proficiency.

The multiple comparison analysis showed that the differences among four major grounding acts were also all significant, and *cont* showed the longest duration for the listener's gaze among all the grounding act categories. With *cont* utterances, the speakers were adding new pieces of information to their own previous utterances, and in doing so, they were sometimes observed using a filled pause to hold the speech floor while bringing order to their ideas. Such characteristics of *cont* utterances might have drawn the listener's attention to the speaker. In contrast, *ack* showed the shortest duration for the listener's gaze, suggesting that utterances just acknowledging the previous utterance without adding new information did not draw the listener's visual attention to the speaker.

We observed significant first-order interaction between language difference and grounding acts, and sub-effect tests showed significant simple main effects of language difference in all the major grounding act categories except *ack*. The results suggest that linguistic proficiency affected the listener's gazing activities only for utterances presenting new pieces of information. Similarly, we observed significant first-order interaction between topic and grounding acts, and sub-effect tests showed significant or marginally significant simple main effects of topic in all the major grounding act categories except *ack*. The results suggest that the topic also affected the listener's gazing activities but only in the case where utterances presented new pieces of information.

Another interesting finding is that there was no significant interaction between the factors of language difference and topic. It suggests that in the current corpus settings, linguistic proficiency and topic independently affected the listener's gazing activities.

These findings suggest that linguistic proficiency, conversation topic, and grounding all affect the listener's gazing activities, and that these factors should be considered when attempting to design better HCI, HRI, and CSCW systems. It is also likely that the effects of these factors may not be just simple and independent but rather interlaced: our experimental results suggest that linguistic proficiency and grounding factors affect each other, and so do topic and grounding factors. Further detailed analyses are necessary to establish system design guidelines that reflect these factors.

## 5 Summary

We analyzed the effect of linguistic proficiency and conversation topic on the listener's gaze in four major grounding acts. The results showed that the duration of a listener's gaze is longer in second language (L2) conversations, in goal-oriented conversations, and during utterances presenting new information. The results also showed that both language proficiency and topic independently affect the duration of the listener's gaze in utterances presenting new information. These results suggest that linguistic proficiency, conversation topic, and grounding factors all affect a listener's gazing activities, supporting our hypotheses. The results are expected to contribute to HCI, HRI, and CSCW system design that reflects the interaction context and the linguistic proficiency of users.

## References

- Mehrabian, A., Ferris, S.R.: Inference of attitudes from nonverbal communication in two channels. *J. Consul. Psychol.* **31**(3), 248–252 (1967)
- Mehrabian, A., Wiener, M.: Decoding of inconsistent communications. *J. Pers. Soc. Psychol.* **6**(1), 109–114 (1967)
- Clark, H.H., Brennan, S.E.: Grounding in communication. In: Resnik, L.B., Levine, J.M., Teasley, S.D. (eds.) *Perspectives on Socially Shared Cognition*, pp. 503–512. APA Books (1991)
- Clark, H.H.: *Using Language*. Cambridge University Press, Cambridge (1996)
- Clark, H.H., Krych, M.A.: Speaking while monitoring addressees for understanding. *J. Mem. Lang.* **50**, 62–81 (2004)
- Argyle, M., Lallijee, M., Cook, M.: The effects of visibility on interaction in a dyad. *Hum. Relat.* **21**, 3–17 (1968)
- Duncan, S.: Some signals and rules for taking speaking turns in conversations. *J. Pers. Soc. Psychol.* **23**, 283–292 (1972)
- Holler, J., Kendrick, K.H.: Unaddressed participants' gaze in multi-person interaction: optimizing reciprocity. *Front. Psychol.* **6**, 14, Article no. 98 (2015). <https://doi.org/10.3389/fpsyg.2015.00098>
- Kendon, A.: Some functions of gaze-direction in social interaction. *Acta Physiol.* **26**, 22–63 (1967)
- Kalma, A.: Gazing in triads: a powerful signal in floor apportionment. *Br. J. Soc. Psychol.* **31**(1), 21–39 (1992)
- Lerner, G.H.: Selecting next speaker: the context-sensitive operation of a context-free organization. *Lang. Soc.* **32**(02), 177–201 (2003)
- Beattie, G.W.: Floor apportionment and gaze in conversational dyads. *Br. J. Soc. Clin. Psychol.* **17**, 7–15 (1978)
- Rutter, D.R., Stephenson, G.M., Ayling, K., White, P.A.: The timing of looks in dyadic conversation. *Br. J. Soc. Clin. Psychol.* **17**, 17–21 (1978)
- Kleinke, C.L.: Gaze and eye contact: a research review. *Psychol. Bull.* **100**, 78–100 (1986)
- Ho, S., Foulsham, T., Kingstone, A.: Speaking and listening with the eyes: gaze signaling during dyadic interactions. *PLoS ONE* **10**(8), e0136905 (2015). <https://doi.org/10.1371/journal.pone.0136905>
- Jokinen, K., Furukawa, H., Nishida, M., Yamamoto, S.: Gaze and turn-taking behavior in casual conversational interactions. *ACM Trans. Interact. Intell. Syst. (TiiS)* **3**(2), 1–30 (2013)
- Ishii, R., Otsuka, K., Kumano, S., Yamato, J.: Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Trans. Interact. Intell. Syst.* **6**(1), 31, Article no. 4 (2016). <https://doi.org/10.1145/2757284>
- Vertegaal, R., Slagter, R., Verr, G., Nijholt, A.: Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In: *CHI 2001 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 301–308. ACM Press, Seattle (2001)
- Ijuin, K., Umata, I., Kato, T., Yamamoto, S.: Difference in eye gaze for floor apportionment in native- and second-language conversations. *J. Nonverbal Behav.* **42**, 113–128 (2018)
- Bavelas, J.B., Coates, L., Johnson, T.: Listener responses as a collaborative process: the role of gaze. *J. Commun.* **52**(3), 566–580 (2002). <https://doi.org/10.1111/j.1460-2466.2002.tb02562.x>
- Hirvenkari, L., Ruusuvoori, J., Saarinen, V.M., Kivioja, M., Peräkylä, A., Hari, R.: Influence of turn-taking in a two-person conversation on the gaze of a viewer. *PLoS One* **8**(8), e71569 (2013). <https://doi.org/10.1371/journal.pone.0071569>
- Cassell, J., et al.: Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In: *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1994)*, pp. 413–420. ACM, New York (1994). <https://doi.org/10.1145/192161.192272>

- Garau, M., Slater, M., Bee, S., Sasse, M.A.: The impact of eye gaze on communication using humanoid avatars. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2001), pp. 309–316. ACM, New York (2001). <https://doi.org/10.1145/365024.365121>
- Heylen, D., van Es, I., Nijholt, A., van Dijk, B.: Controlling the gaze of conversational agents. In: van Kuppevelt, J.C.J., Dybkjær, L., Bernsen, N.O. (eds.) *Advances in Natural Multimodal Dialogue Systems*. TLTB, pp. 245–262. Springer, Netherlands (2005). [https://doi.org/10.1007/1-4020-3933-6\\_11](https://doi.org/10.1007/1-4020-3933-6_11)
- Rehm, M., André, E.: Where do they look? Gaze behaviors of multiple users interacting with an embodied conversational agent. In: Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., Rist, T. (eds.) *IVA 2005*. LNCS (LNAI), vol. 3661, pp. 241–252. Springer, Heidelberg (2005). [https://doi.org/10.1007/11550617\\_21](https://doi.org/10.1007/11550617_21)
- Sidner, C.L., Kidd, C.D., Lee, C., Lesh, N.: Where to look: a study of human-robot engagement. In: *IUI 2004: Proceedings of the 9th International Conference on Intelligent User Interfaces*, pp. 78–84 (2004). <https://doi.org/10.1145/964456.964458>
- Bennewitz, M., Faber, F., Joho, D., Schreiber, M., Behnke, S.: Integrating vision and speech for conversations with multiple persons. In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2523–2528 (2005). <https://doi.org/10.1109/IROS.2005.1545158>
- Kuno, Y., Sadazuka, K., Michie Kawashima, M., Yamazaki, K., Yamazaki, A., Kuzuoka, H.: Museum guide robot based on sociological interaction analysis. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007)*, pp. 1191–1194. ACM, New York (2007). <https://doi.org/10.1145/1240624.1240804>
- Foster, M.E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., Petrick, R.P.A.: Two people walk into a bar: dynamic multi-party social interaction with a robot agent. In: *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI 2012)*, pp. 3–10. ACM, New York (2012). <https://doi.org/10.1145/2388676.2388680>
- Lala, D., Inoue, K., Kawahara, T.: Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In: *2019 International Conference on Multimodal Interaction (ICMI 2019)*, Suzhou, China, 14–18 October 2019, 9 p. ACM, New York (2019). <https://doi.org/10.1145/3340555.3353727>
- Jaber, R., McMillan, D., Belenguer, J.S., Brown, B.: Patterns of gaze in speech agent interaction. In: *1st International Conference on Conversational User Interfaces (CUI 2019)*, Dublin, Ireland, 22–23 August 2019, 10 p. ACM, New York (2019). <https://doi.org/10.1145/3342775.3342791>
- Veinott, E., Olson, J., Olson, G., Fu, X.: Video helps remote work: speakers who need to negotiate common ground benefit from seeing each other. In: *Proceedings of the SIGCHI Conference on Computer Human Interaction (CHI 1999)*, pp. 302–309. ACM, New York (1999)
- Schegloff, E.A., Jefferson, G., Sacks, H.: The preference for self-correction in the organization of repair in conversation. *Language* **53**(1977), 361–382 (1977)
- Traum, D.R.: *A computational theory of grounding in natural language conversation*. Ph.D. dissertation, University of Rochester (1994)
- Hosoda, Y.: Repair and relevance of differential language expertise. *Appl. Linguist.* **27**(2006), 25–50 (2006)
- Yamamoto, S., Taguchi, K., Umata, I., Kabashima, K., Nishida, M.: Differences in interactional attitudes in native and second language conversations: quantitative analyses of multimodal three-party corpus. In: *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, pp. 3823–3828 (2013)
- Umata, I., Yamamoto, S., Ijuin, K., Nishida, M.: Effects of language proficiency on eye-gaze in second language conversations: toward supporting second language collaboration. In: *Proceedings of the International Conference on Multimodal Interaction (ICMI 2013)*, pp. 413–419 (2013)

- Yamamoto, S., Taguchi, K., Ijuin, K., Umata, I., Nishida, M.: Multimodal corpus of multiparty conversations in L1 and L2 languages and findings obtained from it. *Lang. Resour. Eval.* **49**, 857–882 (2015). <https://doi.org/10.1007/s10579-015-9299-2>
- Rossano, F.: Gaze in conversation. In: Stivers, T., Sidnell, J. (ed.) *The Handbook of Conversation Analysis*, pp. 308–329. Wiley-Blackwell, Malden (2013). <https://doi.org/10.1002/9781118325001.ch15>
- Kendrick, K.H., Holler, J.: Gaze direction signals response preference in conversation (2017)
- Rossano, F., Brown, P., Levinson, S.C.: Gaze, Questioning, and Culture, pp. 187–249. Cambridge University Press, Cambridge (2009). <https://doi.org/10.1017/CBO9780511635670.008>
- Stivers, T., Rossano, F.: Mobilizing response. *Res. Lang. Soc. Interact.* **43**, 3–31 (2010)
- Umata, I., Ijuin, K., Kato, T., Yamamoto, S.: Floor apportionment function of speaker’s gaze in grounding acts. In: *Proceedings of ICMI 2019: International Conference on Multimodal Interaction (ICMI 2019 Adjunct)*, Suzhou, China, 14–18 October 2019, 7 p. ACM, New York (2019). <https://doi.org/10.1145/3351529.3360660>
- ELAN. <http://www.lat-mpi.eu/tools/elan>



# The Confidence in Social Media Platforms and Private Messaging

Jukka Vuorinen<sup>1</sup>, Aki Koivula<sup>2</sup>(✉), and Ilkka Koiranen<sup>2</sup>

<sup>1</sup> Unit of Information Systems Science, University of Turku, 20014 Turku, Finland

jukka.vuorinen@utu.fi

<sup>2</sup> Unit of Economic Sociology, University of Turku, 20014 Turku, Finland

{akjeko, ilalko}@utu.fi

**Abstract.** In this paper, we focus on social media users and examine the factors predicting users' confidence in platforms in case of private messaging. For the social media platforms, social ties and information that flows through the contacts are valuable assets, which must be considered in the development of the services (such as messaging applications) in order to attract users. We use nationally representative data derived from surveys targeted at 15- to 74-year-old Finns (N = 3,724). The measures included user's confidence in platform services in social messaging, trust in social ties on social media, size of social media networks, a wide selection of measures related to internet and social media behavior, and demographic factors. The results of the analysis supported the hypothesis that high confidence in platforms is strongly dependent on the social resources of users. Network size and trust in social ties were crucial variables in determining the confidence in social media platforms as a secure channel of private messages. The results also amplified that trust in social media networks has independent explanation power in the platform confidence apart from behavioral and demographic factors. The findings are significant in terms of understanding the contemporary information society and dynamics between platform services and users. The markets of social media platforms and other agents in the sector are dependent on the social resources of users, and especially on the social trust of users.

**Keywords:** Social media · Messaging · Privacy · Trust · Social ties · Survey

## 1 Introduction

Different messaging applications (e.g., Whatsapp, Facebook messenger) provide a possibility to strengthen and uphold social ties through communication. In the past decade, internet users have seized this opportunity. Mostly, this has put pressure on the reliability and confidentiality of the platforms providing such services including private messages. Privacy makes it possible to deepen and intensify the social relationships especially in terms of sharing sensitive information with a strictly confined and carefully chosen group of people. More precisely, sharing sensitive information or secrets decreases the social distance between the parties of communication [1].

From the user point of view, in order to share sensitive information, the service provider needs to be considered as a trustworthy actor, and additionally, the security posture of the platform needs to be seen trustworthy. The reputation of a platform as a reliable actor enabling confidentiality is a significant asset in maintaining the high number of users that is a necessity for the success of the platform. For example, the recent privacy concerns, such as the cases of The Cambridge Analytica and Google+ data leaks, have indicated that the deteriorations of users' confidence in the platform may eventually result in abandoning the services. After the data breach, in which hackers were able to invade Google+'s system, Google announced that they are shutting down the whole platform. In this respect, platforms need to ensure they do not misuse private information themselves and, they need to protect the users' private content from the third parties, such as hackers.

However, despite the recent privacy concerns, the most important social media and messaging platforms are still popular and actively in use, and, to some extent, they have been able to continue to gain popularity in developing countries [2]. This raises the question of whether there are underlying factors that maintain users' confidence in the platform despite the public security risks manifested in protecting of user's privacy. Trust is a tricky phenomenon in terms of how it draws attention. As there is trust, high confidence in the platform, then trust and social ties create and intensify activity on the platforms. Thus, trust brings about content (messages, updates, posts, comments) which in turn creates reactions.

Furthermore, every private message sent through the system (re-)builds and (re-)establishes trust in the system and makes it seem the standard way of communicating. It materializes the fact that the platform is usable in terms of sharing sensitive information. As something becomes repeated it becomes mundane, normal and the standard way of acting. When it becomes the obvious way of acting, it does not draw attention anymore but rather it becomes invisible. Normality, in turn, hinders critical questions about the safety of the service [3]. Thus, as more and more users use a platform, it forms an echoing message of safety. However, when a platform becomes compromised, all the attention is on the platform.

In this study, we seek to catch these vivid social ties by estimating how actively users are engaged to the platform through several attributes. We are interested in how the size of the social media networks and trust in these networks explain the level of confidence in social media platforms. The size of the networks gives us a rough but applicable measure to assess how much potential (social contacts) network contains. In addition, by assessing how much users trust in their social ties gives us another aspect of estimating in how active way users are engaged to these networks. Secondly, we are interested in how different behavioral elements are affecting to interconnections of the level of confidence in social media platforms and primary types of engagement mentioned above. These behavioral elements are measured with the frequency of internet use, the frequency of social media use and skill set for social media use. Additionally, we measure how privacy skills and messaging application preferences are affecting to the level of confidence in social media platforms and primary types of engagement. By splitting the engagement of users to different elements, we can assess more accurately how different components affect the level of confidence in social media platforms.

## 2 The Modes of Connection and Interaction

As an individual becomes connected to a social media service, they obtain an unparalleled ability to connect other individual actors. To become a part of social media is to become a part of a collective, an assemblage [4]. The ability to connect is the essential dynamics of social media platforms. The platform mediates and makes connections; it is social in the sense of association [4]. However, social ties (such as a friend connection) that are established on different social media platforms (Facebook) are not always socially active in terms of direct communication, but they may have several different situation and context-dependent functions. In other words, social media is a broad collection of different types of social ties. Moreover, we argue that these ties contain different forms of potential that is possible to be actualized into different kinds of social actions and translated to different outcomes.

However, we are not sure whether it is necessary and accurate to discuss social media “friends” in terms of “social capital”, or even resources [5, 6]. At least, for this article, we need to clarify the concepts of a friend, capital and resource. The allegory of “friends as capital” is a slightly problematic analogy because the term capital refers to a stash, a stockpile, savings – something you can quietly go back to, something from which you fetch or draw and then consume. While social media can function as a repository of known friends, it, however, provides merely a possibility to activate a friendship, to re-establish the connection that does not guarantee successful activation by itself. For us, following the line from Georg Simmel [1], to Chicago School of interactionist [7], and Actor-Network theory [4], a friendship is made and enacted through interaction. There is no friendship unless there is an actual exchange, e.g., communication. In other words, a friendship is not a stable property but a constantly becoming process that is acted upon. What is a “friend” that is never contacted? An inert, dead piece in a stockpile? Perhaps it is a perishing acquaintance. Plausibly, it can be both. A hanging friend in a stockpile is in fact a virtual foundation of a friendship that offers a possibility of engaging interactive communication in which friendship becomes actual. Being a friend is an event, not a stubborn structure.

We begin with a premise that in an emerging connection energy flows. The emergence of connection is the initial step of actualization that is the movement of energy. No matter whether it was energy in the form of a message (a rush of electrons) or a thought (a chemical-electric event), a wave of the hand (physical movement), the connection is activated, actualized. A channel is opened. A bridge that connects two banks [8, p. 73]. However, the first act, the initial burst of energy, does not necessarily always lead to reciprocal interaction but the outcome and response is always open, not determinate. This uncertainty materializes in many different ways: a friend might have become different, and thus is not the same as in the past. All the processes are that of nature as they include a piece of unpredictability;  $p$  never reaches value 1. “Friend” is sought to be held together by a stabilizing system that is social media service but there is always nonknowledge and chaos involved, a possibility of transformation [8]. Summarized, between friends energy flows, yet it guarantees nothing but instability.

Regarding social media platforms, the interaction of friends is a necessity for their existence [9, pp. 155–160]. The silence of dried out services such as MySpace, Friendster, Orkut, Google+ forces us to pay attention to what the social in social media is. In simple



terms, in alignment with their business logic, the critical factor of being alive is the buzz of users who share, comment, and react. It is noise that users make. The social media platforms and their services also differ in terms of fulfilling users' social needs [10]. Social networking sites, such as Facebook, are merely based on users' self-expression, which is why socialization process often occurs through only following "others" and public commenting [11]. Instead, the messaging applications are more emphasizing on maintenance and development of relationships.

The social is based on activity, on interaction, on the uncertainty of outcome. Previous studies have indicated that the active and intensive use of social media is dependent on the intimacy of user's social interactions [12, 13]. However, users' activity in social media is not straightforwardly related to the intimacy in social relationship manifesting on social media [14]. In other words, activity – for example, messaging – does not always mean sharing sensitive information. Furthermore, the platforms are not only dependent on active social media users, but rather on the integrated and trustful social networks of their users. The platforms breathe through the communication that takes place on them by bridging the social networks of users [15]. Nonetheless, they provide nothing but space and channels for users (to shout out messages and reach other users), which should be - and seems to be - enough to attract more users.

The logic of social media is relatively simple: new users create more noise and call for other users. As the social sites become occupied, they start to hum the noise that is created by the users. The platforms are places of noise that organize, or more precisely that live of the organization of users' noise. Importantly, at this point, using turns around. The user no longer occupies the entitled position of a parasite - the one who merely enjoys services of the host (free space and channels) without paying [8] - but the user becomes a parasitized actor that is drained of information (including reactions) so that they can be analyzed, targeted, fed to algorithms [16]. Through this analysis, the social media service providers cease to be mere internet companies and become advertisement companies selling space for advertisement or selling profiles that is in fact organised noise.

In this sense, social media platforms require and live on the social ties of users to whom the platforms provide a possibility of establishing and maintaining social ties. However, platforms are not static structures on which social action would merely take place, but as mediators, they participate actively as a channel. The place of the channel, the social media platform, lies in between the communicators. It is the place messenger as of Hermes in the Ancient Greek [8]. In an academic discussion, social media platforms have been criticized with concepts of exploitation and alienation.

Initially applied by Marx, "exploitation means the process through which capitalists enrich themselves by selling commodities produced by workers and returning only part of the value of those commodities in wages" [17]. Thus, because users are generally paid little or nothing in return for the value they create for social media platforms, the rate of exploitation approaches infinity [18]. Alienation on the other hand roughly refers to the process whereby the worker is made to feel foreign to the products of own labor. At this moment, users acting on different social media platforms may not be aware that they are creating value with their social actions on these platforms [19]. In this sense, exploitation

and alienation of users can be comprehended as being in mutual dependence on social media platforms.

However, we argue that social media platforms role is not directly comparable with industrial 20th century's capitalist's position, as social media platforms are more dependent from their users. Moreover, it is trust that the sites have to build, and it is not the users that are dependent on their social media hosts. It goes the other way around. Users are hosts. The parasite social site lives attached to the body its victim. The parasite is literally carried around in the pockets of people. The parasite eats processing cycles, reserves memory, drains the battery.

Thus, given the fact that platforms cannot afford to lose their users, some scholars argue that on social media platforms users have to be more likely de-alienated to make the exploitation to happen [17, 19, 20]. Thus, the connection between exploitation and alienation in social media platforms can be understood as some trade-off where users are needed to be de-alienated so that they could implement themselves and gain their full potential in those networking sites [17, 20]. Primarily, by expressing themselves freely, users are producing better - more exciting and alluring - content in both quantitative and qualitative matter, which will eventually turn to more value for platforms. In simple terms, this means more users, more noise to be analyzed, more users to be used.

The social of the media is a link – a string – a social tie. However, in order to keep on the noise of users, social media platform cleverly hides itself. The fuzz is never about the platform itself - it is not the product that is in focus. The focus lies on other users. The channel itself should not get too much attention. As a fluently functioning channel, it becomes almost invisible. The platform came visible and problematized if it does not work or if there are privacy issues. Mark Zuckerberg in U.S. Congress hardly went unnoticed. Moreover, it was because Facebook did not function fluently but was used actively.

Nevertheless, the parasite is invited, and social media is used. Users host it and carry it around. We give it our attention. Admittedly, it provides a neat place for communication, a channel to reach masses. As social media platforms allure users, they provide various services, e.g. private messaging. In terms of actualizing friendship, private messaging is a useful tool. It is no doubt an actualization - a movement of energy - of friendship. Furthermore, what are the energy bursts - the messages - like? In terms of social, they can be sensitive or mundane, the former shrinking the social distance more than the latter. To share sensitive or confidential messages, the user must trust the channel. In other words, in order to make the social gap smaller, the audience of the message must be confined. The secrets that bring people together, bind the ties, are not for everyone but only for a carefully picked number of people. "For your eyes only" is a strong social statement. I want to share this with you. Love excludes. It is only you and me. It is a decision, it cuts. To include one is to exclude others [8, 21]. Thus, private messaging is required to be trusted regarding that confidential stays confidential.

These notions encapsulate the dynamics of social media platforms; visible social media networks constituted through accepted friend requests or phone numbers added to contact lists are not the bread and butter for platforms, but platforms stay live on social activities that take place on these platforms. Platform cherish by staying under the radar and feast with valuable data streams extracted from these interactions. If nothing

is exchanged between users, meaning that social does not materialize, then confirmed ties may stay visible but the platform dies. What would TripAdvisor be without user reviews [22]. The whole social media becomes though activity.

### 3 Method

#### 3.1 Data

Our data based on two different surveys collected at the same time in late 2017 and early 2018. The first part of data were sampled randomly 8000 aged 18 to 74 from the Finnish population register to the mail survey. With 31% response rate, the final data include 2470 observations. Secondly, in order to guarantee sufficient number of social media users, we improved the data by including 1,254 respondents (also aged 18 to 74) from the non-probability sample collected of a nationally representative online panel. Accordingly, the final data included 3,724 respondents, of whom 66% were from the probability sample [23].

In this study, we focused on social media users that accounted for 74% of the total data. We considered the potential sampling error related to two different data sources by providing a robustness check for the main effects. We also controlled for the demographic bias regarding the age with a weight variable constructed by calibrating the sample's age distribution to correspond with the official population distribution of Finnish social media users according to Official Statistics of Finland [24].

#### 3.2 Measures

Table 1 provides information on the measurements and descriptive statistics for all the variables used in the further analyses. As for the dependent variable, we used a variable elicited from the statement: "I can trust that social media platforms (such as Facebook) will not publicize my messages." The answer options ranged from 1 to 5 in which was given 1 completely disagree, 3 do not agree or disagree, and 5 completely agree. In this study, we recoded the variable as binary by combining categories 1–3 and 4–5 to predict those who agree (initial categories 4 and 5). In other words, we focus on those who can trust the social media platform in private messaging.

Our primary independent variables are the network size on social media and trust in networks on social media. We measured both variables from the same angle using the initial questions "To what extent do you have friends and acquaintances on social media?" and "To what extent do you trust your friends and acquaintances on social media?". The answer options for both questions ranged from 1 (not at all) to 5 (very much). These variables were used as continuous in further analyses.

We considered for a set of potential confounding variables related to the internet and social media behavior of participants. First, we took into account the general features of internet usage and privacy skills. The frequency of internet usage was measured by asking how often the respondents use internet by using 5-point scale: 1 (never), 2 (less than weekly), 3 (weekly), 4 (daily), 5 (many times per day). We asked privacy skills with three different statements after the question “To what extent following Internet actives describe you”. Given answers ranged from 1 (Not at all) to 5 (Very well). The initial statements were: 1) I know how to use private browsing settings online (For example, incognito mode) 2) I know how to delete my online browsing history 3) I know how to turn my location services on and off.

Second, to account for the effects of social media usage, we controlled for the participants’ social media skills, activity in using social message applications (SMA) and the preferred SMA. Initially, we measured activity by asking the frequency how often the respondents use SMAs with similar scale as in asking the frequency of internet usage. Social media skills were measured by following the validated measure of social skills [25]. We constructed the preferred SMA application by two questions initially asked how often respondents use Facebook and WhatsApp via the same scale from 1 (Never) to 5 (Many times per day). We combined the information of the variables for constructing a new variable to measure whether participants use either Facebook messenger (1), WhatsApp (2), or both of them (3).

We also controlled participants’ age, gender, and education throughout analyses. We measured age via an open-ended question in which the respondents reported their year of birth. Education was categorized as binary by differentiating those have achieved at least tertiary level from those having primary or secondary level education.

### 3.3 Techniques

In the first phase of the empirical study, we assessed the direct effects of that network size and trust in networks on the confidence in social media platform. We also conducted a robustness check for the sample effect by modeling separately for both the probability and nonprobability samples. We conducted the statistical tests by using logistic regression models and presented the results of main effects odds ratios.

We also estimated the indirect effects of independent variables through confounding variables with the KHB tests [26, 27]. The tests were conducted in a step-by-step manner by holding the sociodemographic variables as covariates in each model. We reported the indirect effects in the text as mediation percentages and logit-coefficients. We performed the analyses with Stata 15 by using the KHB -package and illustrated the results by utilizing the user-written packages of *coefplots* [28] and *graphic schemes* [29].

**Table 1.** Descriptive information of the applied measures

| Measures  | N    | M    | SD   | Min | Max |
|---|------|------|------|-----|-----|
| <u>Dependent</u>                                  |      |      |      |     |     |
| Platform confidence in private messaging (binary) | 2772 | 0.26 | 0.44 | 0   | 1   |
| <u>Independent</u>                                |      |      |      |     |     |
| <i>Social resources on social media</i>           |      |      |      |     |     |
| Social ties on social media                       | 2758 | 2.87 | 0.92 | 1   | 5   |
| Trust in social ties on social media              | 2748 | 3.42 | 0.95 | 1   | 5   |
| <u>Internet and social media behavior</u>         |      |      |      |     |     |
| <i>Frequency of usage*</i>                        |      |      |      |     |     |
| Use the Internet?                                 | 2765 | 4.34 | 0.6  | 1   | 5   |
| Use the instant messenger applications            | 2752 | 3.56 | 0.98 | 1   | 5   |
| <i>Skills</i>                                     |      |      |      |     |     |
| Privacy skills (sum variable, alpha = 0.81)       | 2765 | 3.76 | 1.13 | 1   | 5   |
| Social media skills (sum variable, alpha = 0.88)  | 2761 | 4.12 | 0.85 | 1   | 5   |
| <i>Application**</i>                              |      |      |      |     |     |
| Facebook messenger                                | 2760 | 0.08 | 0.27 | 0   | 1   |
| WhatsApp  | 2760 | 0.34 | 0.47 | 0   | 1   |
| Both  | 2760 | 0.58 | 0.49 | 0   | 1   |
| <u>Demographics</u>                               |      |      |      |     |     |
| <i>Gender</i>                                     | 2766 | 0.52 | 0.5  | 0   | 1   |
| <i>Age</i>  | 2767 | 47.9 | 15.8 | 18  | 75  |
| <i>High education</i>                             | 2726 | 0.59 | 0.49 | 0   | 1   |

\*How often do you do the following?

(1 = Never, 2 = Less than weekly, 3 = Weekly, 4 = Daily, 5 = Many times per day)

\*\*How often do you use the following social media services?

(coded 1 = Facebook Messenger, 2 = WhatsApp, 3 = Both)

## 4 Results

Table 2 shows the results for the first hypothesis. Social ties had a substantial effect on the confident in social media platform (OR = 1.48,  $p < .01$ ). Trust in networks also positively predicted high confidence in the used platform (OR = 1.60,  $p < .01$ ).

According to the results of robustness check, social ties had similar effects on the probability sample ( $OR = 1.44, p < .01$ ) and the nonprobability sample ( $OR = 1.58, p < .01$ ). The effect of trust in networks varies slightly between the probability sample ( $OR = 1.68, p < .01$ ) and the non-probability sample ( $OR = 1.47, p < .01$ ). However, according to the interaction analysis, the differences were not statistically significant in either case after the equating of demographic variables.

Next, we added the first set of behavioral variables to the base model. As seen, internet privacy skills and frequency of internet, skills did not affect the association between interest variables and the confidence in social media platform. In addition, the covariate effects of those variables were weak.

Afterward, we complemented the model with the second set of behavioral variables related to social media usage. We found that social media usage skills and frequency of SMAs usage have a positive effect on the confidence in social media platform. Interestingly, internet privacy skills had also negative effect on platform confidence. Finally, those preferred only WhatsApp as SMAs application reported being less likely confident with the platform they used in social messaging.

After conventional logistic regression, we conducted a more in-depth analysis of confounding effects of covariate variables. The results are shown in Table 3. We found that neither the frequency of internet usage nor the privacy skills have confounding effects on the main associations. Instead, the effects of social media usage influenced significantly on both associations. The frequency of SMAs usage confounded 13% of the effect of network size ( $b = 0.05, p < 0.01$ ) and 9% of the effect of trust in networks ( $b = 0.04, p < 0.01$ ). Social media skills also confounded the effects of main independent variables; 13% of the effect of network size ( $b = 0.05, p < 0.01$ ) and 11% of trust in networks ( $b = 0.05, p < 0.01$ ). Simultaneously, the preferred application found to be a significant factor, when it confounded over 30% ( $b = 0.14, p < 0.01$ ) of the effect of network size and 15% of the effect of trust in networks ( $b = 0.08, p < 0.01$ ).

Figure 1 illustrates the main results of predictive analyses. We may suggest that the size of social media networks and trust in social media networks positively contribute to the platform confidence of social media users in social messaging. Additionally, we propose that trust in social networks is more effective factor, and not as dependent on the other behavioral factors, as compared to the size of social media networks.

**Table 2.** Predicting the confidence in social media platform according to the size of social media networks, trust in social ties and covariate variables, Odds ratios and standard errors with statistical significances

| Dependent variable                                |                  |                  |                  |                  |                  |                  |
|---|------------------|------------------|------------------|------------------|------------------|------------------|
| Confidence in social media platform               | M1               | M2               | M3               | M1               | M2               | M3               |
| Social ties on social media                       | 1.48**<br>(0.07) | 1.48**<br>(0.07) | 1.25**<br>(0.07) |                  |                  |                  |
| Trust in social ties on social media              |                  |                  |                  | 1.60**<br>(0.08) | 1.59**<br>(0.08) | 1.46**<br>(0.08) |
| Privacy skills                                    |                  | 1.01<br>(0.06)   | 0.86*<br>(0.06)  |                  | 1.02 (0.07)      | 0.85*<br>(0.06)  |
| Frequency of Internet usage                       |                  | 1.09<br>(0.1)    | 0.98<br>(0.09)   |                  | 1.10 (0.10)      | 0.98<br>(0.09)   |
| Social media skills                               |                  |                  | 1.33**<br>(0.09) |                  |                  | 1.33**<br>(0.09) |
| Frequency of SMAs usage                           |                  |                  | 1.15*<br>(0.07)  |                  |                  | 1.16*<br>(0.08)  |
| Preferred application: (ref = Facebook messenger) |                  |                  |                  |                  |                  |                  |
| WhatsApp  |                  |                  | 0.49**<br>(0.1)  |                  |                  | 0.47**<br>(0.09) |
| Both  |                  |                  | 1.05<br>(0.2)    |                  |                  | 1.08<br>(0.2)    |
| Observations                                      | 2,707            | 2,693            | 2,657            | 2,698            | 2,685            | 2,649            |

Odds Ratios, Standard errors in parenthese  
 Models control for age, gender and education  
 \*\*p < 0.01, \*p < 0.05

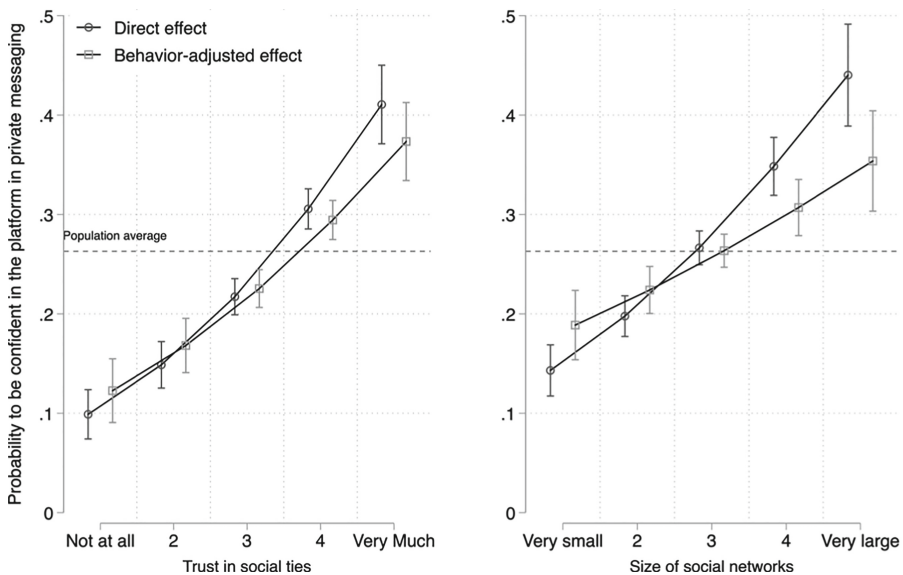
**Table 3.** Effects of social ties and trust in social ties on social media via confounders on the confidence in social media platform, decomposed (KHB) logit coefficients and standard errors.

| Dependent variable                       | Interest variable           |         | Interest variable                    |         |
|--|-----------------------------|---------|--------------------------------------|---------|
|  | Social ties on social media |         | Trust in social ties on social media |         |
| Platform confidence in private messaging | B                           | SE      | B                                    | SE      |
| Interest variable's effect via           |                             |         |                                      |         |
| Internet privacy skills                  | 0.002                       | (0.006) | 0.003                                | (0.005) |
| Frequency of Internet usage              | 0.005                       | (0.005) | 0.005                                | (0.005) |
| Social media skills                      | 0.05**                      | (0.013) | 0.05**                               | (0.011) |
| Preferred application                    | 0.14**                      | (0.022) | 0.08**                               | (0.013) |
| Observations                             | 2,707                       |         | 2,693                                |         |

Logit coefficients (B), Standard errors (SE) in parentheses

Models control for age, gender and education

\*\*p < 0.01, \*p < 0.05



**Fig. 1.** Probability to be confident in the social media platform in case of private messaging according to the size of social media networks (social ties) and the level of trust in social media networks (trust in social ties). Predicted probabilities with confidence intervals.



## 5 Discussion

First, the results of the analysis support the hypothesis that the high level of confidence in platforms is strongly dependent on how engaged users are to their social networks. The size of network (i.e. the number of social media friends) and trust in the networks were significant variables when we analyzed the confidence in social media platforms as mediators of private messages. Secondly, the results show that trust in social media contacts – friends – has independent explanation power in the platform confidence. Similarly, behavioral elements of engagement, namely frequency of social media platform usage and social media skills, had a positive correlation to the level of confidence in platforms. Additionally, while some may argue that users engaged to these platforms because of high confidence to the platform, the effect is much higher when trust in platforms is used as an independent variable.

The trust in friends seems to go beyond the trust in the platform. In theoretical terms, the importance of actualization itself neutralizes doubts towards the mediating platform. Seizing into the potential of a social tie is always giving up of some privacy. In this sense, technology in between loses its significance if it can deliver the messages seamlessly, in other words, if the platform functions appropriately. We argue, that the trust that is placed in the social media friends and the desire to contact the friends provides such a focus that it blurs the privacy doubts relating to technology. Users' desire to make noise - communicate - suits the mediating social platforms. Every single message brings about data which can be analyzed and thus profited. To be able to use the users, the platforms require data - noise - as much as possible. Thus, the production of data is encouraged. All the platforms are thus inclusive.

However, we should ask who and what is excluded. In a sense, the platforms themselves exclude nothing but people who are unable to use the platform and who are not willing to agree with the terms of use. It is not merely users who can be excluded, but the platforms and some of their features can be excluded from using. As our analysis shows, if the users do not trust in their contacts they exclude the private messaging service or use it more cautiously. The cautiousness with social media friends spreads to the platform as well.

Our analysis also showed that those who had a comprehensive set of skills related to privacy issues and are in that sense aware of possible outcomes, do not trust platforms in such extent than those who are not aware of these issues. This notion underlines that when privacy concerns come visible platform's parasitic role reveals itself. When the platform's role as the active subject is not noticed, users do not feel constraints coming from the third party for their social interaction. Invisibility and insensibility are those features which make both platform and parasite effective; if you do not notice the existence of it, you do not make any actions aiming to dispose of it.

Similarly, trust in platforms associates with platform preferences. Those who favor Facebook Messenger over WhatsApp had significantly higher trust in platforms. Thus, platforms' differences in features related to privacy issues may explain this substantial difference. Since the year 2016 all messages sent via WhatsApp are encrypted end-to-end so that third parties - including WhatsApp itself - are not able to see content [30]. In Facebook Messenger, end-to-end encryption (named as "Secret Conversation") is an

optional complement, which user need switch on in every single time when she wants the conversation to be encrypted [31].

As the same company, Facebook Inc. own both of these platforms, it shows how privacy protocols are utilized to lure different segments of users. WhatsApp is bringing privacy issues on the table and is prominently detaching its marketing strategy from the user-generated content (but not from data concerning users networks, ties or frequency of social interaction). Facebook Messenger, on the other hand, is more actively trying to diminish the visibility of privacy issues by for example burying privacy settings in the complex maze of web page architecture. On Facebook Messenger it is possible to encrypt content; to take such action indicates that a privacy issue - the lurking parasite - is recognized. For those are not aware of privacy issues, the presence of a parasite stays hidden.

Our study has some limitations related to the survey period and the interpretation of causality. We conducted the survey before the biggest frenzies in privacy issues faced by Facebook and Google in 2018. By using the applied cross-sectional dataset, we could not validate the mechanism between social media resources and the confidence in the social media platform. This study, however, opened an avenue for further research to focus on how privacy concerns about Facebook and Google have contributed to the association between trust dimensions. We may, for example, ask if it is possible, that trust in social media networks maintain trust in the platform even though the platform has publicly been affected by confidence issues. This hypothesis needs additional development of the applied method by monitoring the variation of participants' trust with longitudinal panel data.

## References

1. Simmel, G.: *The Sociology of Secrecy and of Secret Societies*. Books on Demand (1906)
2. Pew Research Center: *Social Media Use Continues to Rise in Developing Countries, but Plateaus Across Developed Ones* (2018)
3. Tetri, P., Vuorinen, J.: Dissecting social engineering. *Behav. Inf. Technol.* **32**(10), 1014–1023 (2013)
4. Latour, B.: *Reassembling the Social: An Introduction to Actor-Network-Theory*. OUP, Oxford (2005)
5. Chambers, D.: Networked intimacy: algorithmic friendship and scalable sociality. *Eur. J. Commun.* **32**(1), 26–36 (2017)
6. Chambers, D.: Self-presentation online. In: *Social Media and Personal Relationships*, pp. 61–81 (2013)
7. Garfinkel, H.: *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs (1967)
8. Serres, M.: The Parasite. *Clin. Dermatol.* **10**(1), 255 (2007)
9. van Dijck, J.: *The Culture of Connectivity: A Critical History of Social Media*. Oxford University Press, Oxford (2013)
10. Quan-Haase, A., Young, A.L.: Uses and gratifications of social media: a comparison of Facebook and instant messaging. *Bull. Sci. Technol. Soc.* **30**(5), 350–361 (2010)
11. van Dijck, J.: 'You have one identity': performing the self on Facebook and LinkedIn. *Media Cult. Soc.* **35**(2), 199–215 (2013)
12. Ellison, N.B., Steinfield, C., Lampe, C.: Connection strategies: social capital implications of Facebook-enabled communication practices. *New Media Soc.* **13**(6), 873–892 (2011)

13. Hsu, C.-W., Wang, C.-C., Tai, Y.-T.: The closer the relationship, the more the interaction on Facebook? Investigating the case of taiwan users. *Cyberpsychol. Behav. Soc. Netw.* **14**(7–8), 473–476 (2011). <https://doi.org/10.1089/cyber.2010.0267>
14. Sutcliffe, A.G., Binder, J.F., Dunbar, R.I.M.: Activity in social media and intimacy in social relationships. *Comput. Hum. Behav.* **85**, 227–235 (2018)
15. Ellison, N.B., Steinfield, C., Lampe, C.: The benefits of Facebook ‘friends:’ social capital and college students’ use of online social network sites. *J. Comput. Commun.* **12**(4), 1143–1168 (2007)
16. Introna, L.D.: Algorithms, governance, and governmentality: on governing academic writing. *Sci. Technol. Hum. Values* **41**(1), 17–49 (2015)
17. Rey, P.J.: Alienation, exploitation, and social media. *Am. Behav. Sci.* **56**(4), 399–420 (2012)
18. Fuchs, C.: Labor in informational capitalism and on the internet. *Inf. Soc.* **26**(3), 179–196 (2010)
19. Reveley, J.: Understanding social media use as alienation: a review and critique. *E-Learning Digit. Media* **10**(1), 83–94 (2013)
20. Fisher, E.: How less alienation creates more exploitation? Audience labour on social network sites. *TripleC* **10**(2), 171–183 (2012)
21. Pyyhtinen, O.: *The Gift and its Paradoxes: Beyond Mauss*. Ashgate Publishing Ltd. (2014)
22. Orlikowski, W.J., Scott, S.V.: What happens when evaluation goes online? Exploring apparatuses of valuation in the travel sector. *Organ. Sci.* **25**(3), 868–891 (2013)
23. Sivonen, J., Koivula, A., Saarinen, A., Keipi, T.: Research report on the Finland in the digital age - survey. University of Turku, Department of Social Research, Turku (2018)
24. Official Statistics of Finland: use of information and communications technology by individuals (2016)
25. van Deursen, A.J.A.M., Helsper, E.J., Eynon, R.: Development and validation of the Internet Skills Scale (ISS). *Inf. Commun. Soc.* **19**(6), 804–823 (2016)
26. Breen, R., Karlson, K.B., Holm, A.: Total, direct, and indirect effects in logit and probit models. *Sociol. Methods Res.* **42**(2), 164–191 (2013)
27. Karlson, K.B., Holm, A., Breen, R.: Comparing regression coefficients between same-sample nested models using logit and probit: a new method. *Sociol. Methodol.* **42**(1), 286–313 (2012)
28. Jann, B.: Plotting regression coefficients and other estimates. *Stata J.* **14**(4), 708–737 (2014)
29. Bischof, D.: New graphic schemes for Stata: plotplain and plottig. *Stata J.* **17**(3), 748–759 (2017)
30. WhatsApp Inc.: WhatsApp Security. Privacy & Terms protocol (2018). <https://www.whatsapp.com/security/>
31. Facebook Inc.: Help Center (2018). <https://www.facebook.com/help/>

## Author Index

- Åberg, Erica II-16  
Abidin, Zaenal II-3  
Aciar, Gabriela II-147  
Aciar, Silvana II-147, II-158, II-236  
Ahangama, Supunmali II-616  
Ahmed, Wasim II-447  
Alfaro, Rodrigo I-624  
Alharthy, Faowzia II-459  
Allende-Cid, Héctor I-624  
Ancán, Oscar II-483  
Ando, Masaya II-549  
Andreadis, Alessandro I-289  
Androniceanu, Armenia II-261  
Ardelean, Andreea I-442  
Asa, Yasuhiro II-247  
Assenmacher, Dennis I-201, II-71  
Avvenuti, Marco I-376
- Balmaceda Castro, Iván II-158, II-236  
Bascur, Camila II-171  
Bath, Peter A. II-447  
Becker, Jörg II-71  
Beier, Michael I-3  
Belavadi, Poornima I-215  
Berthelot-Guiet, Karine II-278  
Bian, Yuanyuan II-471  
Blekanov, Ivan S. I-19, I-433  
Blockside, Jade I-614  
Bodrunova, Svetlana S. I-19, I-433  
Bookhultz, Shane I-243  
Botella, Federico II-205, II-236  
Bracci, Margherita I-289  
Brunk, Jens II-71  
Burbach, Laura I-215  
Burkhardt, Steffen I-568
- Calderon Maureira, Juan Felipe I-110  
Calero Valdez, André I-27  
Callegaro, Mario I-130  
Cano, Sandra II-181, II-236  
Cao, Xueni I-45, I-168, II-295  
Cardenas, Luis II-501
- Carvajal, Victoria II-226  
Chang, Teng-Wen II-471  
Chen, Sijing I-407  
Cheng, Mei-I I-614  
Chou, Huichen II-29  
Clever, Lena I-201  
Coman, Adela I-442, I-583  
Coto, Mayela II-147  
Couper, Mick P. I-130  
Cresci, Stefano I-376
- Danilova, Yulia I-433  
De Choudhury, Munmun II-87  
Delgado, Dania II-193  
Demant, Jakob I-278  
Demartini, Gianluca II-447  
Díaz, Jaime II-483  
Dremel, Anita I-459  
Dudley, Alfreda I-419, II-459  
Dutta, Sarmistha II-87
- Elortegui, Claudio I-624  
Ernala, Sindhu Kiranmai II-87
- Fang, Ying I-45, I-168, II-295  
Farhan Mardadi, M. II-3  
Feldvari, Kristina I-459  
Fernández-Robin, Cristóbal II-501  
Fernández Valdés, Gregorio II-355  
Fietkiewicz, Kaja J. I-227  
Früh, Sebastian I-3  
Fujino, Hidenori II-511  
Fujioka, Takuya II-247
- Galkute, Milda I-56  
Gang, Yu II-559  
Garcia, David I-542  
Garrido, Eduardo II-598  
Georgescu, Irina II-261  
Gomez-Mejia, Gustavo II-309  
Gomez-Vasquez, Lina II-526  
Gonzalez, Cristian I-624

- Grigore, Ana-Maria I-442  
 Grimme, Christian I-201  
 Gromova, Tamara I-647  
 Guidi, Stefano I-289
- Halbach, Patrick I-215  
 Hamam, Doaa II-538  
 Hardjasa, Louisa I-475  
 Hawdon, James I-243  
 Hidayanto, Achmad Nizar II-3  
 Hoang, Minh-Duc II-44  
 Huang, Hung-Hsuan I-489  
 Hysaj, Ajrina II-538
- Ijuin, Koki I-504, I-658  
 Ilhan, Aylin I-513  
 Illandara, Kanishka H. II-616  
 Inan, Dedi I. II-3  
 Ishida, Toru II-29  
 Ishii, Hirotake II-247, II-431  
 Ito, Ayaka II-549  
 Ito, Kyoko I-72, II-247  
 Izumi, Tomoko I-81, II-559
- Jamet, Erick II-598  
 Jokinen, Kristiina I-504  
 Jones, Brian M. I-534
- Kaakinen, Markus I-278, I-542  
 Kadigamuwa, A. S. T. M. R. D. S. II-616  
 Katagiri, Yuho I-551  
 Kato, Tsuneo I-658  
 Kato, Yuko II-511  
 Keipi, Teo II-569  
 Kessling, Philipp I-568  
 Kiessling, Bastian I-568  
 Kinnunen, Jani II-261  
 Kishi, Yasutaka I-72  
 Kitajima, Yuzuki II-325  
 Kitamura, Takayoshi I-81, II-559  
 Kobashi, Honoka II-511  
 Koh, Karen Wei Ling I-600  
 Koiranen, Ilkka I-669, II-569  
 Koivula, Aki I-669, II-16, II-569  
 Koltsova, Olessia I-261
- Konishi, Tetsuma I-81  
 Kröger, Niclas I-95  
 Kukkonen, Iida II-16  
 Kurushima, Takashi II-431
- Le, Linh II-44  
 Le, Trang II-44  
 Lee, Joon Suk I-349  
 Leman, Scotland I-243  
 Leoveanu, Valentin Mihai I-583  
 Li, Xiaodong I-45, I-168, II-295  
 Lim Ding Feng, Ethan I-600  
 Lin, Donghui II-29  
 Liu, Changqing II-57  
 Lu, Xing II-57  
 Lu, Zhicong II-57  
 Luthfia Fitriani, Amira II-3
- Maracine, Robert I-442  
 Marchigiani, Enrica I-289  
 McCoy, Scott II-501  
 Mendoza, Marcelo I-321  
 Miller, Bryan Lee I-278  
 Miltsov, Alexandre I-305  
 Mitra, Tanushree I-243  
 Miyake, Shin II-336  
 Miyano, Nao II-511  
 Mochizuki, Rika II-431  
 Montero-Liberona, Claudia II-355  
 Morales, Jenny II-205, II-236  
 Muñoz, Martin II-226  
 Murai, Yuichi II-549
- Nakatani, Yoshio I-81, II-559  
 Nakayama, Johannes I-215  
 Nakazawa, Atsushi I-475  
 Namatame, Takashi I-551, II-325, II-336,  
 II-374, II-389  
 Neo, Zhi-Wei I-600  
 Nguyen, Anh-Tuan II-44  
 Nguyen, Hoang D. I-600, II-44  
 Nguyen, Thi-Thanh I-600  
 Niemann, Marco II-71  
 Nigmatullina, Kamilla I-433  
 Nishida, Shogo I-72

- Nishida, Toyooki I-489  
 Nonaka, Mei II-374  
 Numata, Takashi II-247
- Oksanen, Atte I-278, I-542  
 Okunari, Taiga II-511  
 Orellana Quiñones, Mayron I-110  
 Otake, Kohei I-551, II-325, II-336, II-374,  
 II-389
- Pajunen, Tero II-16  
 Palma, Wenceslao I-624  
 Parlàngeli, Oronzo I-289  
 Pimentel Varas, Gianluigi II-355  
 Platonov, Konstantin I-647  
 Plettenberg, Nils I-215  
 Pohl, Janina Susanne I-201  
 Porshnev, Alexander I-305  
 Preuss, Mike I-359  
 Prince, Emma II-584  
 Providel, Eliana I-321  
 Puente, Anibal I-110
- Quiñones, Daniela II-171, II-181, II-193,  
 II-205, II-217, II-226, II-236, II-598
- Rai, Roshan I-614  
 Ranganathan, Shyam I-243  
 Rapp, Maximilian I-95  
 Räsänen, Pekka II-569  
 Riehle, Dennis M. II-71  
 Rodriguez, Sebastian I-624  
 Rojas P., Luis A. I-56, I-110  
 Rojas, Luis II-217  
 Romero-Hall, Enilda II-526  
 Roncagliolo, Silvana II-193, II-226  
 Rusu, Cristian II-158, II-171, II-181, II-193,  
 II-205, II-217, II-226, II-236, II-598  
 Rusu, Virginia Zaraza II-598  
 Rusu, Virginia II-193, II-226, II-236
- Sagal M., Victor A. I-56  
 Saha, Koustuv II-87  
 Saito, Retsuya II-389  
 Sakamoto, Yoshiki II-247, II-431  
 Sandu, Mihaela Cornelia I-583  
 Santander, Pedro I-624  
 Sarpila, Outi II-16
- Savolainen, Iina I-278, I-542  
 Sbaffi, Laura II-447  
 Scheerer, Samira I-95  
 Scheibe, Katrin I-184, I-636  
 Scheiner, Christian W. I-335  
 Selkälä, Arto I-130  
 Shalihah, Rizkah II-3  
 Sharkova, Ekaterina I-647  
 Sharma, Eva II-87  
 Sharmin, Farzana II-401, II-415  
 Shimoda, Hiroshi II-247, II-431  
 Shutsko, Aliaksandra II-108  
 Sim, Kellie I-600  
 Simmons, Mariah I-349  
 Sinyavskaya, Yadviga I-261  
 Sirola, Anu I-278, I-542  
 Smoliarova, Anna I-19, I-433, I-647  
 Stanarević Katavić, Snježana I-459  
 Stöcker, Christian I-359, I-568  
 Sublime, Jérémie I-156  
 Sultan, Mohammad Tipu II-401, II-415  
 Sun, Qinghua II-247
- Takashima, Yuki II-431  
 Takemoto, Muneo II-549  
 Tan, Hong-Ray I-600  
 Tarasov, Nikita I-19  
 Tardelli, Serena I-376  
 Tarutani, Tomoya II-511  
 Terpilovskii, Maxim I-261  
 Tesconi, Maurizio I-376  
 Trautmann, Heike I-201, II-71  
 Truyol, Maria Elena I-110  
 Turner, Mark II-584
- Uchida, Hitoshi II-549  
 Ueda, Kimi II-247, II-431  
 Umata, Ichiro I-658  
 Uotani, Takumi II-431  
 Urs, Bogdan Alexandru II-236  
 Urs, Ilie II-236
- Valdez, André Calero I-215  
 Valencia, Katherine II-598  
 Vuorinen, Jukka I-669
- Wang, Ruiquan II-126  
 Wang, Wenru I-600  
 Wang, Yuanqiong II-459

- Watanabe, Masahiro [II-431](#)  
Weerasinghe, Supuni N. [II-616](#)  
Wen, Peihan [II-126](#)  
Wilbanks, Linda R. [I-393](#)  
William De Silva, Aaron [I-600](#)  
Wishwanath, Champika H. P. D. [II-616](#)  
  
Xiao, Lu [I-407](#)  
  
Yagi, Soyoka [II-511](#)  
Yamamoto, Seiichi [I-658](#)  
Yamashita, Naomi [II-29](#)  
Yamawaki, Mizuki [II-247](#)  
Yáñez, Diego [II-501](#)  
  
Zahrah Halim, Atikah [II-3](#)  
Zaleppa, Paige [I-419](#)  
Zambon, Riccardo [I-289](#)  
Zamora, Daniela [II-193](#)  
Zamora, Juan [I-156](#)  
Zhang, Hantian [I-184](#)  
Zhang, Liqun [I-45, I-168, II-295](#)  
Zhu, Liyu [I-45, I-168, II-295](#)  
Zhuravleva, Nina [I-19, I-433](#)  
Ziefle, Martina [I-215](#)  
Zimmer, Franziska [I-184, I-636](#)  
Zúñiga, Constanza [II-598](#)  
Zych, Izabela [I-278](#)