Robert Thomson · Aravind Hariharan ·
Scott Renshaw · Samer Al-khateeb ·
Annetta Burger · Patrick Park ·
Aryn Pyke (Eds.)

LNCS 14972

Social, Cultural, and Behavioral Modeling

17th International Conference, SBP-BRiMS 2024
Pittsburgh, PA, USA, September 18–20, 2024
Proceedings

Springer

Lecture Notes in Computer Science 14972

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Robert Thomson · Aravind Hariharan ·
Scott Renshaw · Samer Al-khateeb ·
Annetta Burger · Patrick Park · Aryn Pyke
Editors

Social, Cultural, and Behavioral Modeling

17th International Conference, SBP-BRiMS 2024
Pittsburgh, PA, USA, September 18–20, 2024
Proceedings

Springer

Editors
Robert Thomson
United States Military Academy
Army Cyber Institute
West Point, NY, USA

Scott Renshaw
Carnegie Mellon University
Pittsburgh, PA, USA

Annetta Burger
Oak Ridge National Laboratory
Oak Ridge, TN, USA

Aryn Pyke
United States Military Academy
Army Cyber Institute
West Point, NY, USA

Aravind Hariharan
The University of the Cumberlands
Williamsburg, KY, USA

Samer Al-khateeb
Creighton University
Omaha, NE, USA

Patrick Park
Carnegie Mellon University
Pittsburg, PA, USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-72240-0 ISBN 978-3-031-72241-7 (eBook)
https://doi.org/10.1007/978-3-031-72241-7

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

Preface

In this 17th year of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, SBP-BRiMS 2024, we highlight the many advances in the computational social sciences. Improving the human condition requires understanding, forecasting, and impacting socio-cultural behavior in both the digital and non-digital world. Increasing amounts of digital data, embedded sensors that collect human information, rapidly changing communication media, changes in legislation concerning digital rights and privacy, spread of 4G technology to developing countries and development of 5G technology, the rise of large language models (e.g., ChatGPT), and other changes are creating a new cyber-mediated world in which the very precepts of why, when, and how people interact and make decisions is being called into question. For example, Uber understood human behaviors vis-à-vis commuting. It then developed software to support this behavior, which ended up saving time (and so capital) and reducing stress, which indirectly created the opportunity for people to evolve new behaviors.

Scientific and industrial pioneers in this area are relying on both social science and computer science to help make sense of the impact of this new frontier. To be successful pioneers, a true merger of social science and computer science is needed. Solutions that rely only on the social sciences or computer science are doomed to failure. For example, Anonymous developed an approach for identifying members of terror groups, such as ISIS, on the X (formerly Twitter) social media platform using state-of-the-art computational techniques. These accounts were then suspended. This was a purely technical solution. The consequence was that those individuals with suspended X accounts just moved to new platforms and resurfaced on X under new IDs. In this case, failure to understand basic social behavior resulted in an ineffective solution.

The goal of this conference is to build this new community of social cyber scholars by bringing together and fostering interaction between members of the scientific, corporate, government, and military communities who are interested in understanding, forecasting, and impacting human socio-cultural behavior. It is the mission of this community to build this new field, its theories, methods, and scientific culture in a way that does not give priority to either social science or computer science, and to embrace change as the cornerstone of the community. Despite decades of work in this area, this scientific field is still in its infancy. To meet this charge, to move this science to the next level, this community must meet the following three challenges: deep understanding, socio-cognitive reasoning, and reusable computational technology. Fortunately, as the papers in this volume illustrate, this community is poised to answer these challenges. But what does meeting these challenges entail?

Deep understanding refers to the ability to make operational decisions and theoretical arguments based on an empirically based deep and broad understanding of the complex socio-cultural phenomena of interest. Today, although more data is available digitally than ever before, we are still plagued by anecdotally based arguments. For example, in social media, despite the wealth of information available, most analysts focus on small samples, which are typically biased and cover only a small time period, to explain all events and make future predictions. The analyst finds a magic tweet or an unusual tweeter and uses that to prove their point. Tools that can provide more data or less biased data are not widely used and are often more complex or time-consuming than what the average analyst would use for generating results. Not only are more scalable technologies needed, but also a better understanding of the biases in the data and ways to overcome them, as well as a cultural change to not accept anecdotes as evidence.

Socio-cognitive reasoning refers to the ability of individuals to make sense of the world and to interact with it in terms of groups and not just individuals. Currently, most social-behavioral models either focus on (1) strong cognitive models of individuals engaged in tasks that produce a small number of agents with high levels of cognitive accuracy but with little if any social context, or (2) light cognitive models and strong interaction models, which results in a model depicting massive numbers of agents with high levels of social realism and little cognitive realism. In both cases, as realism is increased in the other dimension, the scalability of the models fails, while their predictive accuracy on one of the two dimensions remains low. In contrast, as agent models are built where the agents are not just cognitive but socially cognitive, we find that the scalability increases and the predictive accuracy increases. Not only are agent models with socio-cognitive reasoning capabilities needed, but so too is a better understanding of how individuals form and use these social cognitions.

More software solutions that support behavioral representation, modeling, data collection, bias identification, analysis, and visualization are available to support human socio-cultural behavioral modeling and prediction than ever before. However, this software is generally just piling up in giant black holes on the web. Part of the problem is the fallacy of open source, that is, the idea that if you just make code open source, others will use it. In contrast, most of the tools and methods available in Git or R are only used by the developer, if at all. Reasons for its unpopularity with analysts include the lack of documentation, interfaces, and interoperability with other tools, difficulty of linking to data, and increased demands on the analyst's time due to a lack of tool-chain and workflow optimization. A part of the problem is the "not invented here" syndrome. For social scientists and computer scientists alike, it is just more fun to build a quick and dirty tool for your own use than to study and learn tools built by others. Another issue is the insensitivity of people from one scientific or corporate culture to the reward and demand structures of the other cultures that impact what information can or should be shared and when. A related problem is double standards in sharing: universities are expected to share and companies are not, but increasingly universities are relying on intellectual property as a source of funding just like other companies. While common standards and representations would help, a cultural shift from a focus on sharing to a focus on re-use is as or more critical for moving this area to the next scientific level.

In this volume, and in all the work presented at the SBP-BRiMS 2024 conference, you will see suggestions of how to address the challenges just described. The SBP-BRiMS 2024 conference carried on the scholarly tradition of the past conferences out of which it emerged like a phoenix: the Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP) Conference and the Behavioral Representation in Modeling and Simulation (BRiMS) Society's conference.

A total of 54 papers were submitted as regular track submissions. Of these, 24 were accepted to these proceedings, resulting in an acceptance rate of 44%. All papers were single-blind reviewed by 2 to 3 non-chair reviewers. Additionally, any papers from conference committee members were independently reviewed by 2–3 non-committee members. Furthermore, committee members' students were excluded from reviewing their papers.

Overall, the accepted papers come from an international group with papers submitted with authors from many countries, showing the broad reach of the computational social sciences community. The conference has a strong multi-disciplinary heritage. As the papers in this volume show, people, theories, methods, and data from a wide number of disciplines are represented including computer science, psychology, sociology, communication science, public health, bioinformatics, political science, and organizational science. Numerous types of computational methods are used including, but not limited to, machine learning, large language models, language technology, social network analysis and visualization, agent-based simulation, and statistics.

This exciting program could not have been put together without the hard work of several dedicated and forward-thinking researchers serving as the organizing committee, listed on the following pages. Members of the Program Committee, the scholarship committee, publication, advertising, and local arrangements chairs worked tirelessly to put together this event. They were supported by the government sponsors, the area chairs, and the reviewers. Please join me in thanking them for their efforts on behalf of the community. In addition, we gratefully acknowledge the support of our sponsor – the Army Research Office (W911NF-21-1-0102). We hope that you enjoyed the conference and welcome to the community.

September 2024

Robert Thomson
Scott Renshaw
Aravind Hariharan
Samer Al-khateeb
Annetta Burger
Patrick Park
Aryn Pyke

Organization

General Chairs

Kathleen M. Carley	Carnegie Mellon University, USA
Nitin Agarwal	University of Arkansas at Little Rock, USA

Program Committee Chairs

Robert Thomson	United States Military Academy, USA
Samer Al-khateeb	Creighton University, USA
Annetta Burger	Oak Ridge National Laboratory, USA
Patrick Park	Carnegie Mellon University, USA
Aryn A. Pyke	United States Military Academy, USA

Proceedings Co-chairs

Robert Thomson	United States Military Academy, USA
Scott Renshaw	Carnegie Mellon University, USA
Aravind Hariharan	University of the Cumberlands, USA

Agenda Co-chairs

Robert Thomson	United States Military Academy, USA
Kathleen M. Carley	Carnegie Mellon University, USA

Journal Special Issue Chair

Kathleen M. Carley	Carnegie Mellon University, USA

Tutorial Chair

Kathleen M. Carley	Carnegie Mellon University, USA

Publicity Chair

Donald Adjeroh	West Virginia University, USA

Graduate Program Chair

Stephen Dipple	Carnegie Mellon Univetssity, USA

Sponsorship Chair

Nitin Agarwal	University of Arkansas at Little Rock, USA

Travel/Participation Scholarship Chair

Nitin Agarwal	University of Arkansas at Little Rock, USA

Web Chair

Nitin Agarwal	University of Arkansas at Little Rock, USA

Challenge Problem Committee

Samer Al-khateeb	Creighton University, USA
Annetta Burger	Oak Ridge National Laboratory, USA
Patrick Park	Carnegie Mellon University, USA
Scott Renshaw	Carnegie Mellon University, USA
Aravind Hariharan	University of the Cumberlands, USA

Emeritus Steering Committee

Fahmida Chowdhury	National Science Foundation, USA
Rebecca Goolsby	Office of Naval Research, USA
Paul Tandy	Defense Threat Reduction Agency, USA
Edward T. Palazzolo	Army Research Office, USA
Patricia Mabry	Indiana University, USA

John Lavery	Army Research Office, USA
Tisha Wiley	National Institutes of Health, USA

Additional Reviewers

Neelofar Aleem
Walaa Alnasser
Sultan Alsarra
Amirreza Bagherzadehkhorasani
Aaditya Bhatia
Amrita Bhattacharjee
Sayantan Bhattacharya
Andrew Crooks
Stephen Dipple
Alexander Hough
Samuel Johnson
Waquar Kaleem
Hamdi Kavak
Mon Gabriel Lagustan
Christian Lebiere
Huan Liu
Chitaranjan Mahapatra
Martina Therese Reyes
Chirayu Sanghvi
Mainuddin Shaik
Ling Sun
Alexey Tregubov
Joshua Uyheng
Alina Vereshchaka
Friederike Wall
Xinshi Wang
Xinyu Wang
Dana Warmsley
Siyu Wu
Palvi Aggarwal
Hend Alrasheed
Muhammad Arslan
Mihovil Bartulovic
Alexander Caceres-Wright
Casey Doyle
Peng Fang
Juan Fernandez-Gracia
Jennifer Golbeck
Lingyun He
Ujun Jeong
Prakruthi Karuna
James Kennedy
Stephan Leitner
Yuning Liu
Raha Moraffah
David Mortimore
Md. Saddam Mukta
Salem Othman
Samantha Phillips
Rey Rodrigueza
Salim Sazzed
Jeongkeun Shin
Martin Smyth
John Stevenson
Chi-Shiun Tsai
Sharif Ullah
Henry Xu
Saklain Zaman
Emanuel Ben-David
Mert Can Cakmak
Danqing Chen
Vito D'Orazio
Wen Dong
Yash Kangralkar
Murali Mani
Hossein Salemi
Yasas Wijesuriya
Niloofar Yousefi

Contents

Advancements in Tools and Theory

Explicit Stance Detection in the Political Domain: A New Concept and Associated Dataset

Alexander R. Caceres-Wright[1(✉)], Naveen Udhayasankar[1], Grant Bunn[2], Stef M. Shuster[3], and Kenneth Joseph[1]

[1] University at Buffalo, Buffalo, NY, USA
{caceresw,naveenud,kjoseph}@buffalo.edu
[2] North Carolina State University, Raleigh, NC, USA
gbunn@ncsu.edu
[3] Michigan State University, East Lansing, MI, USA

Abstract. Stance detection, defined as the task of classifying an individual's attitude towards a target person or concept, offers the potential to understand political opinions at scale using social media data. However, recent studies have questioned the robustness and accuracy of current stance detection methods, highlighting issues such as generalizability in time and inconsistencies in annotations driven by subtle differences in annotation task design. We argue that central to these challenges is the unresolved question of what constitutes an expression of stance. To address this, the present work introduces a distinction between *explicit* and *implicit* stance expressions, and argue that a focus on explicit stance detection addresses many of the existing concerns with modern stance detection methods. To facilitate research on explicit stance detection, we then present a novel (and public) dataset of over 1000 tweets across 13 stance targets for explicit stance detection and evaluate baseline models to establish a foundation for future research in this area.

Keywords: Stance Detection · Large Language Models · Politics · Social Media

1 Introduction

Stance detection [19], the task of classifying an individual's attitude towards a target, has become one of the most popular tasks in the area of natural language processing [1,10,13]. Of particular interest is the area of *political* stance detection, where scholars have focused on detecting attitudes towards particular candidates [11] and towards broader politically-relevant claims [5]. The promise of effective political stance detection methods is that they may be able to help us understand political attitudes without the implementation of surveys that may be costly, or challenging (e.g. in hard-to-reach populations) [2,12].

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 3–14, 2024.
https://doi.org/10.1007/978-3-031-72241-7_1

However, several recent works have raised concerns about how effectively modern stance detection actually models public opinion. First, detailed analyses of existing methods suggest that existing stance detection methods are not robust to novel datasets, either with respect to changes in 1) who we aim to detect stance towards [16,17] or 2) the time at which we aim to detect stance [14,15]. Second, there are important questions about the extent to which we should expect opinion expressed on social media to actually reflect individuals' opinions, even when controlling for time [7]. Finally, gold standard datasets are subject to variation based on (often unreported) decisions about how to present the task to annotations [6]. For example, presenting annotators with the exact same tweets, but opting to present, or not present, user information alongside the tweet lead to significantly different annotations, even if both designs seem reasonable.

At the heart of these challenges is a core and largely unresolved question in both the NLP literature [7] and an open area of discussion in the relevant sociolinguistics literature [8]: *what should a person, or model, count as an expression of stance*? For example, consider a setting in which we aim to analyze a user's stance towards Donald Trump, given their tweet "I hate Kamala Harris." In the context of the 2024 presidential election, a politically knowledgeable annotator (or model) might reasonably infer that this user is pro-Trump. However, the inference relies upon an understanding that at this point in the 2024 presidential election cycle, there are two presumptive candidates - one for the Democrats and one for the Republicans. Furthermore in the context of the lead-up to the 2020 elections, one may also need to consider the possibility that the user was anti-Trump *and* anti-Harris and supporting a different candidate still in the running for the Democratic nomination [18].

This example highlights the aforementioned temporality concerns, but also brings to light an often recognized [4,11,17], but as-yet-unnamed distinction in the stance detection literature. Namely, there is a difference between *explicit* stance—where a user references the stance target and their stance towards them ("I love Trump")—versus *implicit* stance, where an annotator (or model) must *infer* stance based on available information. Designing stance detection models based only on *explicit* stance expressions has the limitation of lower recall, in that explicit user expressions of stance are relatively rare, whereas stance can (in theory) be inferred for any user [11]. But there are also a number of benefits-namely, we can expect that a model trained to detect only explicit mentions of stance is much more likely to avoid challenges of temporal drift and variations in annotation design.

To that end, the present work makes the following contributions:

- We introduce the concept of *explicit* stance detection, differentiating it from implicit stance detection and noting its' relative benefits and drawbacks
- We develop a novel and publicly available dataset[1] for explicit stance detection, consisting of over 1000 tweets across 13 different stance targets

[1] https://docs.google.com/spreadsheets/d/1ux2ap-vStSqhZ32VWtmQerrlo3qcDBl73PET9ECS7BA/edit?usp=sharing.

– We develop and evaluate a number of initial, straightforward baseline models for explicit stance detection for others to build upon in future work

2 Data

In this section, we describe the development of our annotated dataset for explicit stance detection. We define a tweet as expressing explicit stance when two conditions are met: 1) the tweet (explicitly) mentions the stance target, and 2) the tweet unambiguously expresses a stance towards the target. In cases where a tweet does not meet condition 1), stance should be labeled as "Not relevant," i.e. not relevant to the task. In cases where a tweet meets condition 1) but not condition 2), the tweet is labeled as "Neutral."

To begin, we draw on the publicly available (upon request) dataset from Shuster et al. [18], who collected over 500M tweets sent during 2020 leading up to the U.S. presidential election. Their data is of interest precisely because we expect *implicit* stance detection to be challenging in such a setting: users varied widely in their attitudes towards individual candidates, and thus (as in our introductory example) inference of stance towards one candidate given stance towards any other candidate is challenging. Here, then, a model that infers stance only from *explicit* stance expressions is particularly useful and desirable.

Table 1. Statistics for the annotated dataset for explicit stance detection. All agreement scores are Krippendorf's alpha; the final two percentage columns reflect the final, gold-standard dataset

Candidate	Mentions Agreement	Full Agreement	Agreement w/o Neutral	% Neutral	% Where Filter Failed
Michael Bennet	0.80	0.55	0.68	44.90	6.12
Joe Biden	−0.01	0.57	0.72	25.64	0.00
Pete Buttigieg	0.61	0.71	0.91	44.90	6.12
Amy Klobuchar	1.00	0.46	0.63	31.58	0.00
Deval Patrick	0.96	0.43	0.63	20.20	26.26
Bernie Sanders	0.92	0.43	0.65	20.20	5.05
Tom Steyer	1.00	0.58	0.56	29.67	0.00
Donald Trump	0.55	0.46	0.83	30.53	4.21
Andrew Yang	0.56	0.57	0.81	31.18	4.30
Tulsi Gabbard	1.00	0.71	0.80	29.00	0.00
Michael Bloomberg	1.00	0.52	0.58	23.71	0.00
John Delaney	1.00	0.48	0.70	40.00	0.00
Elizabeth Warren	0.79	0.49	0.83	26.80	2.06

From the dataset published by [18], we used a simple keyword-based filtering approach to identify all tweets that explicitly mentioned one of the twelve candidates in the 2020 DNC Primary, as well as Donald Trump. Stance targets are listed in the first column of Table 1. To do so, we created a dictionary containing a list of terms relevant to each candidate. If a tweet's text did not contain any of these terms it was excluded from the final dataset. This filtering step resulted in a set of 17,429,630 unique, non-retweeted tweets, or roughly 30.35% of the set of unique non-retweets in the original dataset. Once we had this filtered dataset, we then sampled approximately 100 tweets per politician. From this sample, we then had three separate annotators label each tweet for two categories: 1) if the tweet did in fact mention the candidate, and 2) if so, the stance of the tweet towards that particular candidate. Following prior work on creating accurate, curated datasets, annotators of the data were the authors of the paper [6,7].

For (explicit) stance annotation, we used the standard three-label approach of positive, negative or neutral. Once all annotators completed their annotations, we calculated Krippendorff's alpha [9] to determine the level of agreement across the three annotators. Finally, for each tweet, we then calculated a single, final gold label by the principle of majority rules; i.e. by picking the label that at least two of the three annotators agreed on. In the limited number of cases where all three of the annotators disagreed, we asked an annotator who had not seen those tweets before to label those tweets and used these labels as gold.

Table 1 shows summary statistics for our explicit stance detection dataset. In the first column, we show annotator agreement (in terms of Krippendorf's alpha) for whether or not a tweet meets condition 1 above; i.e. whether or not the tweet actually mentions the candidate. The table reflects several important points.

First, detecting a mention of an account was not always easy. One measure of this is presented in the last column in Table 1, which displays the number of tweets where our keywords filter failed, meaning our filtering identified it as a tweet that mentioned that candidate when in fact it did not. Most obvious in this setting are issues surrounding Deval Patrick, whose last name is popular and thus filters in many irrelevant tweets. However, in most cases, our keyword filter was effective in identifying tweets relevant to candidates. Perhaps more interesting is that, as the "Mentions Agreement" column shows, even for human annotators, determining whether or not a tweet explicitly mentioned a candidate was challenging, especially for some candidates. This could be due to the unfamiliarity of some of the annotators with specific keywords our filter used, as well as differing opinions on what they considered to be explicit mentions. For example, some annotators considered decontextualied hashtags added at the end of tweets, such as '#YangGang' and '#Berniecrats', as explicit references to Andrew Yang and Bernie Sanders, respectively, while others did not. Ultimately, we opted to include these as explicit references in our dataset unless they had absolutely nothing to do with the rest of the tweet content (e.g. in the few instances of hashtag spamming we observed).

Second, even when annotators agreed that a candidate was mentioned, determining whether stance was expressed was hard for some candidates. When including the neutral category, Krippendorf alphas ranged from 0.43–0.71, in line with agreement on other annotation tasks for social media [3] but lower than when we a priori expected given an explicit mention of a candidate. These agreements increased, however, to a range of 0.56–0.91 when we exclude cases where the final gold-standard label was neutral. This reflects that allowing neutral labels vastly increase the amount of disagreement between the annotators. Finally, and related, is that our explicit stance detection dataset contains significantly more neutral labels (roughly 30% of all annotations) than other existing datasets in the literature, which often have 10% or fewer neutral labels.

In summary, our explicit stance detection dataset shows that there are interesting challenges to be addressed in explicit stance detection versus implicit stance detection, namely in the context of the disproportionate number of cases in explicit stance detection (relative to the standard stance detection task) where a candidate is mentioned but where an explicit stance is not expressed. We now turn to an initial exploration of explicit stance classification, defining a suite of simple but reasonable baseline models and evaluating them.

3 Methods

In addition to contributing this new dataset to the research community, we develop a suite of three baseline methods to ground future work aiming to use our data to make predictions about expressions of explicit stance. Notably, we consider only open-source approaches to classification that are fully reproducible by the research community, i.e. that do not rely on a private entity opting not to depreciate a particular model version. More specifically, we consider two different models using three different approaches to leveraging large language models (LLMs) to accomplish these two tasks: a prompt-based use of Meta's Llama-2 model[2] [20] ("Llama-2 Generative"), a multiple choice pipelining approach with Llama-2[3] ("Llama-2 Pipeline"), and a similar multiple choice approach using the recently released Deberta-Polistance model,[4] a Deberta model fine-tuned for stance detection in a political setting.

The task of explicit political stance detection can be subdivided into two tasks: politician identification and stance classification. We compare two uses of Llama-2 in order to assess whether or not we can use simple prompts to see whether a generative approach based on prompting can solve both identification and stance classification in one step. In the Llama-2 Generative model, we prompt the model to provide both 1) which candidates were mentioned and 2) what the stance is towards that candidate. With respect to model hyperparameters, we experiment with a variety of 1) prompts, 2) temperatures, and 3) character limits (on the latter point, in order to assess how quickly we can complete the task). Hyperparameters evaluated are presented in the Appendix; we here take an optimistic view on performance and select for reporting results from the best-performing hyperparameter combination, because even with these optimistic scores the model still underperforms the other approaches.

In our Llama-2 Pipeline approach, we instead tell the model which candidate to provide a stance classification for. We do so via use of the `pipeline` abstraction provided by the `transformers` python library. More specifically, we evaluate the probability of the phrases "Positive towards [target]," "Neutral towards [target]," and "Negative towards [target]" in the model's next word prediction, and

2 https://huggingface.co/meta-llama/Llama-2-13b-chat-hf.
3 https://huggingface.co/meta-llama/Llama-2-13b-hf.
4 https://huggingface.co/mlburnham/deberta-v3-large-polistance-affect-v1.1.

take the option with the highest probability. While this abstraction can be useful especially for individuals who do not want to, or have the necessary experience, to directly work with the models to achieve the same result, as we will see there are some cases where directly engaging with the model can achieve a better result.

Finally, in our Deberta-Polistance Pipeline approach, we again use a classification pipeline, and specify a set of labels for the pipeline to choose from. This is a three-step process. First, we format the labels with the candidate in question. Second, we pass the text of the tweets as well as our labels to the model in the form of a fill-in-the-mask prompt. Finally, we extract the stance with the highest next phrase probability according to the model.

For all three of our approaches, we passed each tweet to the model in separately. In the case of Llama's text generation, we tested a number of different prompt options, as well as temperatures and limits on the maximum number of characters that the response could return. Using the zero shot classification pipelines, we only ran through each tweet once, using the same labels as in Llama-2 Pipeline approach. We opt not to consider hyperparameters for this approach. Once we had the labels for each tweet, we then calculated the accuracy, precision, and recall for each model as well as each parameter combination for Llama-2 on a per candidate basis. In the case there was a mismatch between the number of annotations returned by the LLM, which primarily occurred in the case of Llama's text generations, we computed each of precision accuracy and recall in two different ways. First we labelled each tweet that Llama did not provide a stance for as neutral, and second we dropped any tweet from the set of gold labels that did not have a corresponding entry in the LLM labelled tweets.

4 Results

4.1 Detecting Candidate Mentions

As shown in Table 3 even in the best case our Llama-Generate method fails to correctly identify the candidate in almost 20% of the gold label tweets, and for many candidates fails on the identification task significantly more often than this. However, as we can see from Table 4, in the cases where Llama-Generate did successfully identify the politicians explicitly mentioned in the tweet, it was able to correctly identify the stance with a higher degree of accuracy than the Llama-Pipeline method. Failures to correctly identify the candidate were impacted by hyperparameter settings. For example, using a lower limit on the number of characters the model is allowed to return may cut off the part of the response discussing the relevant candidate. However, as we can see in Table 3 across all the different parameter combinations we tried, there were still significant failures even when extending to significantly longer response limits. To this end, another possible explanation could be in our choice of prompt, which is something else we attempted to control for. We tried four different prompts, which can be found in the appendix, and used each of them for each tweet. Again, these prompt options did not have a large effect on the both the number of responses returned nor the metrics which will be discussed later in this paper (Table 2).

Table 2. Percent of Gold Labels Annotated by Llama Generate

Candidate	Percent Annotated	Candidate	Percent Annotated
Bennet	65.97	Patrick	52.92
Biden	77.08	Steyer	70.71
Bloomberg	76.07	Sanders	68.26
Buttigieg	75.09	Trump	53.82
Delaney	83.08	Warren	68.25
Gabbard	72.73	Yang	75.08
Klobuchar	73.84		

In contrast to the challenges faced by Llama-Generate, Table 4 shows that our filter was successful in the majority of cases, struggling to correctly identify only those candidates with more common names, such as Deval Patrick, or candidates such as Donald Trump, which have multiple prominent family members which share their last name thus leading to false positives. In summary, then, our findings for candidate mention performance conform to other relevant recent work on a broader set of computational social science tasks (including stance detection) suggesting that using domain-relevant knowledge (in this case, keywords to detect candidate mentions) is still an effective approach beyond relying on generative models to complete tasks without such knowledge [21].

Table 3. Percent of Label Allocation for Llama Using Generate

Candidate	Gold labels	Avg. Annotations	Gold Pct. Positive	Avg. Pct. Positive Labels	Gold Pct. Negative	Avg. Pct. Negative Labels	Gold Pct. Neutral	Avg. Pct. Neutral Labels
Bennet	98	64.65	7.14	27.55	41.84	24.27	44.90	48.18
Biden	117	90.19	8.55	25.72	65.81	58.25	25.64	16.02
Bloomberg	97	73.79	3.09	26.34	73.20	55.82	23.71	17.84
Buttigieg	98	73.59	7.14	24.21	41.84	50.09	44.90	25.70
Delaney	95	78.93	14.74	29.41	45.26	30.27	40.00	40.32
Gabbard	100	72.73	31.00	37.56	40.00	36.90	29.00	25.54
Klobuchar	95	70.15	6.32	32.66	62.11	37.62	31.58	29.72
Patrick	99	52.39	6.06	28.06	47.47	18.49	20.20	53.45
Steyer	91	64.35	25.27	40.13	45.05	25.54	29.67	34.32
Sanders	99	67.58	67.68	77.01	7.07	11.15	20.20	11.84
Trump	95	51.13	2.11	21.17	63.16	71.44	30.53	7.38
Warren	97	66.20	13.40	30.29	57.73	49.34	26.80	20.37
Yang	93	69.83	32.26	46.26	32.26	17.87	31.18	35.88

Table 4 shows accuracy under two conditions: 1) when we added in a label of neutral for any tweet in the gold annotations not found in the labelled set, and 2) by removing any tweet found in the gold set not found in the labelled set. Examining the results, we can make several general claims.

Table 4. Accuracy in Two Cases

Candidate	LLama Generate Adding Neutral	LLama Generate Dropping Diff From Gold	LLama Pipeline Adding Neutral	LLama Pipeline Dropping Diff From Gold	Deberta-Polistance Adding Neutral	Deberta-Polistance Dropping Diff From Gold
Bennet	0.46	0.44	0.38	0.36	0.61	0.59
Biden	0.56	0.64	0.66	0.71	0.79	0.86
Bloomberg	0.48	0.55	0.23	0.18	0.74	0.79
Buttigieg	0.45	—	0.42	–1.00	0.49	–1.00
Delaney	0.41	0.39	0.42	0.50	0.61	0.67
Gabbard	0.48	0.52	0.36	0.36	0.59	0.59
Klobuchar	0.46	0.50	0.60	0.50	0.71	0.25
Patrick	0.26	0.35	0.26	0.26	0.46	0.47
Steyer	0.48	0.53	0.43	0.38	0.74	0.69
Sanders	0.56	0.71	0.07	0.09	0.71	0.65
Trump	0.42	0.56	0.63	0.33	0.54	0.44
Warren	0.47	0.55	0.42	0.52	0.71	0.62
Yang	0.51	0.56	0.35	0.35	0.67	0.67

First, overall, we find that the Deberta-Polistance model performed better than Llama-2. Table 4 shows the average accuracy for text generation by Llama-2 as well both llama-2 Pipeline and Deberta-Polistance. As we can see, in all cases using Llama in a zero shot classification pipeline performs the worst out of the three models we used, while the polistance performs the best. We also see that while using Llama to generate text is not very successful at identifying the correct candidate, when it does it performs pretty well with identifying the correct stance. This suggests that if another tool is used to identify who is explicitly mentioned in the text, using a text generation such as Llama 2 may allow for improved human readability at a small trade off for accuracy. Additionally, Meta recently released Llama 3 which it reports has even better performance than Llama 2, which may improve the results of completing these tasks.

Second, however, is that this overall pattern masks significant variability across targets. Notably, Llama-Generate outperforms Deberta-Polistance for two of the more important candidates in the 2020 Election (Donald Trump and Bernie Sanders). More broadly, across models, our results is that there is wide variability across different candidates by our three models. As we can see from Table 4 the models struggled the most with Pete Buttigieg and Deval Patrick. This may be due to their relatively lesser known status as compared to the other candidates in our dataset. Even more interesting is that the agreement between our three annotators for Buttigieg is on the higher end of our agreement scores.

This variability, in turn, stems from a third interesting point: models had very different responses to the Neutral label, in ways that impacted their performance. For example, we see that gold labels for Pete Buttigieg had a large proportion of neutral tweets and from Table 3, and in turn that Llama-2 tends to favor positive or negative labels much more than neutral labels. More generally, out of the three stance labels used in this paper (positive, negative, and neutral), neutral was consistently the hardest for both human annotators as well as all three models used. As we can see in Table 3, three out of our thirteen candidates, Bennet, Buttigieg and Delaney, received a stance label of neutral about 40% of the time, while the remaining ten candidates had closer to 20% of their tweets labelled as neutral. While this could be due to the fact that the more prominent a candidate is, the greater the chance that individuals feel strongly in support or against them and are more likely to make those views public. Interestingly, with the exception of Donald Trump and Deval Patrick, the percent of tweets that llama labelled as neutral was fairly similar to the percentage of gold annotations.

5 Discussion

While the core focus of our work is not effective classification but rather the identification of explicit stance detection as a task and the contribution of a public dataset for this task, our classification exercise here presents three interesting questions to be explored in future work:

- Explicit stance detection almost necessarily results in significantly more neutral labels than in stance detection more generally, because inference of user intent is discouraged. This has implications for downstream modeling that are interesting to explore further.
- This disparity in neutral labels is further impacted by candidate status, as is which model performs best. Why models show variability across targets is an interesting point that could be further explored with our data
- Finally, it is interesting to consider how best to make use of relevant domain knowledge for explicit stance detection, as existing work tends to make use of domain knowledge to draw inferences potentially unexpressed in the text itself (and thus not relevant in the explicit setting) [11]

6 Conclusion

The present work is the first to differentiate explicit stance detection as its own concept, arguing that defining and operationalizing explicit stance detection may address certain concerns with existing stance detection models. To help bootstrap the study of explicit stance detection, we have curated a dataset with two distinct data points per datapoint: if the data point does in fact mention the target in question and separately from that the stance of the tweet towards that target in particular. This second point is especially important for the given task since one tweet can mention multiple entities and can have different stances towards each named individual.

Taken together, our findings offer demonstrable evidence of the importance of explicit stance detection, and the potential for large language models to conduct explicit stance detection given expertly curated datasets. We contribute to the literature by offering a framework for using open source models to do so, as well as a dataset which can be used as a benchmark to test. Given the moderate success of both an out-of-the-box model such as Llama-2 and a fine tuned model such as Deberta-Polistance, we leave it to future scholars to continue pursuing how to correctly identify which individuals, if any are explicitly mentioned by a given tweet.

Appendix

We considered the following prompts for Llama-Generate:

1. The following is a tweet from the 2020 US presidential election, give me the politicians it explicitly makes mention of and if it is for, against, or neutral towards each politician it mentions. Do not return the text of the tweet nor any emojis, and return your response in the following format: Politicians Name: 'stance'(for or against or neutral). Separate each politician and stance pair with a semicolon. Here is the tweet:
2. The following is a tweet from the 2020 US presidential election, give me the politicians it explicitly makes mention of and if it is for, against, or neutral towards each politician it mentions. Do not return anything besides the politician(s) mentioned and stance pair(s). Return your response in the following format: Politicians Name: 'stance'(for or against or neutral). Separate each politician and stance pair with a semicolon. Here is the tweet:
3. The following is a tweet from the 2020 US presidential election, give me the politicians it explicitly makes mention of and if it is for, against, or neutral towards each politician it mentions. Do not return anything besides the politician(s) mentioned and stance pair(s). Return your response in the following format: Politicians Name: 'stance': for or against or neutral. Seperate each politician and stance pair with a semicolon. Only include a politician if they are on the following list; Pete Buttigieg, Michael Bloomberg, Joe Biden, Michael Bennet, John Delaney, Tulsi Gabbard, Amy Klobuchar, Deval Patrick, Bernie Sanders, Tom Steyer, Donald Trump, Elizabeth Warren, and Andrew Yang. Here is the tweet:
4. The following is a tweet from the 2020 US presidential election, give me the politicians it explicitly makes mention of and if it is for, against, or neutral towards each politician it mentions. Do not return the text of the tweet nor any emojis, and return your response in the following format: Politicians Name: 'stance': for or against or neutral. Seperate each politician and stance pair with a semicolon. Only include a politician if they are on the following list; Pete Buttigieg, Michael Bloomberg, Joe Biden, Michael Bennet, John Delaney, Tulsi Gabbard, Amy Klobuchar, Deval Patrick, Bernie Sanders, Tom Steyer, Donald Trump, Elizabeth Warren, and Andrew Yang. Here is the tweet:

For temperature, we considered each of: 0.001, 0.25, 0.5, 0.75, 0.8. For character limits, we considered 25, 50, 75, and 125 characters.

References

1. AlDayel, A., Magdy, W.: Stance detection on social media: state of the art and trends. Inf. Process. Manag. **58**(4), 102597 (2021)
2. Conrad, F.G., Gagnon-Bartsch, J.A., Ferg, R.A., Schober, M.F., Pasek, J., Hou, E.: Social media as an alternative to surveys of opinions about the economy. Soc. Sci. Comput. Rev. (2019). https://doi.org/10.1177/0894439319875692

3. Du, Y., Masood, M.A., Joseph, K.: Understanding visual memes: an empirical analysis of text superimposed on memes shared on twitter. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 153–164 (2020)
4. Ebrahimi, J., Dou, D., Lowd, D.: Weakly supervised tweet stance classification by relational bootstrapping. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1012–1017 (2016). http://ix.cs.uoregon.edu/%7Edou/research/papers/emnlp16.pdf
5. Hardalov, M., Arora, A., Nakov, P., Augenstein, I.: A survey on stance detection for mis- and disinformation identification (2022). http://arxiv.org/abs/2103.00242
6. Joseph, K., Friedland, L., Hobbs, W., Lazer, D., Tsur, O.: ConStance: modeling annotation contexts to improve stance classification. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1115–1124 (2017)
7. Joseph, K., Shugars, S., Gallagher, R., Green, J., Mathé, A.Q., An, Z., Lazer, D.: (mis) alignment between stance expressed in social media data and public opinion surveys. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 312–324 (2021)
8. Kiesling, S.F.: Stance and stancetaking. Ann. Rev. Linguist. **8**(1), 409–426 (2022). https://doi.org/10.1146/annurev-linguistics-031120-121256
9. Krippendorff, K.: Computing krippendorff's alpha-reliability (2011)
10. Küçük, D., Can, F.: Stance detection: a survey. ACM Comput. Surv. **53**(1) (2020). https://doi.org/10.1145/3369026
11. Li, A., Liang, B., Zhao, J., Zhang, B., Yang, M., Xu, R.: Stance detection on social media with background knowledge. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 15703–15717. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.972. https://aclanthology.org/2023.emnlp-main.972
12. McGregor, S.C., Mourão, R.R., Molyneux, L.: Twitter as a tool for and object of political and electoral activity: considering electoral context and variance among actors. J. Inf. Technol. Politics **14**(2), 154–167 (2017). https://doi.org/10.1080/19331681.2017.1308289
13. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 31–41 (2016)
14. Mu, Y., Jin, M., Bontcheva, K., Song, X.: Examining temporalities on stance detection towards COVID-19 vaccination. arXiv preprint arXiv:2304.04806 (2023)
15. Ng, L.H.X., Carley, K.: Flipping stance: social influence on bot's and non bot's COVID vaccine stance. arXiv preprint arXiv:2106.11076 (2021)
16. Ng, L.H.X., Carley, K.M.: Is my stance the same as your stance? A cross validation study of stance detection datasets. Inf. Process. Manag. **59**(6), 103070 (2022). https://www.sciencedirect.com/science/article/pii/S0306457322001728
17. Sen, I., Flöck, F., Wagner, C.: On the reliability and validity of detecting approval of political actors in tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1413–1426. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.110. https://www.aclweb.org/anthology/2020.emnlp-main.110
18. Shuster, S.M., Campos-Castillo, C., Madani, N., Joseph, K.: Who supports Bernie? Analyzing identity and ideological variation on Twitter during the 2020 democratic

primaries. Plos One **19**(4), e0294735 (2024). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0294735
19. Somasundaran, S., Wiebe, J.: Recognizing stances in ideological on-line debates. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 116–124. Association for Computational Linguistics (2010). http://dl.acm.org/citation.cfm?id=1860645
20. Touvron, H., et al.: Llama 2: open foundation and fine-tuned chat models (2023)
21. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D.: Can large language models transform computational social science? Comput. Linguist. **50**(1), 237–291 (2024)

Multimodal LLMs Struggle with Basic Visual Network Analysis: A VNA Benchmark

Evan M. Williams(✉) and Kathleen M. Carley

Carnegie Mellon University, Pittsburgh, PA 15213, USA
{emwillia,carley}@andrew.cmu.edu

Abstract. We evaluate the zero-shot ability of GPT-4 and LLaVa to perform simple Visual Network Analysis (VNA) tasks on small-scale graphs. We evaluate the Vision Language Models (VLMs) on 5 tasks related to three foundational network science concepts: identifying nodes of maximal degree on a rendered graph, identifying whether signed triads are balanced or unbalanced, and counting components. The tasks are structured to be easy for a human who understands the underlying graph theoretic concepts, and can all be solved by counting the appropriate elements in graphs. We find that while GPT-4 consistently outperforms LLaVa, both models struggle with every visual network analysis task we propose. We publicly release the first benchmark for the evaluation of VLMs on foundational VNA tasks.

Keywords: Vision Language Models · Graphs · Benchmark

1 Introduction

Large Language Models (LLMs) and Large Vision Language Models (VLMs) are transforming the ways people work and research. There has recently been a surge of interest in understanding how LLMs can be used in network analytic workflows and evaluating their performance on network-level tasks [5]. As myriad Generative AI tools are introduced to assist practitioners with every aspect of data analytics pipelines, understanding the limitations of these tools is important. Network Analysis is used in almost every domain imaginable, and network visualization is often a core component of network analytics pipelines. Misinterpreting a graph can lead to wildly incorrect conclusions about the underlying data, which could lead practitioners to act in sub-optimal ways.[1] We introduce the broad task of zero-shot Visual Network Analysis (VNA), which we define as deriving graph theoretic concepts from network visualizations without relying on human-annotated examples. We introduce a new VNA benchmark and evaluate the performance of GPT-4 [1] and LLaVa1.5-3b [11] on 5 tasks based on 3 foundational network science concepts.

[1] At the time this was first submitted to ArXiv, there were not yet any public visual network analysis tasks. However, two datasets have since been released; we note that none of the tasks introduced in this work were considered in the concurrent works or benchmarks.

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 15–24, 2024.
https://doi.org/10.1007/978-3-031-72241-7_2

We introduce 5 VNA tasks based on three simple and important graph theory concepts: 1) degree, 2) structural balance, and 3) components. We create two tasks related to degree. Given a visualized graph, we ask the multimodal LLM to 1a) identify the maximum degree of the graph, i.e., the largest degree of any node, and 1b) to return the node IDs of all nodes with the maximum degree. We create one task based on structural balance: 2a) given an image of a triad with edges colored to denote a relation type, we ask the multimodal LLM to assess whether the triad is balanced or imbalanced. Finally, given an image of graphs with multiple components, 3a) we ask the LLM to count the number of components in the graph, and 3b) we ask the LLM to count the number of isolates.

Each of these tasks is related to an important graph theoretic concept, but each task is also connected in how it can be solved. Every task we create can be solved by counting the correct element within each graph. For the degree-level tasks, the multimodal LLM needs to count edges incident to the nodes that appear to have the largest number of edges. In the structural balance tasks, the answer can be deduced by either counting the number of positive or negative edges in the triad. The component tasks are very explicitly asking the multimodal LLM to count the number of components and the number of isolates. Consequently, these tasks, as we've constructed them, are highly related to zero-shot object counting [14]. For each of these tasks, we generate synthetic, high-resolution, graph visualizations while prioritizing human readability. We publish the all of the data generated for this project[2]

We find that GPT-4 and LLaVa struggle on all 5 tasks we propose. Across all experiments, the highest accuracy was achieved by GPT-4 on the isolate counting task, where it correctly identified the number of isolates in 67 of 100 graph visualizations. Predicting whether or not triads were structurally balanced was surprisingly one of the most challenging tasks for both LVMs, despite the simplicity of the task. The best performing model—GPT-4—achieved an overall accuracy of 0.51 on the task, on par with random guessing. More work is needed to understand and improve zero-shot LVM performance on visual graph analysis tasks.

2 Related Works

Recent work has explored potential Large Language Model (LLM) applications within social networks [15], and in generating features and predictions for machine learning applications on graphs [4]. Recent work has also explored how best to encode graphs as text for LLM analysis [5]. However, each of these works look solely at LLMs and do not consider VLM usage. Two days prior to the release of this publication on ArXiv, the VisionGraph Benchmark was released [10]. The visiongraph benchmark contains complementary tasks and evaluates

[2] https://figshare.com/articles/dataset/Multimodal_LLMs_Struggle_with_Basic_Visual_Network_Analysis_a_Visual_Network_Analysis_Benchmark/25938448.

VLM performance on many relatively-complex graph tasks, including cycle identification, identifying shortest paths, and identifying maximum flow [10]. Wei et al. concurrently introduced a GITQA (Graph-Image-Text Question Answering) Dataset and an end-to-end framework for general graph reasoning [13].

There has also been a recent surge in interest in zero-shot evaluation capabilities Large Vision Language Model (VLM) evaluations. These evaluations have been applied to a wide range of computer vision tasks [17]. [14] propose the task of zero-shot object counting, which they define as counting the number of instances of a specific class in the input image given only the class name. Llava has been evaluated on various zero-shot tasks, and has performed well on differentiating animals, counting animals, identifying written digits, and identifying pox [9]. Previous work has found that VLMs perform worse than specialized models at object counting, and seemed to perform better at counting cars than counting trees or animals [16].

3 Methods

In all tasks, each graph is independently sent to GPT-4 along with its accompanying text prompt using OpenAI's API. We also feed each graph, independently, to LLaVa, which we run on a local cluster. All LLaVa prompts were modified to include the suggested prompt format on which LLaVa was trained: "USER: <image> \n< prompt >\nASSISTANT:".

3.1 Maximum Degree Tasks

Degree centrality is likely the most widely-used graph centrality statistic. For an undirected graph G, the degree of node i is simply the number of edges incident to node i. For a binary, symmetric, and undirected adjacency matrix A, degree is simply defined as $\sum_i^N x_i$. In social networks, degree centrality often corresponds to popularity or engagement, e.g., in a friendship network, a node with a high degree has many friends.

We consider two different prompts for each graph. In the first prompt, we ask the question using Graph Theory terminology—we call this our "formal prompt". In the second question, we attempt to ask the question in a more human way. We state that the graph is a first-grade friendship network and ask the same questions using the terminology like popular students and total number of friends—we call this the "human prompt". We consider two prompts for each graph. The first asks for the largest degree centrality in the graph and the list of all nodes with that centrality. The second prompt states that we are observing a first-grade friendship network and asks who has the most friends. We exclude exact prompts from this paper due to space constraints, but we make all of prompts publicly available online[3]. LLaVa would return blank fields if given the same formatting stipulations as GPT-4, so those were dropped in

[3] https://github.com/EvanUp/VNA_Benchmark/blob/main/prompts/prompts.csv.

the LLaVa prompts. In the cases where LLaVa made contradictory statements or inaccurate statements, e.g., "A has a degree centrality of 10, while B has a degree centrality of 9. The largest degree centrality is B, which is 9", we selected the most logically coherent option—in this case, (A, 10). In cases where LLaVa or GPT-4 did not return a numeric maximum degree or node IDs, we impute a maximum degree of 0 and assign the response an empty set of node IDs.

3.2 Structural Balance Task

Balance Theory, which dates back to Fritz Heider's 1946 studies of individual cognition and perception of social situations, is a core concept in social network analysis [8]. Heider's original formalization of *cognitive balance* was generalized in the 1950s to *structural balance*, which focuses on a group of people rather than an individual [3,7]. Given a undirected signed graph $\mathcal{G}_{-}^{+}$, where each edge can assume a value of "like" (+) or "dislike" (−), structural balance occurs when i likes j ($i \overset{+}{\leftrightarrow} j$) and i and j agree in their evaluation of k, i.e., $(i \overset{+}{\leftrightarrow} k \wedge j \overset{+}{\leftrightarrow} k) \vee (i \overset{-}{\leftrightarrow} k \wedge j \overset{-}{\leftrightarrow} k)$. Concretely, for signed, undirected, triads, this means that any triad with a total of 1 or 3 (+) edges is balanced. Conversely, triads with 0 or 2 (+) edges are considered unbalanced. Consequently, of the 8 possible signed triads, 4 are structurally balanced and 4 are unbalanced. For more information on structural balance we refer the reader to Chap. 6 of [12]. We considered two different prompts, one without a definition of structural balance and one containing a simple definition of structural balance.

3.3 Component Tasks

A component is a connected graph or subgraph that is not a part of a larger connected subgraph. In undirected graphs, the number of components can be defined as the minimum number of random walkers that it would take to reach every node on a graph. Nodes of degree zero are a special type of component commonly referred to as isolates. We considered two prompts for this task—one without definitions and one with definitions. Again, LLaVa was not generating responses with the formatting requirements included in the prompt, so its prompt was truncated. LLaVa's return format with this prompt was largely consistent and answers could be reliably extracted with a regex. Of the 400 LVM calls for this task, there were 4 instances, all from GPT-4, where no numbers were returned. These instances were imputed with zeros.

4 Data

For each experiment, we programatically generate graphs using the python libraries NetworkX and netgraph. For the degree tasks, graphs were generated with Kamada Kawai layouts, as we qualitatively found these to be the most human-readable for visual node degree tasks. We bolded font weights in the degree and structural balance tasks for the same reason. All graphs are exported as high-quality 300 DPI .png files.

4.1 Maximum Degree Graph Generation

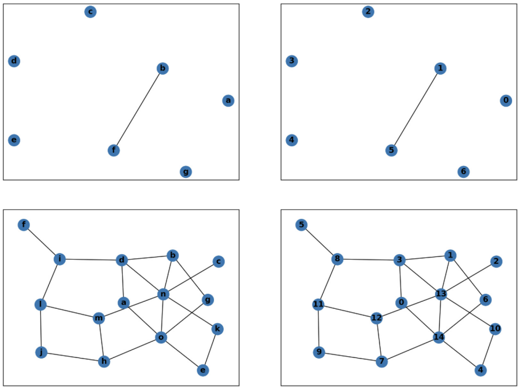


Fig. 1. Degree Task Graph Examples with letter (left) and numeric (right) node IDs.

We provide the LLMs with two degree-related tasks. First, we ask the LLM to find the maximum degree of the graph. Second, we ask the LLM to return the node IDs of all nodes with the maximum degree. In large graphs, this can be a challenging, or even infeasible, task for humans. Consequently, we consider only relatively-sparse graphs ranging from 1 to 20 nodes. We generate 20 Erdos-Renyi graphs, each with the parameter p set to 0.2. As LVMs have previously been found to be prone to typographic attacks [6], we considered that numeric node IDs could impact the ability of LLMs to return numbers. Consequently, we generate two identical versions of each graph: one with numeric node IDs and one with alphabetical node IDs (see Fig. 1).

4.2 Structural Balance Graph Generation

For each of the 8 possible signed triads, we generate 10 graphs, each with random layouts for a total of 80 triad graphs. In each graph, "like" edge relations are colored blue and "dislike" edge relations are colored red. As node IDs are unimportant for this task, we arbitrarily chose to use letter node IDs. We further group types of triads into four "classes" corresponding to the number of "like" (blue) edge relations. This results in two balanced groups—3b and 1b—containing 3 and 1 "like" relationship respectively and two unbalanced groups (0b and 2b in Fig. 2). In Fig. 2, we provide examples of 4 individual triads sampled from each grouping. We note that we allow position to vary across triad images, as it should be irrelevant to this task.

4.3 Component Graph Generation

For each component graph, we independently generate 4 Erdos-Renyi graphs, with p = 0.3 and where the number of nodes for each of the 4 graphs are

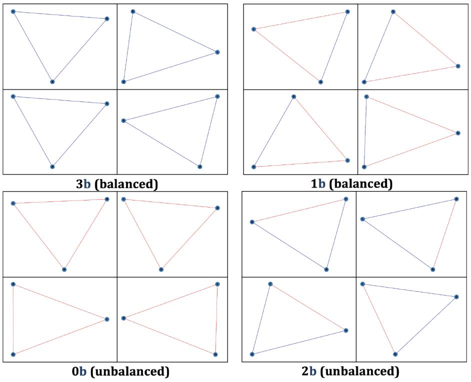


Fig. 2. Triadic Balance Examples. Top row contains a sample of balanced triads, bottom row contains a sample of unbalanced triads. ‘b’ denotes the number of like (blue) relationships in each group. (Color figure online)

randomly drawn integers between 0 and 30 inclusive. We then take the disjoint union between these four graphs, record the number of components and isolates in each graph, and visualize the result. Optimizing for human readability, we elected to visualize these graphs using the netgraph library, as it is optimized to support visual layouts containing multiple components [2]. We chose a small-world layout for each component again for human readability. Component counts across all graphs ranged from 2 to 11 and isolate counts ranged from 0 to 6. We provide four example component graphs in Fig. 3.

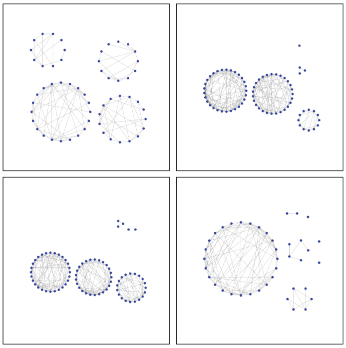


Fig. 3. Components Example Graphs. Read from left to right and top to bottom, these graphs contain 4, 5, 6, and 7 components respectively. The graphs contain 0, 1, 2, and 3 isolates.

5 Results

5.1 Maximum Degree Task Results

We evaluate GPT-4 and LLaVa at both 1) identifying the maximum node degree in a graph and 2) identifying node IDs that have the maximal degree. Despite being able to define degree centrality, LLaVa performed extremely poorly on both tasks. On the "formal" prompt, where we instruct LLaVa to identify nodes the maximum degree and identify all nodes that have the maximal degree, LLaVa failed to correctly identify any maximum degrees. On the human prompts, it exhibited bizarre behavior; it frequently made up its own graphs in its responses and then (often incorrectly) answered the question based on the graph it generated. In both scenarios, its predictions often deviated substantially from ground truth, even with easy examples, as can be seen in the relatively high MSE scores in Table 1. In the top row of graphs in Fig. 1, which very clearly has a maximum degree of one shared by two nodes, LLaVa's 4 runs identified maximum degrees of 3, 4, 10, and 11 respectively.

GPT-4 performed better than LLaVa on all metrics, but still struggled despite the simplicity of the task. A human-centric prompt with Letter IDs performed the best, yielded the best accuracy in identifying maximum degree, but did not have the best Mean Squared Error (MSE) nor the best Mean Jaccard Similarity. The human-centric prompt with letter IDs was the only run where GPT-4 correctly predicted the maximum degree of the graph in the top left corner of Fig. 1. In all other runs, GPT-4 incorrectly identified the maximum degree of the top row graphs as 2.

Table 1. Results of GPT-4 and LLava on Identifying the maximum degree (Accuracy, MSE), and at identifying the set of nodes IDs that have the maximum degree in the graph (mean Jaccard similarity over all graphs)

			Max Degree		Degree IDs
			Accuracy	MSE	Mean Jaccard Similarity
GPT-4	Formal Prompt	Numeric IDs	0.45	**17**	0.491
		Letter IDs	0.35	24	0.53
	Human Prompt	Numeric IDs	0.4	23	**0.54**
		Letter IDs	**0.533**	25	0.533
LLaVa	Formal Prompt	Numeric IDs	0	693	0.075
		Letter IDs	0	328	0.195
	Human Prompt	Numeric IDs	0.15	760	0.065
		Letter IDs	0.1	572	0.054

5.2 Structural Balance Task Results

GPT-4 generally performed well on the cases with 3 (+) edges (b1 in Fig. 2) and 0 (+) edges (b0), achieving accuracies greater than or equal to 0.7 for balanced and unbalanced triads in these categories. However, it did surprisingly poorly across cases with 1 or 2 (+) edges. Balanced triads with 1 edge (b1) with and without a definition in the prompt received accuracies of 0.2 and 0.5, respectively. Unbalanced triads with 2 (+) edges (2b) received accuracies of 0.467 and 0.367 respectively. We note that GPT-4's reasoning was often inconsistent or faulty, even when a clear definition of structural balance was provided in the prompt. For example, on one of the (b0) categories that it incorrectly classified as balanced, GPT-4 returned the justification: "This triad is balanced as all three edges depict "dislike" relationships (even number of "dislike", odd number of "like")".

When provided with a clear definition of what constitutes structural balance, LLaVa predicted that every triad was unbalanced. Without a definition, LLaVa still largely overwhelmingly predicting that triads were unbalanced. In the cases where LLaVa predicted that triads were balanced, it generally offered a perplexing or inaccurate reasoning. For example, in several cases LLaVa stated that triads were balanced because "The blue and red lines are of equal length, indicating a balance between the two relationships". One of its most frequent (faulty) justifications for classifying a triad as unbalanced, accurately or inaccurately, was some variation of "the blue and red lines are not parallel, indicating an imbalance in the relationships (Table 2)."

Table 2. Triadic Balance Results: b denotes the number of 'like' (blue) edges. 3b and 0b report accuracy only on triads containing 3 and 0 (+) relations respectively. 1b and 2b contain accuracy metrics on the triads containing 1 and 2 (+) relations respectively. The balanced column contains accuracy calculated on all balanced triads (3b and 1b), unbalanced is calculated on all unbalanced triads (b0 and 2b) and overall is accuracy calculated over all triads

		Accuracy						
		3b	1b	b0	2b	balanced	unbalanced	overall
GPT-4	No Definitions	0.70	**0.50**	0.80	0.37	**0.55**	0.46	**0.51**
	Definitions	**1**	0.20	0.90	0.47	0.40	0.58	0.49
LLaVa	No Definitions	0.10	0.23	0.80	0.73	0.20	0.75	0.48
	Definitions	0	0	**1**	**1**	0	**1**	0.50

5.3 Component Task Results

The top performing model across all evaluation metrics for both the component and isolate counting tasks was GPT-4 with no definitions included in the prompt. However, we note that the inclusion of definitions resulted in a substantial improvement of LLaVa's MSE, which decreased from over 20,000 to below 11. GPT-4 performed relatively well on the isolate counting task, and provided the correct number of isolates 67 of 100 times. We provide results in Table 3.

Table 3. Component Results. We report Accuracy, MAE, and MSE for each model and prompt condition for both the Component counting and Isolate counting tasks

		Components			Isolates		
		Accuracy	MAE	MSE	Accuracy	MAE	MSE
GPT-4	No Definitions	**0.39**	**1.25**	**3.41**	**0.67**	**0.51**	**0.99**
	Definitions	0.37	1.45	4.97	0.64	0.58	1.3
LLaVa	No Definitions	0.03	59.79	23297.95	0.08	32.43	20652.51
	Definitions	0.05	2.81	10.57	0.17	1.27	2.23

6 Discussion and Limitations

Given GPT-4's strong performance on professional exams like the LSAT and MCAT, it is, at least from a human perspective, surprising that it would struggle with something as simple as counting specific elements of graphs. This phenomenon likely relates, at least in part, to how LVMs process images as patches [14]. Additionally, it's unsurprising that GPT-4 would outperform LLaVa given LlaVa almost certainly contains far fewer (3B) parameters than GPT-4[4]. Nonetheless, more research is needed to understand why LVMs struggle on tasks this simple, and future research could explore the performance of LVMs fine-tuned on graph-related tasks. Additionally, the ways that graph visualization parameter selection and prompt engineering impact LVM performance on VNA tasks are clear and important avenues for future research.

7 Conclusion

We propose the task of zero-shot Visual Network Analysis to evaluate the performance of LVMs on graph analytics tasks. We create a benchmark that includes 5 tasks related to 3 core network science concepts: maximum degree, structural balance, and identifying components. We find that across all tasks, LLMs struggled to identify and count the appropriate element of graphs—an essential skill in analyzing network data. We publicly release all generated data and ground-truth labels.

Acknowledgements. The research for this paper was supported in part by the ARMY Scalable Technologies for Social Cybersecurity and the Office of Naval Research, MURI: Persuasion, Identity, & Morality in Social-Cyber Environments under grants W911NF20D0002 and N000142112749. It was also supported by the center for Informed Democracy and Social-Cybersecurity (IDeaS) and the center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ARMY, the ONR, or the US Government.

[4] the exact number of parameters in GPT-4 is not currently known, but it is likely far larger.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achiam, J., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Brodersen, P.J.N.: Netgraph: publication-quality network visualisations in python. J. Open Source Softw. **8**(87), 5372 (2023). https://doi.org/10.21105/joss.05372
3. Cartwright, D., Harary, F.: Structural balance: a generalization of Heider's theory. Psychol. Rev. **63**(5), 277 (1956)
4. Chen, Z., et al.: Exploring the potential of large language models (llms) in learning on graphs. ACM SIGKDD Explor. Newsl **25**(2), 42–61 (2024)
5. Fatemi, B., Halcrow, J., Perozzi, B.: Talk like a graph: encoding graphs for large language models. arXiv preprint arXiv:2310.04560 (2023)
6. Goh, G., et al.: Multimodal neurons in artificial neural networks. Distill (2021). https://doi.org/10.23915/distill.00030. https://distill.pub/2021/multimodal-neurons
7. Harary, F.: On the notion of balance of a signed graph. Mich. Math. J. **2**(2), 143–146 (1953)
8. Heider, F.: Attitudes and cognitive organization. J. Psychol. **21**(1), 107–112 (1946)
9. Islam, A., Biswas, M.R., Zaghouani, W., Belhaouari, S.B., Shah, Z.: Pushing boundaries: exploring zero shot object classification with large multimodal models. In: 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 1–5. IEEE (2023)
10. Li, Y., Hu, B., Shi, H., Wang, W., Wang, L., Zhang, M.: Visiongraph: leveraging large multimodal models for graph theory problems in visual context. arXiv preprint arXiv:2405.04950 (2024)
11. Liu, H., et al.: Llava-next: improved reasoning, ocr, and world knowledge (2024). https://llava-vl.github.io/blog/2024-01-30-llava-next/
12. Wasserman, S., Faust, K.: Social network analysis: methods and applications (1994)
13. Wei, Y., Fu, S., Jiang, W., Kwok, J.T., Zhang, Y.: Rendering graphs for graph reasoning in multimodal large language models. arXiv preprint arXiv:2402.02130 (2024)
14. Xu, J., Le, H., Samaras, D.: Zero-shot object counting with language-vision models. arXiv preprint arXiv:2309.13097 (2023)
15. Zeng, J., et al.: Large language models for social networks: applications, challenges, and solutions. arXiv preprint arXiv:2401.02575 (2024)
16. Zhang, C., Wang, S.: Good at captioning, bad at counting: benchmarking gpt-4v on earth observation data. arXiv preprint arXiv:2401.17600 (2024)
17. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: a survey. IEEE Trans. Pattern Anal. Mach. Intell. (2024)

Tiny-BotBuster: Identifying Automated Political Coordination in Digital Campaigns

Lynnette Hui Xian Ng(✉), Mihovil Bartulovic, and Kathleen M. Carley

Carnegie Mellon University, Pittsburgh, USA
lynnetteng@cmu.edu

Abstract. Automated political campaigns in the digital space can influence electoral votes and tilt the balance of power. We developed a compact ensemble approach named Tiny BotBuster to identify automated bot users, then we applied the Combined Synchronization Index to reveal the political actors working together. We applied our techniques to the 2024 Indonesian Elections, revealing groups of coordinated digital campaigns. We also characterized the coordination-automation interplay, highlighting the use of automation by political parties.

Keywords: bot detection · coordination · elections · indonesia

1 Introduction

The digital sphere has facilitated political campaigning, and social media platforms such as X serve as platforms of engagement [16]. Many political campaigns strategically use social media to shape and discern public opinion [11]. The digital campaign environment and its technology have evolved to include automation, whether via simple scripts, APIs, or generative AI to reduce the human effort required for online campaigning. The agent of automation is called a "bot", a software programmed species, and groups of bots working together to spread a common narrative are termed "automated political coordination" [10,18].

Automated political campaigns are of interest to social cybersecurity because bots used in these campaigns can cause a wide range of issues [4]: they effectively manipulate public opinion at scale, instigate quantitative misrepresentations of popularity, and create instability in voter alignment [9].

Past work reported bots influencing the UK Brexit campaign [1,3]. Bots were observed to be responsible for a large number of tweets during the past US elections, flooding the discourse with campaign information [13,21]. In Asia, state-sponsored political campaigns attempt to influence the Taiwan-China rhetoric, especially during US-hosted democracy summits [8].

The research for this paper was supported by the following grants: Cognizant Center of Excellence Content Moderation Research Program, Office of Naval Research (Bothunter, N000141812108) and Scalable Technologies for Social Cybersecurity/ARMY (W911NF20D0002). The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied.

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 25–34, 2024.
https://doi.org/10.1007/978-3-031-72241-7_3

We expand the study of political campaigns into the SouthEast Asia (SEA) region, examining the 2024 Indonesian Elections. As a technique for identifying automated users, we have developed Tiny-BotBuster, a compact ensemble approach for the efficient detection of automated users. We combine bot detection algorithms with algorithms for online coordination [12,21], identifying coordinated political campaigns and characterizing their strength.

2 Methodology

2.1 Identifying Automation with Tiny-BotBuster

To identify automated users, we make use of the concept of bot detection. Bots are defined as automated programs [20], thus in our use case, finding bots will lead us to uncover the automated users.

Past bot detection algorithms use longitudinal data to classify temporal patterns [5], language models for understanding tweet texts, or require extensive data collection to construct network graphs [7]. Tiny BotBuster evaluates bot probability based on user and tweet metadata, thus achieving its compactness by requiring a smaller feature set, and needing less data. It is a compact yet efficient and accurate bot detection algorithm. Table 1 details the features used in the model.

Figure 1 presents the architecture of Tiny-BotBuster, which is implemented in Python's scikit-learn[1]. Tiny first expresses its feature set in a vector form, before passing those features through a Stacked Classifier that consists of a Gradient Boosted and a Random Forest set up. The logits from the Stacked Classifier are passed through a Logistic Regression, which returns a probability that is thresholded at the 0.7 level to classify the user into bot/human classes. The best parameters of the Stacked Classifier were found in a two-fold iteration using a RandomSearchCV across the train set. The first iteration chose the best parameters from 500 random models across the 1.5 mil parameter space. The second iteration chose the best parameters from 500 random models from a 10k parameter space that is close to the best model from the first iteration. The 0.7 threshold is derived from past work that involved statistical analysis of stable probability scores over a large number of users across time [14]. With this architecture, Tiny can be run on CPUs, increasing its portability across setups.

Training and Evaluation. We used manually annotated data from the OSOME repository, a data repository used to train many commercial bot detectors[2]. We performed a stratified train:test split of 85:15 and tested Tiny-BotBuster for each dataset. We use BotHunter [2] and BotBuster4Everyone [13] as baseline comparisons. BotHunter is a tiered random-forest model, while BotBuster4Everyone is a set of random forests evaluated with a mixture-of-experts architecture. For evaluation, we used the weighted-F1 accuracy score, for consistency with baseline

[1] https://scikit-learn.org/stable/.
[2] https://botometer.osome.iu.edu/bot-repository/datasets.html.

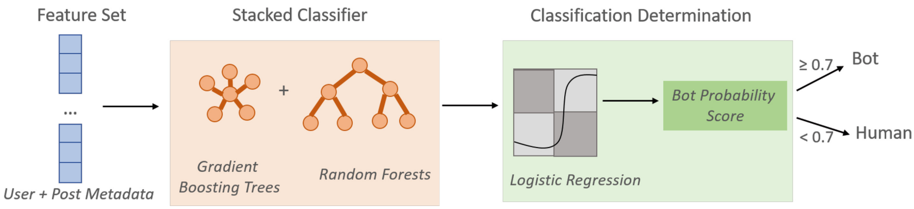


Fig. 1. Architecture Diagram

models. This score balances out the number of each class (bot, human) within the evaluation dataset. In all experiments, we used a CPU-based computer with an Intel i7 3.60 GHz CPU. Table 2 reflects the final parameters of the model.

Results. Figure 2 shows the F1 scores and timing runs for the models. Tiny-BotBuster consistently ran with a shorter runtime and higher accuracy. In terms of size, Tiny is the most compact, at 142 MB. Comparatively, its cousins are larger in size: BotHunter at 716 MB and BotBuster4Everyone at 309 MB. This compactness allows Tiny to cater to applications that have restricted computation and memory footprint. These evaluations set new benchmarks for bot detection algorithms in terms of runtime and size. However, unlike BotBuster4Everyone, Tiny lacks the ability to determine automation of users with incomplete data, and currently only works on the X platform.

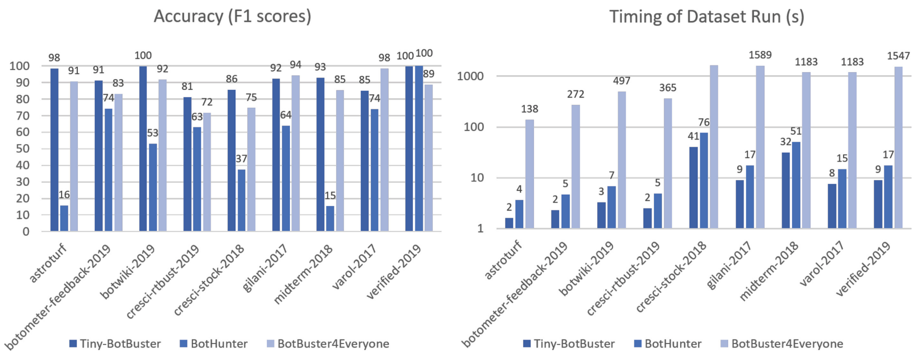


Fig. 2. F1 scores and timing runs for the models. Tiny-BotBuster surpassed baselines in delivering higher accuracy with a shorter runtime

2.2 Identifying Coordination with Combined Synchronization Index

Coordination is generally measured by a high level of synchronization of social media actions within a short timeframe of 5 min. A common technique of evaluating coordinated behavior is to project behavioral traces of online activity as similarity networks and identify connected components [17,22]. Coordinated groups have been studied in US and UK elections, serving as congregations for amplifying political campaigns [15,19].

We first construct the all-communication network, which represents users as nodes and links as a communication attribute between users (e.g., retweet, quote). We then parse this network into the Combined Synchronization Index (CSI) implementation of the ORA software[3] to retrieve coordinated networks. The Index evaluates collective action between users based on three action types: semantic, referral, and social synchronicity [12]. Users that exhibit a high amount of synchronicity (i.e., 1 stdev above the mean synchronicity) are termed to be coordinated. This threshold filters out low-weighted edges and disconnected nodes, leaving behind connected components that are highly coordinated.

3 Case Study

We present a case study of the 2024 Indonesian General Elections. We collected data from X with the hashtag #2024indonesianelections, during the week of the elections (7-18 Feb 2024). In total, we have 86,471 tweets from 12,928 users, of which 98% of the tweets are geolocated by X to be from Indonesia. We identify automation with Tiny-BotBuster, and coordination with CSI. Here we describe the presence of automated coordination within political campaigns.

Overall, 34% of the users present as bots. However, not all users that are classified as bots by Tiny BotBuster are bots. These users are part of political campaigns, and therefore can be organizations exhibiting bot-like behavior such as frequent tweets. One such case is the suite of DPP accounts, one of the larger Indonesian political parties, which manages a series of accounts, each targeting a different citizen demographic.

The discourse presents three main lines of digital communication, expressed by the three clusters of the All Communication Network (Fig. 3(d)). Using the Latent Semantic Analysis topic analysis method on the tweets of the users within these clusters, the themes of communication correspond to the three key presidential candidates: Group H, consisting of 19.4% bots, is the campaign for PrabowoGibran which promotes Prabowo Subianto and his running mate Gibran Rakabuming of the Gerindra party. Group I, consisting of 22.3% bots, is the campaign for GanjarMahfud, which promotes Ganjar Pranowo and his running mate Mahfud MD of the PDI-P party. Group J, consisting of 19.9% bots, is the campaign for AniesMuhaimin, which promotes Independents Anies Baswedan and his running mate Muhaimin Iskandar. The Presidential election was won by a landslide by the Gerindra party with 58% of the votes. The incumbent party

[3] http://www.casos.cs.cmu.edu/projects/ora/index.php.

PDI-P presents the most proportion of bot accounts while the winning party presents the least proportion. The results suggest that the winning party was confident in its candidate's victory based on its evaluation that its offline grassroots efforts were effective, requiring less digital campaigning. In contrast, the incumbent party attempted to boost its chances by using digital campaigning to sway votes.

The All-Communication Network cannot effectively showcase campaign clusters, hence we turn to the coordination networks. There is minimal referral coordination, i.e., minimal sharing of common URLs. In comparison, past work within the US context shows extensive referral coordination across multiple events [12,22]. This difference showcases the uniqueness of the SEA political discourse. Therefore, it is important to study the region since online communication techniques do not necessarily generalize from the Western to SEA front.

Across the three action types, the bot-human coordination strength is the highest. This is a reflection of how much automated accounts interact with, and potentially influence humans. This is consistent with past work which also finds the bot-human pair has the highest amount of coordination [12].

Automation vs. Coordination. Figure 3 annotated clusters of coordinating users which are described in the following paragraphs.

Flocking Humans have high coordination, low automation. This refers to chatty humans that share the same conversation topics. These flocking humans are embedded within clusters of bot-human coordination (e.g., A, B, D, F), reflecting that the discourse bounces off both species of users. Cluster A and D discuss the GanarMahfud campaign; Cluster F discuss the PKB campaign; and Cluster B discusses the general political landscape.

Bot Farms have high coordination, high automation. Cluster C is an isolated bot farm that does not interact with users outside the group and mostly tag each other. The users belong to PKB, the National Awakening Party, which is an Islam-based political party, and Anies-Baswedan campaign (of which PKB-affiliated Iskandar is the running mate). These bots, while aggressive, seem to not to be connecting to the general discourse. Other bot farms are embedded within clusters of bot-human coordination (e.g., A, B, D, F), suggesting that the bot farms successfully penetrate the human social circle. The GanarMahfud campaign coordinates socially and semantically, in Clusters A and D. The key bots in these networks (measured by CSI) are news networks such as MetroTV and political figures. This suggests that the campaign uses traditional TV outlets for political advertising and that political figures are Cyborgs, relying partially on automation to automate their messaging. Cluster F is mainly bots from the PKB political campaign, with automated users that are directly associated with the PKB party (i.e. usernames prefixed with PKB_) having high betweenness centrality values within the network, indicating that these users may be humans that leverage automation to schedule their tweets and are central to message dissemination. Finally, Cluster B contains political pundits as bot users, voicing their political views, and echoing the views of their community (seen with high social coordination).

Isolates have low coordination. They can be bots or humans depending on the amount of automation present. In our data, we do not observe isolated users, in the original dataset (see Fig. 2d), suggesting that the discourse is well connected across conversational topics.

Political Bridges have a mixture of coordination and automation. Cluster E is a group of humans that bridge between two clusters using common hashtags with clusters D and F. The cluster is made up of mostly political commentators and observers, of which 37.8% are automated users that blast and aggregate political news, akin to news bots.

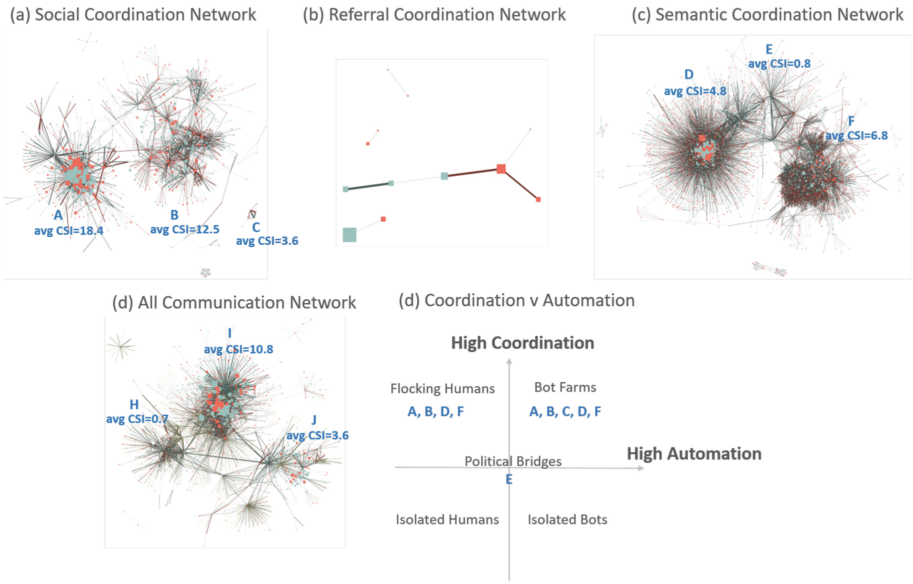


Fig. 3. Network graphs showing All-Communication and coordination graphs. Red nodes are automated users, green are human users. Nodes are sized by Combined Synchronization Index. Clusters are annotated by average user CSI, and indicated by letters which are elaborated in Sect. 3 (Color figure online)

4 Discussions and Conclusions

Digital campaigning represents a distinctive shift in political strategy. The online campaign extends a party's structures and increases the visibility of a political campaign's digital infrastructure and activity, allowing not only official campaign narratives but also satellite campaigns by grassroots and activists [6].

To increase the volume and velocity of a digital campaign, many parties rely on algorithmic political communication, i.e., the deployment of automated users with specific communicative actions to disseminate information [6]. These automated political bots have been designed to manipulate public opinion in the US elections [6]. Our study reflects a similar concept within the Indonesian elections, where political campaigns employ bots. In fact, official political party handles are often misclassified as bots, suggesting the considerable use of automation for those handles during the campaigning period.

Finally, political bots take part in political coordination, whether overtly or covertly, soliciting monetary contributions, votes, or flooding the zone with desired ideologies [6,8]. Within the Indonesian context, political bots flood the zone through extensive promotion of their political candidate, using a large number of hearts and stars emojis, e.g., "❤️❤️❤️❤️❤️❤️❤️❤️❤️❤️❤️ #GanjarMahfud2024...", or "❣️❣️❣️ #aniesbasw[...]". These bots also repeat the same messages multiple times. For example, the message "[...] TOP GAN, lanjoot.. ☕ Pa 5G Okkeh.. Publik antusias 5th Ganjar Okkeh !!, #GanjarMahfudM3nang #GanjarMahfud2" was repeated 10 times, each message tweeted 12 h apart, by the same bot user. This resonates with techniques observed in studies within Taiwan-Chinese discourse of Repeater Bots using repetition of the same phrase to drill home the message. [8].

Our findings show three key points: first, 34% of users are reflected as automated bots. This is a larger percentage than reflected by BotPercent, which estimates 7–14% of bots across countries [20]. However, BotPercent does not estimate across SEA countries, leaving our observation as the current statistic. Additionally, we observe some disconnect between the extent of automation used versus the actual electoral results, in the Presidential race. In the legislative race, the percentage of bots used showed a broad but non-absolute correlation with the percentage of seats won in the House of Representatives. The winning party, PDI-P, led with 16.7% of seats and used the most bots (22.3%). In contrast, the Golkar party, which secured the second-largest number of seats (15.3%), had barely any bots or users active within the social media discourse. Therefore, the correlation between bot usage and election success is not consistent across all parties. These results echo past work of active but ineffective bots in the past Singapore elections [23], suggesting that automated campaigns may not be as effective in the SEA compared to the Western context, and other factors have stronger contributions to a party's dominance in the elections.

Second, the nature of political coordination in the Indonesian elections is primarily social and semantic based and is segregated into clear clusters. In contrast, observed coordination within several US-based events show extensive coordination across the three action types, and a lack of distinct clusters [12]. This reflects a need to study the same phenomenon outside commonly studied Western context, as its presentation differs across regions.

Third, the extensive bot-human coordination provides cause for concern because these automated users tend to amplify and spread disinformation [11], and are also used by state-sponsored actors to spread ideology [8]. The more

Table 1. Features and data type used in Tiny BotBuster

Feature	Description	Data Type
name_length	number of characters in name	integer
account_age	number of days account is alive	integer
has_description	presence of non-empty description in user bio	boolean
source	platform used for posting (e.g., Twitter for Android, TweetDeck)	string
has location	presence of non-empty location string	boolean
string_entropy	character entropy of screen name	float
bot reference	presence of indications of word 'bot' in username or description	boolean
friends_count	number of friends the user has	integer
followers_count	number of followers the user has	integer
friends_followers_ratio	the ratio of the number of friends to followers the user has	float
favorites_count	the number of favorites the post has	integer
statuses_count	the number of statuses the user has posted	integer
status_isretweet	whether the status posted is a retweeted status	boolean
last_status_hasthags	number of hashtags in the status	integer
last_status_mentions	number of mentions in the status	integer
status_possibly_sensitive	presence of sensitive phrases in the post	boolean
has_default_profile	presence of default profile pictures	boolean
tweets_per_day	number of tweets the user posts per day	integer
emojis_in_name	number of emojis in user name	integer
emojis_in_description	number of emojis in user description	integer
verified	presence of verified flag	boolean

Table 2. Final Parameters for Tiny BotBuster

RandomForestClassifier		GradientBoostingClassifier	
num estimators	1600	num estimators	300
max depth	10	max depth	10
bootstrap	False	learning rate	0.2
		min samples split	5
		max features	sqrt
		subsample	0.8

extensive the bot-human coordination, the more likely the ideology can sway human opinions through the repeated exposure of similar narratives.

There are several limitations that nuance this work. First, Tiny BotBuster is trained on the OSOME dataset, which largely contains data from the Western hemisphere. Since the algorithm is largely based on user and tweet metadata rather than language dependent features like text, and due to the lack of data for the SEA region, we opted to use the available training data. Nonetheless, we manually verified the results, and we reflected our discussion of it within the last section. Second, the case study data was collected using the new paid Twitter API, which provides only a small proportion of the discourse. While limited, this data still provides us insight into the structure of digital campaigns in the SEA region within legal boundaries.

Future research involves expanding Tiny-BotBuster across more social media platforms, and defining a typology to combine the type of bots against the extent of coordination, thereby providing greater insight into the interactions between organic and inorganic users.

In this work, we expand the study of online political discourse into automated political coordination, serving as a primer to studying digital campaigns. This work focuses on the Indonesian Elections, a SouthEast Asian region, highlighting the differences and similarities with Western-based studies. The combination of Tiny-BotBuster and the Combined Synchronization Index provides an efficient way to filter large amounts of data for automated groups, enabling analysts to quickly zoom in on groups of interest.

References

1. Badawy, A., Addawood, A., Lerman, K., Ferrara, E.: Characterizing the 2016 Russian IRA influence campaign. Soc. Netw. Anal. Min. **9**, 1–11 (2019)
2. Beskow, D.M., Carley, K.M.: Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In: Conference paper. SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, vol. 3 (2018)
3. Bruno, M., Lambiotte, R., Saracco, F.: Brexit and bots: characterizing the behaviour of automated accounts on twitter during the UK election. EPJ Data Sci. **11**(1), 17 (2022)
4. Carley, K.M.: Social cybersecurity: an emerging science. Comput. Math. Organ. Theory **26**(4), 365–381 (2020)
5. Chavoshi, N., Hamooni, H., Mueen, A.: Debot: twitter bot detection via warped correlation. In: ICDM, vol. 18, pp. 28–65 (2016)
6. Dommett, K., Temple, L.: Digital campaigning: the rise of facebook and satellite campaigns. Parliament. Affairs **71**(suppl_1), 189–202 (2018)
7. Feng, S., et al.: Twibot-22: towards graph-based twitter bot detection. Adv. Neural. Inf. Process. Syst. **35**, 35254–35269 (2022)
8. Jacobs, C.S., Ng, L.H.X., Carley, K.M.: Tracking china's cross-strait bot networks against Taiwan. In: Thomson, R., Al-khateeb, S., Burger, A., Park, P., A. Pyke, A. (eds.) SBP-BRiMS, pp. 115–125. Springer, Heidelberg (2023). https://doi.org/10.1007/978-3-031-43129-6_12
9. Keller, T.R., Klinger, U.: Social bots in election campaigns: theoretical, empirical, and methodological implications. Polit. Commun. **36**(1), 171–189 (2019)
10. Khaund, T., Kirdemir, B., Agarwal, N., Liu, H., Morstatter, F.: Social bots and their coordination during online campaigns: a survey. IEEE Trans. Comput. Social Syst. **9**(2), 530–545 (2021)
11. Kreiss, D.: Digital campaigning. In: Handbook of Digital Politics, pp. 118–135. Edward Elgar Publishing (2015)
12. Ng, L.H.X., Carley, K.M.: A combined synchronization index for evaluating collective action social media. Appl. Netw. Sci. **8**(1), 1 (2023)
13. Ng, L.H.X., Carley, K.M.: Assembling a multi-platform ensemble social bot detector with applications to us 2020 elections. Soc. Netw. Anal. Min. **14**(1), 1–16 (2024)

14. Ng, L.H.X., Robertson, D.C., Carley, K.M.: Stabilizing a supervised bot detection algorithm: how much data is needed for consistent predictions? Online Soc. Netw. Media **28**, 100198 (2022)
15. Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., Tesconi, M.: Coordinated behavior on social media in 2019 UK general election. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, pp. 443–454 (2021)
16. Opeibi, T.: The twittersphere as political engagement space: a study of social media usage in election campaigns in Nigeria. Dig. Stud.s/Le champ numérique **9**(1) (2019)
17. Pacheco, D., Hui, P.M., Torres-Lugo, C., Truong, B.T., Flammini, A., Menczer, F.: Uncovering coordinated networks on social media: methods and case studies. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, pp. 455–466 (2021)
18. Rheault, L., Musulan, A.: Efficient detection of online communities and social bot activity during electoral campaigns. J. Inf. Technol. Polit. **18**(3), 324–337 (2021)
19. Suresh, V.P., Nogara, G., Cardoso, F., Cresci, S., Giordano, S., Luceri, L.: Tracking fringe and coordinated activity on twitter leading up to the us capitol attack. arXiv preprint arXiv:2302.04450 (2023)
20. Tan, Z., et al.: Botpercent: estimating bot populations in twitter communities. arXiv preprint arXiv:2302.00381 (2023)
21. Tardelli, S., Nizzoli, L., Avvenuti, M., Cresci, S., Tesconi, M.: Multifaceted online coordinated behavior in the 2020 US presidential election. EPJ Data Sci. **13**(1), 33 (2024)
22. Tardelli, S., et al.: Temporal dynamics of coordinated online behavior: stability, archetypes, and influence. Proc. Natl. Acad. Sci. **121**(20), e2307038121 (2024)
23. Uyheng, J., Ng, L.H.X., Carley, K.M.: Active, aggressive, but to little avail: characterizing bot activity during the 2020 Singaporean elections. Comput. Math. Organ. Theory **27**(3), 324–342 (2021)

SAARTHI: Smart Auto Assessment and Roadside Technical Help Interface

Chirayu Sanghvi(✉) and Alina Vereshchaka

Department of Computer Science and Engineering,
State University of New York at Buffalo, Buffalo, USA
{chirayus,avereshc}@buffalo.edu

Abstract. Emergency vehicle accidents pose significant challenges to operational efficiency and financial stability within emergency services, impacting organizations and communities. These incidents result in substantial repair costs, prolonged vehicle downtime, and potential legal liabilities, straining crucial public safety resources. Additionally, issues like inflated repair costs, inefficient roadside assistance, and lengthy insurance processes compound these challenges. The Smart Auto Assessment and Roadside Technical Help Interface (SAARTHI) provides a comprehensive end-to-end solution, integrating with emergency service protocols to leverage computer vision technologies for damage assessment, consistent repair cost estimation, immediate repair coordination, towing services, and faster insurance processes. By addressing these issues, SAARTHI enhances the efficiency and reliability of emergency response systems, ensuring continuous service coverage and improved operational readiness. This socially beneficial solution strengthens emergency services and promotes community safety and well-being. SAARTHI demonstration and source code are available at sites.google.com/view/saarthi-home.

Keywords: Emergency vehicles · Roadside Assistance · Convolutional Neural Networks · Object detection · Instance Segmentation · Salient object detection · End-to-end framework · Non-Maximum Suppression

1 Introduction

Emergency vehicle accidents, particularly those involving ambulances and fire trucks, pose operational and financial challenges. Immediate consequences include vehicle downtime, increased operational costs, and potential legal liabilities. The total number of ambulance crashes, including minor "fender benders," has been estimated at 6,500 per year [1]. Fire truck crashes occur at a rate of approximately 30,000 per year, having potentially dire consequences for the vehicle occupants and the community if the fire truck was traveling to provide emergency service [2]. Each year, there are approximately 300 fatalities in the U.S. that occur during police pursuits [3]. These crashes often occur at high speeds, at night, and on local roads. The economic impact of vehicle crashes is substantial, with the government paying an estimated $35 billion annually [4].

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 35–46, 2024.
https://doi.org/10.1007/978-3-031-72241-7_4

These accidents lead to issues such as inflated repair costs, delays in vehicle recovery due to inefficient roadside assistance, and prolonged insurance processes that extend vehicle downtime. These challenges strain emergency services financially and impede their ability to provide timely and effective responses.

In this paper, we introduce the Smart Auto Assessment and Roadside Technical Help Interface (SAARTHI), an end-to-end framework that addresses these challenges. SAARTHI leverages advanced artificial intelligence (AI) technologies, including object detection, instance segmentation, and salient object detection, to improve the way emergency vehicle damages are assessed and repairs are managed. This platform assesses damage from user-uploaded images, classifies the damages into six categories-dent, scratch, crack, lamp broken, glass shatter, and tire flat, and further distinguishes them as major and minor damages, also providing immediate repair cost estimates. While current processes for dealing with emergency vehicle damages are generally functional, SAARTHI aims to enhance existing protocols by offering rapid damage assessment tools, supporting decision-making for fleet managers and maintenance departments.

One of the issues in emergency vehicle operations is the variability and inconsistency in repair estimates for similar damages. There are inflated repair costs due to the lack of standardized pricing. For example, two similar ambulance accidents might result in vastly different repair quotes from different service providers, causing potential overpayment. Implementing a standardized repair cost estimation system can aid government agencies in planning their budgets for upcoming years by reducing overpayment for repairs.

Another challenge is the efficiency of roadside assistance, which is crucial for keeping emergency vehicles operational. Delays in returning a fire truck to service after an accident can leave the fire department short of essential resources, impacting their ability to respond effectively to emergencies.

To address these issues, we introduce SAARTHI, a comprehensive framework that manages the entire lifecycle of an emergency vehicle from accident to return to service. This end-to-end AI-powered framework can be directly implemented to expedite vehicle damage management. Key features of SAARTHI include:

- Real-time damage assessment. We implemented Resnet-based [5] mask-RCNN [6] model utilizing MMDetection [7], and U-net to perform real-time damage assessments from user-uploaded images.
- Detailed reporting and documentation. The system generates detailed reports and documentation of the damage, including images, statistical graphs, estimated repair costs, and other critical details.
- Repair cost estimation. By integrating Non-Maximum Suppression [8,9] and considering the base prices of different types of damages along with their impact factors, we provide accurate repair cost estimations.
- Immediate assistance via chatbot. SAARTHI includes a chatbot feature that provides immediate assistance to users.

The rest of the paper is structured as follows: Sect. 2 covers related work. Section 3 presents our approach and algorithms for damage detection and repair cost estimation. Section 4 details the SAARTHI framework. Section 5 provides experimental results, and Sect. 6 concludes the paper.

2 Background

Advanced AI technologies, such as object detection [10] and instance segmentation [11], offer robust solutions for assessing vehicle damage [12]. There is abundant research that has utilized Convolutional Neural Networks (CNNs) for detecting damages in vehicles [13,14]. CNNs have shown great promise in accurately identifying various types of vehicle damage. However, one of the significant challenges in this field has been the accurate detection of overlapping damages. Overlapping damages complicate the task of isolating individual damages, making it difficult to provide precise repair cost estimates.

In the SAARTHI framework, after implementing the initial damage assessment using object detection and instance segmentation, we apply the Non-Maximum Suppression (NMS) technique [15] to eliminate the overlapping detection of damages. NMS is crucial for refining the results of damage detection by ensuring that only the most significant damages are considered. By focusing on the most significant damage, we can more accurately identify the estimated repair cost. Detecting vehicle damage, particularly with irregular shapes and flexible boundaries, poses significant challenges. Scratches and cracks often have similar contours and colors, leading to misclassification [16]. To address this, we use Salient Object Detection (SOD) methods, which refine boundaries and segment objects with irregular shapes. SOD locates all salient objects without classifying them, ensuring accurate assessment of dents, scratches, and cracks by focusing on the location and extent of the damage.

3 Methodology

3.1 Damage Detection And Segmentation For Emergency Vehicle

The aim of car damage assessment is to accurately detect, classify, and contour damages on vehicles, aligning with the objectives of instance segmentation and object detection. We implement a Mask R-CNN [6] model with a ResNet-50 [5] backbone using the MMDetection [7,17] toolbox (Fig. 1). The model is initially pre-trained on the COCO dataset [18] and fine-tuned on the CarDD dataset [12], which consists of approximately 4000 images, manually annotated with bounding boxes and masks to improve the performance on vehicles assessments.

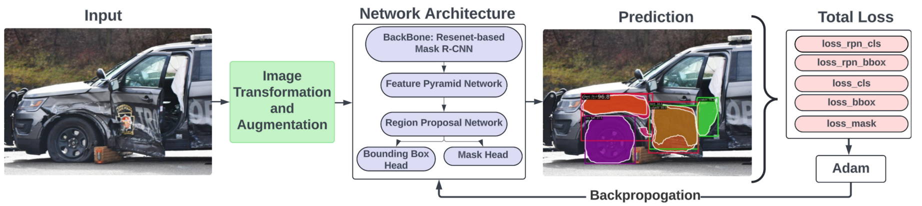


Fig. 1. Network architecture of the Mask R-CNN model with a ResNet-50 backbone, used for damage detection and segmentation in emergency vehicles

To enhance the training process, we implement a custom hook that dynamically modifies the data augmentation pipeline, incorporating steps to increase model robustness. Images are loaded with their annotations (bounding boxes and masks), randomly resized with scales from 0.1 to 2.0. Then randomly cropped, flipped to handle orientation variations, and padded to 640×640 pixels. This augmentation strategy introduces diverse transformations, improving the model's generalization.

3.2 Non-Maximum Suppression for Bounding Box Filtering

To refine our object detection results and prepare data for repair cost estimation, we apply the Non-Maximum Suppression (NMS) [8,9] algorithm. NMS reduces the number of overlapping bounding boxes by retaining only the most relevant ones, ensuring each detected object is represented by a single bounding box, thus enhancing the accuracy of subsequent cost estimation.

The NMS algorithm computes the Intersection over Union (IoU) between bounding boxes to measure overlap. It selects the box with the highest confidence score and suppresses all other boxes with an IoU greater than a specified threshold. This process is repeated until only the most significant bounding boxes remain (Algorithm 1). Figure 2 shows the NMS results on a sample input.

Algorithm 1 Non-Maximum Suppression (NMS)

```
function NonMaximumSuppression(boxes, scores, IoUThreshold)
    boxes ← {b_1, ..., b_N}                                   ▷ List of detection boxes
    scores ← {s_1, ..., s_N}                                  ▷ Corresponding detection scores
                  ▷ IoUThreshold: Maximum allowable overlap between bounding boxes
    idxs ← np.argsort(scores)[::-1]
    selected_idxs ← []
    while idxs.size > 0 do
        current_idx ← idxs[0]
        selected_idxs.append(current_idx)
        if idxs.size == 1 then
            break
        end if
        rest_idxs ← idxs[1:]
        rest_boxes ← boxes[rest_idxs]
        ious ← [ComputeIoU(boxes[current_idx], rest_boxes[i]) for i in range(len(rest_idxs))]
        idxs ← rest_idxs[ious < IoUThreshold]
    end while
    return selected_idxs
end function
```

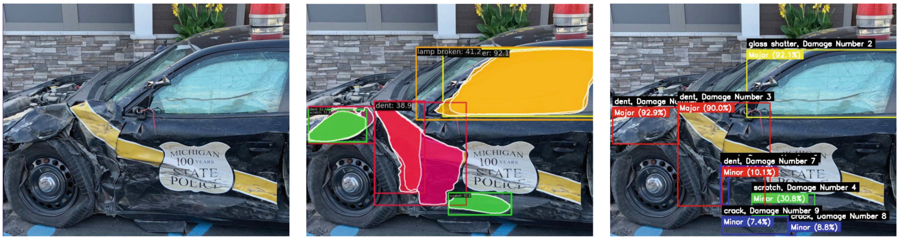
For bounding boxes boxA and boxB defined by corners (x_1, y_1, x_2, y_2):

$$\textbf{Intersection Area} = \max(0, \min(x_2^A, x_2^B) - \max(x_1^A, x_1^B)) \times \max(0, \min(y_2^A, y_2^B) - \max(y_1^A, y_1^B))$$

$$\textbf{Area of boxA} = (x_2^A - x_1^A) \times (y_2^A - y_1^A) \qquad \textbf{Area of boxB} = (x_2^B - x_1^B) \times (y_2^B - y_1^B)$$

$$\textbf{IoU} = \frac{\text{Intersection Area}}{\text{Area of boxA} + \text{Area of boxB} - \text{Intersection Area}} \tag{1}$$

– Area of Intersection is the overlapping area between two bounding boxes
– Area of Union is the total area covered by the two bounding boxes.



(a) Original image (b) Damage detection (c) NMS result

Fig. 2. Non-Maximum Suppression (NMS) results applied on a police vehicle

3.3 Estimating Repair Cost Using Detected Damages

To estimate repair costs for car damages, we employ an algorithm that processes detected bounding boxes, labels, and confidence scores. The methodology involves two primary steps: NMS (Sect. 3.2) and cost calculation.

NMS filters redundant bounding boxes to ensure each damage type is uniquely represented. Subsequently, the repair cost is computed based on predefined base costs for each damage type. This base cost is adjusted by a severity factor, which is derived from the area of the bounding box, the associated confidence score, and a normalization factor to adjust severity. The final repair cost is the sum of these adjusted costs. Repair cost is calculated by:

$$\text{Cost} = \text{Base Cost} \times \left(1 + \frac{(\text{Area} \times \text{Score})}{\text{Normalization Factor}}\right) \tag{2}$$

This approach ensures precise and reflective repair cost estimation based on the severity of detected damages, facilitating accurate repair assessments.

3.4 Salient Object Detection For Emergency Vehicle Damages

Salient Object Detection (SOD) enhances the detection of car damages by refining the boundaries of irregular and slender shapes. SOD focuses on locating all salient objects in an image, highlighting the most noticeable and severe areas of damage, and reducing noise by filtering out less relevant parts. This is useful in complex scenes where it is more important to identify key areas of damage than to classify them.

We apply a modified U-Net [19] model to refine the boundaries of detected damages. Using the CarDD [20] dataset, images are resized to 256×256 pixels and augmented with random resizing and flipping. The U-Net model was trained on an NVIDIA Tesla T4 GPU with a batch size of 32 for 250 epochs, using an Adam optimizer with a learning rate of 0.001 and weight decay of 1e-5. The loss function is Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss).

4 SAARTHI Framework

The SAARTHI framework (Fig. 3) integrates advanced AI techniques for vehicle damage assessment and repair coordination, streamlining the process from accident to repair completion. In this context, "users" refers to emergency vehicle drivers or fleet managers, depending on the organization's protocol.

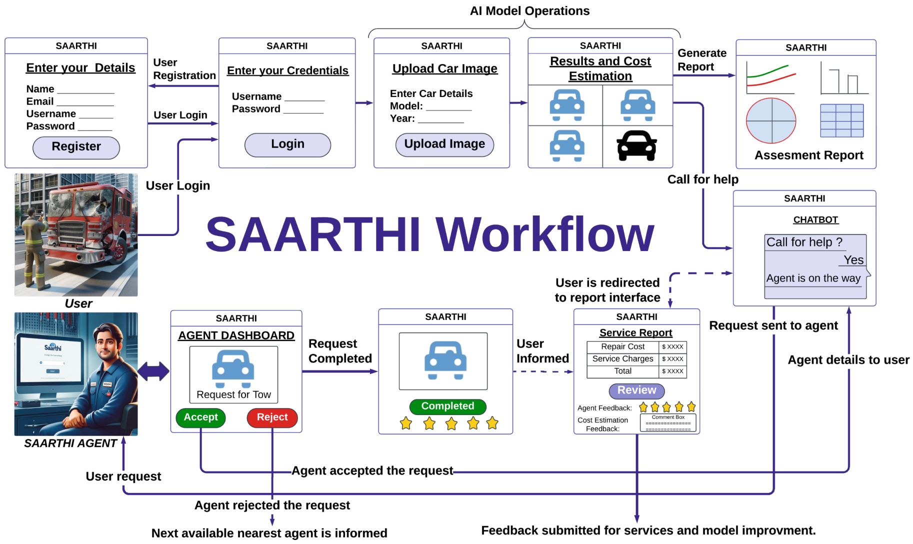


Fig. 3. SAARTHI Workflow. Please refer to Sects. 4.1–4.6 for more details

4.1 User Registration and Login

Users register on the SAARTHI platform. Once logged in, they can view their damage assessment history and past help requests on a personalized dashboard. Here they can also start a new assessment by uploading an image of the vehicle.

4.2 Image Upload and AI Operations

Users can upload an image of their damaged emergency vehicle along with optional details such as the car model and year. This image serves as the input for AI model operations, which include the following steps:

1. **Instance Segmentation and Object Detection** (Sect. 3.1): The uploaded image undergoes instance segmentation and object detection to identify and classify different types of damage on the vehicle, including dent, scratch, crack, lamp broken, glass shatter, and tire flat.

2. **Non-Maximum Suppression (NMS)** (Sect. 3.2): NMS is applied to eliminate redundant bounding boxes, ensuring each type of damage is represented only once.
3. **Cost Estimation** (Sect. 3.3): The system estimates the repair cost based on the identified damages, using predefined base costs adjusted by the area and confidence score of each detected damage.

The assessment allows users to view the original image, the damage-detected image, the image after applying NMS, and the result of SOD. The assessment report includes statistical graphs, providing a comprehensive understanding of the damages. The report also includes a table of estimated repair costs with labels for individual damages added in the total estimated repair cost.

4.3 Request Handling and Agent Coordination

Users can initiate a chatbot conversation for further assistance. Developed using JavaScript and the BotUI library [21], the chatbot offers options such as creating and downloading a PDF of the assessment report and connecting with the nearest SAARTHI agent based on the users location. Users can request a tow or on-the-spot repairs. The system uses integrated Google Maps API [22] to fetch the users location and create a request on the SAARTHI agent dashboard.

4.4 SAARTHI Agent

SAARTHI agents are emergency vehicle repair experts who have their own accounts on the SAARTHI portal. They are available 24/7 to provide assistance at any time of the day. Agents are registered on the portal by an administrator after verifying their identity and credentials. During registration, the agents details including shop location (latitude and longitude), city, phone number and personal informations are collected. Each agent is assigned a unique ID to facilitate tracking and coordination.

4.5 Nearest Agent Selection

When a user requests assistance, the system identifies the nearest available agent based on distance as follows:

1. **Calculate Distance**: Distance is calculated by Haversine formula [23,24]:

$$d = 2r \cdot \arctan 2\left(\sqrt{a}, \sqrt{1-a}\right) \tag{3}$$

 where $a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)$ and $\Delta\phi$ and $\Delta\lambda$ are the differences in latitude and longitude, and r is the Earth's radius.
2. **Sort Agents**: Agents are sorted based on their distance from the user. The closest agent is contacted first.
3. **Handle Availability**: If an agent rejects the request or is at full capacity, the system moves to the next closest agent.

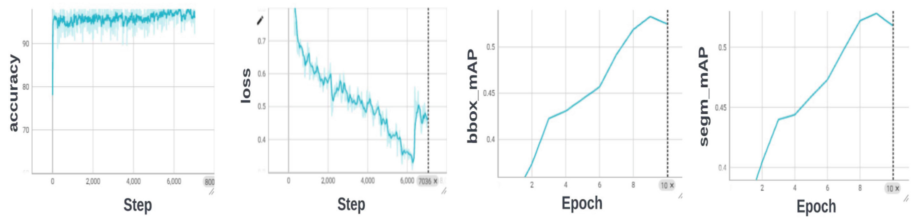
4.6 Report

Once the repair is completed, the user is notified and the detailed cost breakdown, including repair costs and service charges, is presented for review and then added to the monthly report of the user's organization.

5 Results

5.1 Object Detection and Instance Segmentation

For this research we used CarDD dataset [20]. Models were trained using an NVIDIA Tesla T4 GPU with a batch size of 4 for 10 epochs. The learning rate was set to 8e-05, using the Adam optimizer with a weight decay of 0.05. The loss functions included RPN classification loss, RPN bounding box regression loss, main classification loss (Cross-Entropy Loss), main bounding box regression loss (L1 Loss), and mask prediction loss (Mask Loss). The total loss, which is the sum of these individual losses, provided a comprehensive measure of the model's performance across different stages of object detection and segmentation. Figure 4 presents key metrics evaluating the model's performance per batch during training, with a batch size of 4.



(a) Accuracy (b) Loss (c) bbox mAP (d) seg. mAP

Fig. 4. Training Phase Result Metrics. Accuracy (Fig. 4a) shows a consistent upward trend, indicating improving model performance. The loss (Fig. 4b) declines steadily, reflecting increasingly accurate predictions and convergence. Bounding Box Mean Average Precision (bbox mAP) (Fig. 4c) indicates improving object detection performance, with steady increases in precision and recall. Segmentation Mean Average Precision (seg. mAP) (Fig. 4d) shows consistent improvement in identifying and segmenting objects

Testing Phase. For the testing phase of the SAARTHI framework we evaluated the model's performance using Mean Average Precision (mAP) and Intersection over Union (IoU) metrics. mAP measures the average precision across different classes, indicating better model accuracy with higher values. IoU measures the overlap between predicted and ground truth bounding boxes, with specific thresholds (e.g., 0.5, 0.75) determining correct predictions.

The overall bounding box mAP (bbox mAP) was 0.538 (Table 1), with high precision at IoU thresholds of 0.5 and 0.75. The model showed good accuracy for large, medium, and small objects. For segmentation masks (seg. mAP), the overall mAP was 0.519 (Table 2), also demonstrating high precision at IoU thresholds of 0.5 and 0.75, and good performance across different object sizes. Figures 6 and 8 illustrate the model's performance with examples of original images and damage detection outputs. These results indicate that the model performs well, particularly for larger objects and at an IoU threshold of 0.5. Figure 5 represents results of damage detection and segmentation for an input image.

Table 1. bbox mAP

mAP	Value
Overall	0.538
IoU 0.5	0.746
IoU 0.75	0.560
mAP Large	0.537
mAP Medium	0.309
mAP Small	0.314

Table 2. seg. mAP

mAP	Value
Overall	0.519
IoU 0.5	0.716
IoU 0.75	0.525
mAP Large	0.545
mAP Medium	0.236
mAP Small	0.171

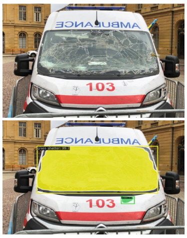
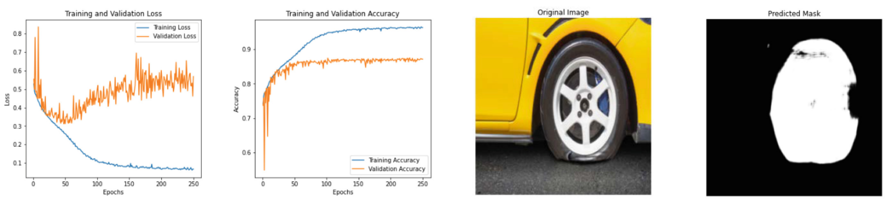


Fig. 5. Results

5.2 Salient Object Detection

Table 3 summarizes the model's performance for salient object detection during testing. The average loss of 0.5826 indicates a good match between predicted and actual values. The model achieves an accuracy of 0.8645, reflecting a high proportion of correct predictions. Precision and recall, both at 0.87, suggest the model effectively identifies true positives with balanced accuracy. The F1-score of 0.87 confirms the overall robust performance of the model in detecting salient objects. Figure 6 represents training phase results (loss, accuracy) and predicted segmentation mask for an input image.



(a) Training Loss (b) Training Accu. (c) Input Image (d) Predicted Mask

Fig. 6. SOD Training phase result 6a 6b and sample input image results 6c 6d

Table 3. Salient Object Detection results metrics (Testing Phase)

Average Loss	Accuracy	Precision	Recall	F1-score
0.5826	0.8645	0.87	0.87	0.87

5.3 SAARTHI User Interface

Figure 7 represents a few snippets of the end-to-end SAARTHI user interface. For more details, demo and code, please refer to sites.google.com/view/saarthi-home

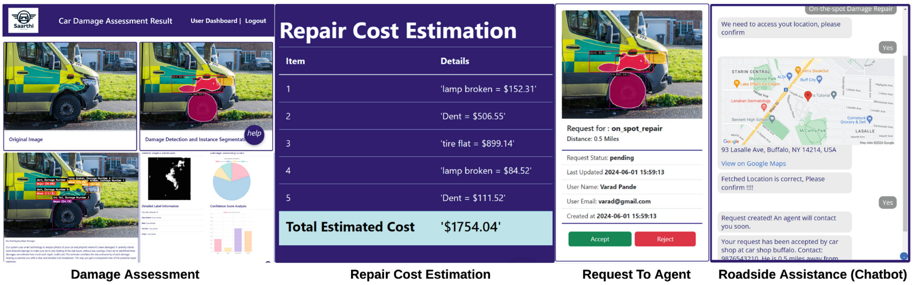


Fig. 7. SAARTHI User Interface

6 Conclusion

The SAARTHI framework effectively addresses the challenges of emergency vehicle accidents by utilizing advanced AI technologies for damage assessment, standardized repair cost estimations, and expedited vehicle recovery. This framework reduces downtime and enhances operational readiness through real-time assessment, detailed reporting, and chatbot assistance. While the focus is on external damages visible in images, future work will integrate telematics data and sensor readings for a holistic approach, including internal damages.

Integrating SAARTHI with IoT technologies for real-time monitoring and predictive maintenance will further improve emergency vehicle readiness and efficiency.

References

1. Sanddal, T., Sanddal, N., Ward, N., Stanley, L.: Ambulance crash characteristics in the us defined by the popular press: a retrospective analysis. Emerg. Med. Int. **2010**, 525979 (2010). https://doi.org/10.1155/2010/525979
2. Donoughe, K., Whitestone, J.J., Gabler, H.C.: Analysis of firetruck crashes and associated firefighter injuries in the united states. In: Annals of Advances in Automotive Medicine. Association for the Advancement of Automotive Medicine. Annual Scientific Conference, vol. 56, pp. 69–76 (2012). https://api.semanticscholar.org/CorpusID:17994834

3. Rivara, F., Mack, C.: Motor vehicle crash deaths related to police pursuits in the United States. Inj. Prev. **10**, 93–5 (2004). https://doi.org/10.1136/ip.2003.004853
4. Miller, T., Bhattacharya, S., Zaloshnja, E., Taylor, D., Bahar, G., David, J.: Costs of crashes to government, United States. In: 2008, Annals of Advances in Automotive Medicine/Annual Scientific Conference ... Association for the Advancement of Automotive Medicine. Association for the Advancement of Automotive Medicine. Scientific Conference, vol. 55, pp. 347–355 (2011)
5. He, K., Zhang, X., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2016). https://doi.org/10.1109/cvpr.2016.90
6. He, K., Gkioxari, G., DollÃr, P., Girshick, R.: Mask r-cnn (2017). arXiv:1703.06870
7. Chen, K., et al.: Mmdetection: open mmlab detection toolbox and benchmark (2019). arXiv:1906.07155
8. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms – improving object detection with one line of code. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE (2017). https://doi.org/10.1109/iccv.2017.593
9. Hosang, J., et al.: Learning non-maximum suppression. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2017). https://doi.org/10.1109/cvpr.2017.685
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation (2014). arXiv:1311.2524
11. Hafiz, A.M., Bhat, G.M.: A survey on instance segmentation: state of the art. Int. J. Multimedia Inf. Retr. **9**(3), 171–189 (2020). https://doi.org/10.1007/s13735-020-00195-x
12. Wang, X., et al.: Cardd: a new dataset for vision-based car damage detection. IEEE Trans. Intell. Transport. Syst. **24**(7), 7202–7214 (2023). https://doi.org/10.1109/tits.2023.3258480
13. Author, B.: Vehicle damage detection and analysis using convolutional neural networks. IEEE Trans. Intell. Transport. Syst. (2020). https://ieeexplore.ieee.org/document/9752971
14. Author, C.: Damage classification in vehicles using deep learning techniques. IEEE Trans. Veh. Technol. (2021). https://ieeexplore.ieee.org/document/10105039
15. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1704.04503 (2018)
16. Zhou, Q., et al.: An automatic surface defect inspection system for automobiles using machine vision methods. Sensors **19**, 644 (2019). https://doi.org/10.3390/s19030644
17. MMlabs and mmdetection contributors. MMdetection documentation (2024). https://mmdetection.readthedocs.io/en/latest/
18. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Wang, X., Li, W., Wu, Z.: CarDD dataset. https://drive.google.com/file/d/1bbyqVCKZX5Ur5Zg-uKj0jD0maWAVeOLx/view?pli=1
21. BotUI contributors, BotUI—A JavaScript framework to create conversational UIs (2024). https://botui.org/docs

22. Google Developers, Google Maps API (2024). Accessed 28 May 2024. https://developers.google.com/maps/documentation
23. Robusto, C.C.: Haversine formula. Am. Math. Monthly **64**(1), 57–59 (1957)
24. Nichat, M.: Landmark based shortest path detection by using dijkestra algorithm and haversine formula. Int. J. Eng. Res. Appl. (IJERA) **3**(3), 162 (2013). ISSN: 2248-9622

Comparing Similarity and Homophily-Based Cognitive Models of Influence and Conformity

Robert Thomson[1(✉)] and Christian Lebiere[2]

[1] United States Military Academy, West Point, NY 10996, USA
robert.thomson@westpoint.edu
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA
cl@cmu.edu

Abstract. The majority of theories and models of social influence tend to focus on the social-behavioral level and implicitly discount the potential role of cognitive explanations to ground out social phenomena in cognitive operations. The present study describes a preliminary simulation of social influence and conformity using two possible cognitive mechanisms: the first a homophily-based model that weighs belief updating based on generating a latent trust magnitude, and the second is a novel similarity-learning mechanism that weighs memory retrieval via similarity. While the homophily model was able to capture both influence and conformity effects based on degree of initial strength in the belief and crystallization of the belief, the similarity-based model always conformed. Both models exhibited initial learning and larger belief updating while settling to a relatively-stable state over time. Future implications for modeling influence via cognitive factors are discussed.

Keywords: cognitive modeling · similarity · homophily · influence

1 Background

Over the past 40 years in the United States, there appears to be a societal shift towards more polarized viewpoints and less discourse on substantive topics. This tendency has been further exacerbated by the rise of internet chat forums and social media, where both natural and algorithmically-manipulated *echo chambers* [6] (for a counter-opinion, see [9]) provide venues for like-minded people to strength (i.e., to *confirm*) their preexisting beliefs. The preponderance of research into this behavior has focused on higher-level social and behavioral theories, without as much focus on the underlying cognitive mechanisms which may drive this behavior. Many theories posit dual-process or multi-process behavioral models (e.g., System 1 vs System 2[16]; Elaboration-Likelihood [23]) which may be seen as false dichotomies [25] when viewed from the lens of cognitive operations. The goal of the present paper is to two-fold: (1) is to describe how belief

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 47–57, 2024.
https://doi.org/10.1007/978-3-031-72241-7_5

updating occurs from a cognitive perspective, and (2) is to present two cognitive models using separate mechanisms (similarity learning vs homophily) to describe the process of belief change as applied to a synthetic social influence dataset.

The paper will proceed as follows: we will first describe several relevant theories of social influence and belief change, and show how they may be modeled from various perspectives. We then provide a description of how cognitive models store, retrieve, and revise beliefs, as well as describe two possible modeling paradigms to understand how latent factors (such as *trust*) can be modeled and how they impact belief updating. This study does not seek to define trust, but to use the general concept of *trust* as a construct which includes respect and credibility in the trustee. Finally we describe a synthetic social influence dataset based on the plot of Shakespeare's Othello, to see how the model's opinion on a given topic changes based on the dynamics of incoming messages from various sources. Finally, implications for extending these models to larger and real-world datasets and similar domains (e.g., cognitive biases) will be discussed.

1.1 Theories of Belief Change

While there has been a preponderance of social psychological research on the role of beliefs (as well as attitudes and social norms) on behavior change (e.g., the Theory of Planned Behavior; Triandis model; and Schwartz model; for an overview see [3]) the focus of this paper is targeted to understanding the cognitive underpinnings of belief change regardless of concomitant changes in performance. For instance, changes in performance may be due to belief revision, but may also be the result of transformations of the underlying mental models/schema or categorical/normative shift in the underlying belief concepts [5]. For instance, people's shifting perception on what constitutes the scientific consensus on a topic can influence their support for public action on said topic. As such, we will focus on specific cognitive factors that influence belief change and how they may be modeled by a cognitive architecture.

Perhaps one of the most studied factors in belief change is that of similarity. Homophily is the effect where the influence of another person on your beliefs is based on how similar their beliefs are to your own [8,10]. In other words, the more a person's beliefs are perceived as similar to your own, the more you will tend to **trust** them and the more their beliefs tend to influence you. Furthermore, authoritative sources are more likely to influence those whose beliefs have less experience or whose beliefs have not crystallized [24], with more credible sources and more plausible explanations being more influential [20]. Numerous other social and personality factors have also been associated as mediators of belief change (see [7] for a recent review).

Prior modeling efforts are primarily network models of homophily which use similarity, and often a hard-coded form of confirmation bias [8] to weight the model's initial belief/opinion, and then to weight belief updating based on some measure of similarity between positions [11]. These simulations are often used to show the time-course of polarization in a population, which tends to fractionate into subgroups of similar-minded members. While interesting, this paper focuses on the cognitive underpinnings of this belief change. One advantage of modeling cognition is that we can introspect into the model to see how beliefs update in the model's *mind* before they are expressed as behavior change.

1.2 Cognitive Explanations of Belief Change

Cognitive psychology focuses on memory processes to explain how information is processed, stored, and retrieved. This is often described in terms of the recency and frequency of the information being experienced, the similarity and spread from prior stored knowledge, and order effects based on the structure of the task and environment [2,27]. Cognitive biases can be explained as being the result of the interplay between memory processes, task requirements, and the interaction between these two. Similar effects have already been seen in the belief change literature, with recency and frequency of interaction influencing the degree of belief change [4]. In fact, the patterns seen in belief change were similar to the patterns of category change seen in a novel category learning task [12].

While traditional fact-based memories decay, strength of belief is less likely to decay over time, but instead gets revised based on new information [14,28] (which is consistent with theories of interference-based 'decay' [26]). This may be due in part to the fact that the premises which support (or reject) the current belief are decaying at similar rates, limiting the impact of temporal decay. Of note, an existing framework exists which is consistent with these features. The Knowledge Revision Component Framework (KReC) [17] assumes that information in long-term memory is permanent. This framework assumes that prior beliefs often may be re-activated when being presented with updated knowledge and hinder the acceptance of this new information. For belief change to occur, it is essential to activate the 'correct' belief whenever the 'incorrect' belief is activated.

1.3 Cognitive Models of Decision-Making

The present research uses the dynamics of memory retrieval from the ACT-R cognitive architecture [1,27]. ACT-R is an integrated theory of memory that models decision-making using a hybrid symbolic-subsymbolic architecture. At the symbolic level, memories are stored as instances (also called chunks) consisting of unordered *slot-value* pairs of information where the *slot* is like an attribute and the *value* is the representation (i.e., content) of the attribute. The subsymbolic level associates numeric strengths of activation to symbolic elements, with activation dynamics driving what gets recalled during memory retrieval. ACT-R models have been used to explain complex decision-making tasks and cognitive biases across numerous domains (for a review of the theory and detailed explanation of the mechanisms, see [27]).

Memory retrieval occurs when the most active instance in memory is selected through an explicit request which queries declarative memory for a set of slots and values. A memory's activation consists of a summation of its base-level activation $B(m)$, mismatch penalty to the current retrieval specification $\sum_f P \times \delta(f, m)$ and momentary noise θ (see Eq. 1). The odds of retrieving m

increase linearly with the number of associated traces, while the probability of retrieving a memory declines over time according to a power function. Similarity is implemented in the form of a mismatch penalty δ between the value of a feature specified at retrieval and the actual feature value; and represents the dissimilarity between the representation of two values on a scale $-1 < \delta < 0$. For numeric values between 0 and 1, often a linear or ratio similarity (i.e., .1 to .2 is the same mismatch as .4 is to .8) are used, while for unbounded numbers often a log- similarity is most appropriate.

$$A(m) = B(m) + \sum_{f} P \times \delta(f, m) + \theta \tag{1}$$

While retrievals will return the most active memory, ACT-R allows for a special mechanisms known as *blending* [18] to retrieve memories that are an interpolated mixture of features in declarative memory while not necessarily being identical to any of them (see Eq. 2).

$$V = \underset{V_t}{\operatorname{argmin}} \sum_{i=1}^{n} P_i \cdot Sim(V_t, v_{it})^2 \tag{2}$$

In the simplest case where the values are numerical and the similarity function is linear the process simplifies to a weighted average by probability of retrieval. That said, the blending process is still sensitive to the recency, frequency, and order of incoming information and how the retrieval specification matches to prior instances stored in memory. In the case of both standard retrievals and blending, the retrieval specification (i.e., what slots and values are specified) and choice of what to store in memory greatly impacts overall model performance and those aspects can be seen as a kind of *learning strategy* which can be evaluated.

2 Current Study

To compare similarity and trust-based cognitive models, a synthetic dataset was developed in which each model consumes messages from a variety of sources with differing proportions of belief agreement (i.e., the proportion of messages supporting the model's initial belief). The dataset was inspired by previous research on influence and conformity [4,21,28], which shows that the presence of a confederate may impact whether a target is likely to conform to a given (incorrect) belief. To avoid stigma surrounding affective topics, the present modeling effort is called *Othello* and names match the plot of Shakespeare's play. In the play, Othello is deceived by his trusted advisor Iago into believing that his wife Desdemona is having an affair. In the play, most people are deceived by Iago with the exception of her handmaid Emilia who remains true to Desdemona.

2.1 Synthetic Dataset Structure

The structure of the data has five features for each message: a time, a source (*messenger*), a destination (*recipient*), a topic, and a stance (*pro or anti*). To constrain the present study, there is only one recipient (Othello) and one topic (Desdemona). The remaining structure is as follows: the data consists of 1800 messages across 9 different messengers (200 messages each from). Each messenger has a set proportion of 'pro' stances (Iago 10%, Roderigo 20%, Cassio 20%, Montano 30%, Lodovico 40%, Brabantio 40%, Gratiano 50%, Bianca 50%, and Emilia 90%). Each 'trial', the model processes a single message from a single source. In this study, the message order was randomized so there is no explicit bias or stance-switching within a given messenger. As can be seen, the messages are biased against Desdemona, with only Emilia consistently supporting her. The initial stance of Othello will be presented in each model's description below.

2.2 Present Models

Trust-Based Model. The trust-based model is written using PyACTUp, a lightweight and scalable modeling framework based on the ACT-R cognitive architecture [1], encompassing its core memory retrieval functionality without committing to the complete architectural theory of how information is processed between various brain regions (*buffers* in ACT-R parlance). This allows for rapid prototyping of memory-based models that requires no particular commitment to specific sensory and procedural processes. This model utilizes homophily-inspired trust metric that weights how belief strength is updated (see Eq. 3).

$$B_{(t)} = \frac{B_{(t-1)} + U_{(t)}}{2} \; where \;\; U_{(t)} = \left(M_{(t)} - B_{(t-1)}\right) * {T_{(t)}}^{w} \tag{3}$$

The current belief $B_{(t)}$ is based on average of the prior belief $B_{(t-1)}$ and belief update factor $U_{(t)}$, where the belief update factor is based on the difference between the message strength $M_{(t)}$ and prior belief multiplied by the weighted trust ${T_{(t)}}^{w}$. The trust weight w is a parameter which should be set by human data (when such data is available), and in the present model is set to 2 to provide non-linear dynamics (trust is hard to gain and easy to lose). Averaging the current belief and belief update factor provides a natural exponential curve which can be biased by the current trust in the messenger.

The basic model functionality is a message receiving and processing loop (see Algorithm 1), where the model *receives* a message, determines its internal position on the message's topic, determines its trust in the messenger, and then updates its trust in the messenger and belief in the topic. This process iterates and the dynamics of belief and trust can be evaluated over time and as more messages are received. Beliefs and trust are both messenger- and topic- specific, so the model can in theory trust a messenger on a specific topic (e.g., in one they have authority in) without necessarily trusting them in another, although due to the model's mismatch penalty there can be some bleed over between the opinions of other messengers and other topics on the model's current decision-making.

In terms of parameters and other model settings, both the belief and trust values scale from 0 to 1. Mismatch penalties are set to linear similarities for numeric values (belief and trust), and to the maximal penalty (with a weight = 5) for non-numeric values (different messengers and topics).

Algorithm 1. Algorithmic Flow of the Trust-Based Cognitive Model

```
procedure INGEST(M)                                          ▷ Ingest Message
    BLEND B_(t−1) for Topic t and Stance s                   ▷ Get Prior Belief
    if Message Stance == Topic Stance then
        TrustFactor T = 1
    else
        TrustFactor T = 0        ▷ Determine Whether Model Stance is Similar
    end if
    LEARN M                                          ▷ Store Message in Memory
    BLEND T_(t)                                          ▷ Get Current Trust
    LEARN T                                              ▷ Store Trust Factor
    UPDATE B_(t)                      ▷ Update Belief According to Equation 3
    LEARN B_(t)                          ▷ Store the Updated Belief in Topic s
end procedure
```

Trust is encoded and retrieved entirely by the default blending dynamics. When a message is received the stance of the message is compared to the model's current stance. When stances are similar an instance in memory is stored with the messenger, topic, and trust = 1; while stances that are dissimilar are stored with trust = 0. This allows for power law of learning as well as local biases due to recency, frequency, and order effects [19]. Current belief is retrieved by blending and updated based on Eq. 3: changes in belief strength are weighted by the degree of trust in the messenger and how well it accords with our current beliefs. This also provides a cognitive explanation of confirmation bias.

Similarity-Based Model. The similarity-based model is written in ACT-UP, a similar implementation to PyACTUp in the Lisp programming language. This model conceives of trust as the similarity between a message recipient and the messenger. The assumption is that one tends to trust people viewed as similar to ourselves, and distrust people viewed as dissimilar. This view maps straightforwardly to the common concepts of in-group and out-group. The difference is that similarities are a basic cognitive concept impacting many mechanisms of the cognitive architecture such as partial matching and blending that control access to declarative information and form the basis of intuitive judgments [13].

Up to now, similarities between concepts were set in ACT-R by the modeler. Similarities between quantities were usually set according to regular formulas such as linear similarities for probabilities and log-ratio similarities for magnitudes [19]. Similarities between abstract concepts could be set to reflect their overlap in ontologies [22] or in the distributed representations in neural networks [15]. Here, we propose a new mechanism for learning similarities based on patterns in knowledge structures. The intuition is that the similarity between values in the corresponding slots of two chunks reflects the similarity between values in the other slots. The idea is that our cognitive systems seeks to maximize the coherence, congruence or consistency of our knowledge representation.

As for other learning mechanisms of the ACT-R cognitive architecture such as associative learning, the learning of similarities will follow a Bayesian approach of starting with a prior value that is updated from experience. Formally, if chunks C and D include two slots containing values $< X,A >$ and $< Y,B >$ respectively, then the similarity S_{XY} between values X and Y can be learned from the similarity S_{AB} between values A and B according to the following equations:

$$S'_{XY} = \frac{\omega_{XY} S_{XY} + p_C p_D S_{AB}}{\omega_{XY} + p_C p_D} \; and \; \omega'_{XY} = \omega_{XY} + p_C p_D \quad (4)$$

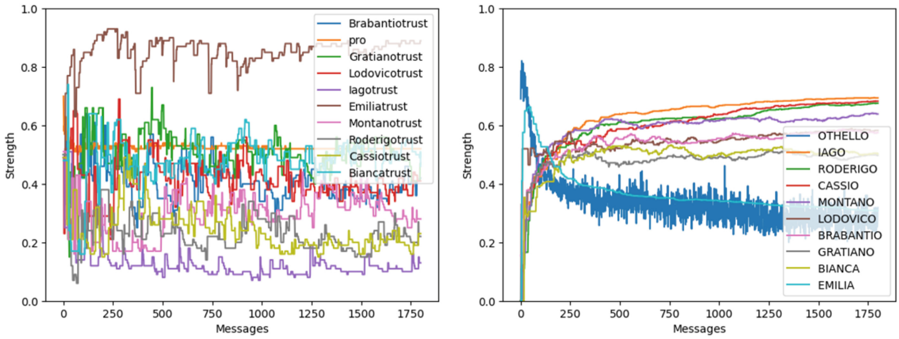
when ω_{XY} is the weight of S_{XY} and p_C and p_D are the probabilities of chunks C and D, respectively, as determined by their activation reflecting factors such as recency, frequency and match to context, using a Boltzmann distribution.

The similarity-based model is otherwise quite similar to the trust-based model. Beliefs are represented by chunks containing two slots: the source and the belief. The source is either self if resulting from the model's own stance computation, or some external source if the result of a communication. The belief is represented on a scale of $[0, 1]$ corresponding to the anti-pro spectrum. For each new message, the model first advanced the time then generates its own stance by blending over all belief chunks, with a requirement that its source match the self. This is the key step where learned similarities between sources, specifically itself and other external sources, impact the generation of the model's own stance. The model then learns its own stance as well as the received belief from the other source, then updates the similarity between itself and the source according to the equation above applied to the two new chunks that were just learned.

Parameters are left to their default values used in previous models: the activation noise is 0.25 while the blending temperature is 0.5 and the similarities between beliefs is a linear function. For similarity learning, the initial similarity is –1, which is the maximum difference, and the initial weight is 1.

2.3 Results and Discussion

Both models ran multiple simulations on varying degrees of initial belief strength (how close to 1 their initial belief state is from .7 to 1.0 in .1 increments) and the homophily model also investigated how prior experience (e.g., how crystallized the initial belief is between no-experience to experienced). As seen in Figs. 1a and 1b, at the .7 initial belief state the homophily model is still resilient to the influence of the negative and ambivalent opinions, while the similarity-based model is not. In both cases, the model dynamics tend to settle after several hundred trials (see Table 1), which is consistent with the structure of the dataset in that the dynamics of the trials do not change over time. If this were extended to a multi-agent system, then how the models flip would have downstream emergent properties on the overall beliefs of the various agents. In this case, as the incoming messages do not change, it is appropriate for the model's subsymbolic dynamics to settle once sufficient experience is gained.



(a) **Homophily Model.** Lowered trust from the initial disagreement non-Emilia sources makes the model resilient to flipping. The only occurs when the model's initial belief strength is very high or the model's belief has crystallized.

(b) **Similarity Model.** Lower similarity reduces the relative influence of those sources but the sheer number of trials slowly pulls the models towards the opposite stance in all conditions. This is due the volume of counter-messages.

Fig. 1. Model output with 70% initial belief. *'Pro'* position is Othello's trust in the Homophily Model and is shown as *'Othello'* in the Similarity Model.

Table 1. Runs labeled **H** are from the homophily model while runs prefaced with **S** are from the similarity model. **NE** means that the model is not experienced. The not experienced model from .7-.9 flips belief while the others do not.

Name	Trust	Iago	Rode	Cass	Mont	Lodo	Brab	Grat	Bianca	Emilia
S.7	.31	.69	.67	.68	.63	.58	.57	.49	.50	.30
S.8	.22	.69	.66	.67	.63	.57	.57	.49	.49	.32
S.9.	.28	.68	.66	.66	.62	.56	.57	.49	.49	.31
S1.	.25	.67	.66	.66	.62	.57	.56	.50	.49	.32
H.7	.52	.13	.22	.23	.28	.41	.42	.47	.46	.89
H.8	.54	.13	.22	.23	.28	.41	.41	.47	.46	.89
H.9	.55	.13	.22	.23	.28	.41	.41	.47	.46	.89
H1.0	.56	.13	.22	.23	.28	.41	.41	.47	.46	.89
H.7NE	.36	.86	.78	.77	.72	.59	.58	.52	.54	.12
H.8NE	.38	.86	.77	.77	.72	.59	.57	.53	.53	.12
H.9NE	.36	.86	.78	.77	.72	.59	.58	.53	.54	.12
H1.NE	.54	.13	.22	.23	.28	.41	.42	.47	.46	.89

In terms of the learning of similarities in the similarity-based model, they reflect the compatibility between stances of the model and the messages received from the various sources. The similarity to Emilia is particularly interesting since it illustrates the dynamics of the similarity learning process. Initially, since the model is initialized with a strong pro stance similar to Emilia, the similarity between the two jumps to become the strongest while the similarities to the other sources start quite low. However, as the model stance quickly becomes more anti reflecting the strong pattern of received messages, the similarity to Emilia quickly decreases to become the lowest of all while the similarity to the other sources gradually rises to reflect the degree of their anti stance.

Conversely, in the homophily-based model, the lower trust reduces the relative contribution of those messages to the point that, if the model either (1) has experience with prior belief (represented by reinforcing the initial position in the past), or (2) has a very high *'pro'* initial belief to begin with, then it is resistant to conforming to the overall group position. That said, with enough consistent counter-messaging from sufficient sources (without a confirming message source), it would be possible to flip even an experienced model.

3 Conclusion and Future Plans

The present study simulated the effects of influence and persuasion on two different cognitive models: one based on trust-weighted belief updating and the other based on similarity. While the trust-weighted belief model displayed resilience to conformity based on prior belief strength and prior experience similar to that seen in previous literature, the similarity-based model was influenced by the imbalanced negative messages and conformed over time. Of interest, both models settled into similar dynamics after several hundred trials. The homophily-based trust model explicitly retrieved trust as a latent variable and weighted the degree of belief updating based on the current trust. By embedding this theory in a cognitive model, the results naturally encompass local deviations due to frequency, recency, and order of incoming information.

Future work will focus on more systematic investigation of the models and extending them into multi-agent simulations to investigate emergent social dynamics such as those seen in network models of social influence.

Acknowledgements. This research was supported by the Office of Naval Research MURI Award N0001422MP00465 and Award 10407118530 from the DoD Basic Research Office. The views expressed in this work are those of the authors and do not necessarily reflect the official policy or position of the United States Military Academy, U.S Army, Office of Naval Research, or U.S. Government.

References

1. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychol. Rev. **111**(4), 1036 (2004)
2. Anderson, J.R., Schooler, L.J.: Reflections of the environment in memory. Psychol. Sci. **2**(6), 396–408 (1991)
3. Bamberg, S., Schmidt, P.: Incentives, morality, or habit? predicting students' car use for university routes with the models of Ajzen, Schwartz, and Triandis. Environ. Behav. **35**(2), 264–285 (2003)
4. Brechwald, W., Prinstein, M.: Beyond homophily: a decade of advances in understanding influence. J. Res. Adolesc. **21**(1), 166–179 (2011)
5. Chi, M.T.: Three types of conceptual change: belief revision, mental model transformation, and categorical shift. In: International Handbook of Research on Conceptual Change, pp. 89–110. Routledge (2009)
6. Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proc. Natl. Acad. Sci. **118**(9), e2023301118 (2021)
7. Cusano, L.C.: Thinking styles, conspiracist belief, and the mediating role of the dunning-kruger effect in modeling belief. Ph.D. thesis, San Jose State (2024)
8. Dandekar, P., Goel, A., Lee, D.T.: Biased assimilation, homophily, and the dynamics of polarization. Proc. Natl. Acad. Sci. **110**(15), 5791–5796 (2013)
9. Dubois, E., Blank, G.: The echo chamber is overstated: the moderating effect of political interest and diverse media. Inf. Commun. Soc. **21**(5), 729–745 (2018)
10. Ertug, G., Brennecke, J., Kovács, B., Zou, T.: What does homophily do? a review of the consequences of homophily. Acad. Manag. **16**(1), 38–69 (2022)
11. Frigo, V., Chen, L., Rogers, T.: A cognitive mechanism for the persistence of widespread false beliefs. Technical report, PsyArXiv (2020)
12. Frigo, V., Rogers, T.T.: Evidence for heuristic evidence weighting in real-world beliefs. In: Annual Meeting of the Cognitive Science Society, vol. 45 (2023)
13. Gonzalez, C., Lerch, J.F., Lebiere, C.: Instance-based learning in dynamic decision making. Cogn. Sci. **27**(4), 591–635 (2003)
14. Gupta, S., Karduni, A., Wall, E.: Belief decay or persistence? a mixed-method study on belief movement over time. Comput. Graph. **42**(3), 111–122 (2023)
15. Jilk, D., Lebiere, C., O'Reilly, R., Anderson, J.: Sal: an explicitly pluralistic cognitive architecture. Exp. Theor. AI **20**(3), 197–218 (2008)
16. Kahneman, D.: Thinking, fast and slow. Macmillan (2011)
17. Kendeou, P.: A theory of knowledge revision: the development of the krec framework. Educ. Psychol. Rev. **36**(2), 44 (2024)
18. Lebiere, C.: Blending: an act-r mechanism for aggregate retrievals. In: Proceedings of the Sixth Annual ACT-R Workshop (1999)
19. Lebiere, C., et al.: A functional model of sensemaking in a neurocognitive architecture. Comput. Intell. Neurosci. **2013**, 5–5 (2013)
20. Lombardi, D., Nussbaum, E.M., Sinatra, G.M.: Plausibility judgments in conceptual change and epistemic cognition. Educ. Psychol. **51**(1), 35–56 (2016)
21. Mei, W., Cisneros-Velarde, P., Chen, G., Friedkin, N.E., Bullo, F.: Dynamic social balance and convergent appraisals via homophily and influence mechanisms. Automatica **110**, 108580 (2019)
22. Oltramari, A., Lebiere, C.: Using ontologies in a cognitive-grounded system: automatic action recognition in video surveillance. In: 7th International Conference on Semantic Technology for Intelligence, Defense, and Security (2012)

23. Petty, R.E., Cacioppo, J.T., Petty, R.E., Cacioppo, J.T.: The Elaboration Likelihood Model of Persuasion. Springer, Heidelberg (1986). https://doi.org/10.1007/978-1-4612-4964-1_1
24. Popp, D., Laursen, B., Kerr, M., Stattin, H., Burk, W.K.: Modeling homophily over time with an actor-partner interdependence model. Dev. Psychol. **44**(4), 1028 (2008)
25. Stanovich, K.E., West, R.F.: Advancing the rationality debate. Behav. Brain Sci. **23**(5), 701–717 (2000)
26. Thomson, R., Harrison, A., Trafton, G., Hiatt, L.: An account of interference in associative memory: learning the fan effect. TopiCS **9**(1), 69–82 (2017)
27. Thomson, R., Lebiere, C., Anderson, J.R., Staszewski, J.: A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. J. Appl. Res. Mem. Cogn. **4**(3), 180–190 (2015)
28. Wyer, R.S., Albarracin, D.: Belief formation, organization, and change: cognitive and motivational influences. Handb. Attit. **273**, 322 (2005)

Developing Epidemiological Models with Differentiated Infected Intensity

Niloofar Yousefi(✉), Nitin Agarwal, and Emmanuel Addai

COSMOS Research Center, University of Arkansas - Little Rock, Little Rock, Arkansas, USA
{nyousefi,nxagarwal,eaddai1}@ualr.edu

Abstract. This study investigates the spread of toxic content on social media, a growing concern as online platforms serve as primary information sources. This research aims to enhance the accuracy of models capturing toxicity spread by differentiating between varying levels of toxicity intensity. Two epidemiological models are developed and assessed: the $SEIRS$ model and a novel SEI_mI_hRS model. The latter divides infected users into moderate and highly infected groups to reflect the varying severity of toxic behavior. Both models are tested on six datasets to evaluate their performance. The SEI_mI_hRS model achieves even lower error rates, indicating a more precise representation of toxicity propagation. This research contributes a sophisticated tool for analyzing online toxicity, aiding policymakers and online platforms in developing targeted interventions and enhancing content moderation systems.

Keywords: Hate speech · Toxicity · Epidemiological models · SEIRS · Social media · Propagation

1 Introduction

The internet usage among people is increasing, with many relying on it to access information. In some cases, social media information affects user behavior, emotions, misinformation, and more [1–3]. Unfortunately, hate speech on social media is prevalent, especially during events like political chaos or disease outbreaks, where people express their anger by sharing hateful sentiments online

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-23-1-0011, W911NF-24-1-0078), U.S. Office of Naval Research (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency, Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 58–68, 2024.
https://doi.org/10.1007/978-3-031-72241-7_6

[4–6]. Understanding the spread of harmful content is essential for addressing its negative consequences as online information sharing grows. Thus, researchers have employed various methods to investigate the spread of toxicity on social media [7]. Some researchers have applied epidemiological modeling, traditionally used for disease spread, to understand toxicity propagation online. By treating a toxic post as a "disease" they analyze its dissemination using these models.

In this study, we assess an epidemiological model termed SEIRS (Susceptible-Exposed-Infected-Recovered-Susceptible). Subsequently, we develop a novel model that incorporates various infected groups, namely, moderate and high infected users in the SEI_mI_hRS (Susceptible-Exposed-Moderate-High Infected-Recovered-Susceptible) model. We divide the Infected compartment into two distinct states: high infected and moderate infected. This division accounts for the varying toxicity levels, acknowledging that the intensity of toxicity can impact the propagation dynamics within the community. Not all instances of infection carry the same level of toxicity; some users show high-intensity toxic behavior, while others display moderate toxicity. By categorizing infected users by toxicity intensity, we can better capture the nuances of how toxicity spreads in the community. This distinction reflects how spread dynamics and impact vary with the severity of the content. By implementing the two models on multiple datasets, our objective is to address the following research question: **RQ:** Will dividing the infected compartment based on varying levels of toxicity intensity (moderate and high) enhance the accuracy of modeling the dissemination of toxicity?

2 Related Work

In order to address the spread of harmful content on social media platforms, Some researchers employed machine learning techniques to investigate the patterns of communication on Reddit and investigate how toxic comments influence propagation [8,9]. DeMarsico et al. [10] developed and validated a tool for measuring online aggression, utilizing a randomly selected Twitter dataset. Furthermore, [11] explored the implications of toxic content on user engagement and retention within online communities. In another research, [12] employs techniques for analyzing and mitigating toxic content on YouTube during the early COVID-19 pandemic by using topic modeling, social network analysis.

2.1 Epidemiological Modeling

A range of studies have applied epidemiological models to understand the propagation of toxicity in social media. The authors in [13] apply SIR and extended SIR models to data on conspiracy theories. In some studies, researchers use epidemiological models such as SEIZ to analyze the spread of toxicity in Twitter (X) [14,15]. In another research, [16] compared various epidemiological models to analyze toxicity, finding that the STRS model best captures toxicity spread. In [17] an epidemic model was used to analyze information cascades, accurately capturing the spread of news/rumors concerning events.

In research by [18] a modified SEIR model was introduced to simulate the COVID-19 epidemic in China, demonstrating that the model effectively predicts epidemic peaks and sizes. Research on rumor spread using the SEIR model reveals that online social networks with changing user populations can accurately depict the characteristics of rumor dissemination [19], and incorporating social media measures into SEIR models enhances the accuracy of infection forecasts [20,21].

One study used an improved SEIRS model to analyze the spread of rumors about agricultural product safety [22]. Another study developed the SEIRS-C model, incorporating a rumor delay mechanism with fuzzy rules to slow down rumor spread [23]. In [24], the authors claimed that fractional SEIRS models effectively fit real data, such as the spread of information from the 2018 college entrance exam in China. In [25], the SEIRS model was applied to understand IoT botnet spreads, revealing that improved user information could mitigate the frequency of IoT botnet attacks. Additionally, [26] used the SEIRS model to investigate community misinformation, emphasizing its utility in depicting dynamic rumor spread stages.

No other researchers have studied the propagation of toxicity by dividing the Infected compartment into two states for high and moderate toxicity, which we investigate in this research. This approach will help researchers and policymakers determine where they should allocate their resources and implement interventions to get more accurate results.

3 Methodology

This section overviews the data collection and the epidemiological models employed.

3.1 Data Collection

We leveraged the Twitter Academic APIs to collect tweets, focusing on hashtags related to anti-COVID-19-related topics and social movements. After identifying prevalent hashtags, we selected relevant ones for data collection. Tweets flagged for misinformation and subsequently removed from X were not included in our dataset. Our dataset encompasses original tweets, and retweets, with no language restrictions.

COVID-19 Related Data: The COVID-19 pandemic sparked intense debates on COVID-19, vaccines, face masks, and lockdowns. The data spans from February 2020 to June 2021, capturing the peak periods of these discussions. Data was collected using various hashtags covering a range of topics, including #f*ckcovid, #f*cklockdown/s, #f*ckyourmask/s, #f*ckvaccine/s, etc.

Social Movement Related Data: We analyzed two social movements during Brazil's unrest from November 1, 2022, to February 25, 2023. Anti-government protests over electoral fraud after the elections led to riots. Pro-Bolsonaro supporters used hashtags like #Brazilwasstolen and #brazilelectionfraud. In response, Pro-Government Protests, advocating for retribution against rioters, used hashtags like #semanistia and #brazilcapitolriots. Table 1 shows the statistical information regarding datasets.

Table 1. Toxicity dataset statistics

Dataset	Average of Toxicity	No. Toxic posts	No. High Toxic	No. Moderate Toxic	Dataset Length
F*Covid	0.91	6,766	4,684	2,082	28,131
F*Mask	0.91	755	538	217	2,423
F*Vaccine	0.89	189	128	61	610
F*Lockdown	0.82	1,091	598	493	1,995
Brazil Anti	0.70	3,530	1,221	2,309	405,160
Brazil Pro	0.75	236	105	131	44,415

3.2 Toxicity Detection

We utilize Detoxify, developed by Unitary AI, to compute toxicity scores for tweets in multiple languages [27]. Detoxify assigns a probability score between 0 and 1, indicating the likelihood of toxic language, with a threshold of 0.5 used to distinguish toxic from non-toxic posts [28]. To analyze high and moderate toxicity spread, we introduce another threshold. Posts with a toxicity score above the dataset's average are classified as high toxic, and scores below the average are considered moderate toxic (see Table 1).

3.3 Model Formulation

Epidemiological models are used to investigate the spread of toxicity on social media, introducing the $SEIRS$ and SEI_mI_hRS models for this purpose. The primary difference between the $SEIRS$ and SEI_mI_hRS models is the division of the Infected compartment into two states. The $SEIRS$ model has a single infected (I) compartment, representing users who post toxic content. In the SEI_mI_hRS model, we introduce two separate infected (I) compartments: moderate-intensity toxic posts and high-intensity toxic posts. This differentiation provides a more granular perspective on the escalation of toxic behavior, recognizing that not all toxic engagements are equal.

By comparing the results of the $SEIRS$ and SEI_mI_hRS models, we can determine which model more accurately represents the data. This, in turn, will help policymakers optimize their strategies and allocate resources more effectively.

***SEIRS* Epidemiological Model.** To adapt the $SEIRS$ model to the context of toxicity, we redefine the groups as follows: A **Susceptible' (S)** user has not yet posted tweets using the identified hashtag. An **Exposed' (E)** user has encountered tweets containing the identified hashtag and experiences a delay before posting a tweet using the same hashtag. An **Infected' (I)** user is someone who posts a tweet using a hashtag identified. A **Recovered' (R)** user refrains from using the identified hashtag for a specified period.

We assume that the total number of users is variable, so we consider the recruitment rate (Λ) and the rate at which users leave autonomously (μ). When a user recovers, they re-enter the susceptible group at a rate of δ. Figure 1 shows the $SEIRS$ model. β is the effective contact rate, σ is the rate at which the exposed become infected, and θ is defined as $\frac{\beta I}{N}$. Therefore, the ordinary differential Eq. 1 is:

$$\begin{cases} dS(t) = \left[\Lambda + \delta R - \frac{\beta IS}{N} - \mu S\right] dt, \\ dE(t) = \left[\frac{\beta IS}{N} - (\mu + \sigma)E\right] dt, \\ dI(t) = \left[\sigma E - (\mu + \gamma)I\right] dt, \\ dR(t) = \left[\gamma I - (\mu + \delta)R\right] dt, \end{cases} \tag{1}$$

where at any given time (t), we define the quantity N by

$$N(t) = S(t) + E(t) + I(t) + R(t), \tag{2}$$

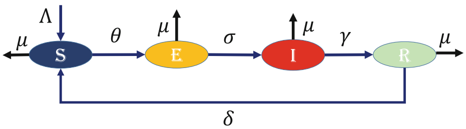


Fig. 1. Transfer diagram for the toxicity spread for $SEIRS$ Model

***SEI$_m$I$_h$RS* Epidemiological Model.** For the SEI_mI_hRS model, similar to the $SEIRS$ model, we adapted the model to address toxicity spread by utilizing the same redefined groups introduced for $SEIRS$ (Sect. 3.3). However, in SEI_mI_hRS, we feature two Infected groups. The **High-Infected** (I_h) users are those who post toxic tweets using the specified hashtag, with a toxicity score exceeding the average toxicity score of the entire toxic population. And the **Moderate-Infected** (I_m) users are those who post toxic tweets with a toxicity score falling between the average toxicity score of the toxic population and 0.5. This model incorporates distinct transfer rates to account for varying intensities of toxicity.

In the SEI_mI_hRS model, users are dynamically recruited at rate (Λ) and exit autonomously at rate (μ). Toxicity spreads with an effective contact rate (β), depending on the presence of moderate infected (I_m) and high infected users (I_h). Exposed users become moderate or high infected at rates σ_m and σ_h, respectively. Transitions between moderate and high infection occur at rates ϕ_m and ϕ_h. Both moderate and high infected users recover at rates γ_m and γ_h, respectively. Upon recovery, users rejoin the susceptible population at rate δ. Figure 2 represents the SEI_mI_hRS Model. θ is defined as $\frac{\beta(I_m+I_h)}{N}$. The proposed SEI_mI_hRS is calculated using the Eq. 3.

$$\begin{cases} dS(t) = \left[\Lambda + \delta R - \frac{\beta(I_m+I_h)S}{N} - \mu S\right]dt, \\ dE(t) = \left[\frac{\beta(I_m+I_h)S}{N} - (\mu + \sigma_m + \sigma_h)E\right]dt, \\ dI_m(t) = \left[\sigma_m E + \phi_h I_h - (\mu + \gamma_m + \phi_m)I_m\right]dt, \\ dI_h(t) = \left[\sigma_h E + \phi_m I_m - (\mu + \gamma_h + \phi_h)I_h\right]dt, \\ dR(t) = \left[\gamma_m I_m + \gamma_h I_h - (\mu + \delta)R\right]dt, \end{cases} \tag{3}$$

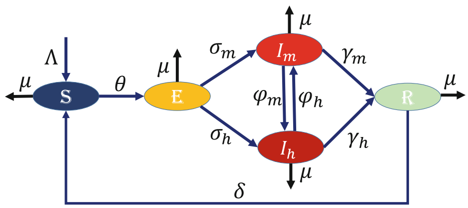


Fig. 2. Transfer diagram for the toxicity spread for SEI_mI_hRS Model

where at any given time (t), N is defined by

$$N(t) = S(t) + E(t) + I_m(t) + I_h(t) + R(t), \tag{4}$$

4 Results

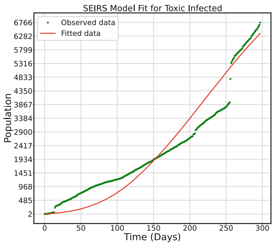
We apply the $SEIRS$ model to all datasets. To assess the accuracy of the model and understand how well it fits the data, we use an error rate. The error rate is computed using the following equation:

$$\text{Error rate} = \frac{\|I_{estimated}(t_i) - I_{observed}(t_i)\|_2}{\|I_{observed}(t_i)\|_2} \tag{5}$$

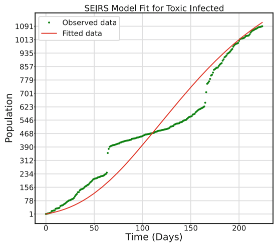
The error rate measures the disparity between the estimated and actual number of infected users at different time points t_i. The $|\cdot|_2$ norm, representing the Euclidean distance, quantifies the overall error relative to the actual data. A lower error rate suggests a better fit of the model to the observed data, validating its accuracy in representing toxicity spread. The $SEIRS$ model is applied to six datasets to assess its performance and evaluate its alignment with observed data, depicted in Fig. 3. The red line shows the Infected compartment, while blue dots represent actual tweets. Table 2 shows the error rates. Consistently low error rates across all datasets highlight the model's robustness in capturing toxicity dynamics, making it a reliable tool for understanding toxicity spread in various contexts.

Table 2. Error Rates Obtained from *SEIRS* Model across Six Datasets.

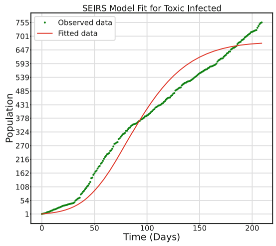
	F*Covid	F*Mask	F*Vaccine	F*Lockdown	Brazil Anti Gov	Brazil Pro Gov
Error rate	0.18	0.11	0.07	0.15	0.09	0.08



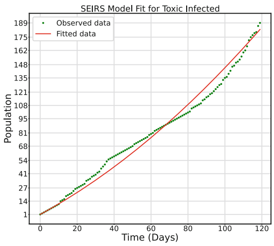
(a) F*Covid data



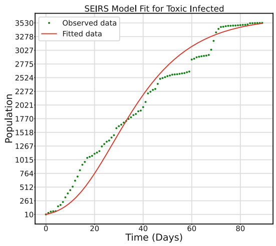
(b) F*Lockdown data



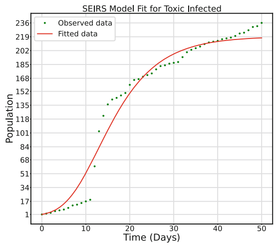
(c) F*Mask data



(d) F*vaccine data



(e) Brazil Anti Gov data



(f) Brazil Pro Gov data

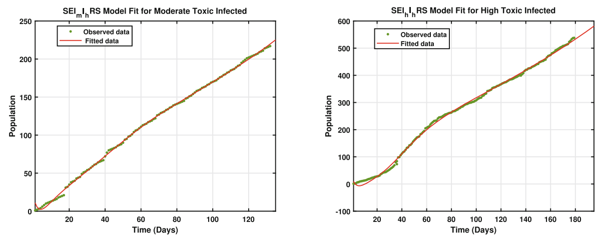
Fig. 3. *SEIRS* Model Fit to the Six Datasets.

After obtaining the results from the *SEIRS* model, to answer the research question, we employed our proposed SEI_mI_hRS model on each dataset. Figure 4 presents the fitted model, categorized by moderate and high toxicity intensity. Table 3 shows error rates for each dataset. The SEI_mI_hRS model achieves a lower error rate for both moderate and high toxicity compared to the *SEIRS* model. For instance, in the F*Covid dataset, the *SEIRS* model yields an error rate of 0.18, while the SEI_mI_hRS model results in error rates of 0.009 for high toxicity and 0.02 for moderate toxicity. Figure 3a and Fig. 4b show the fitted models for this F*Covid. After comparing Tables 2 and 3, similar improvements are observed across all datasets, with consistently lower error rates for the SEI_mI_hRS model. This indicates that dividing the infected compartment based on toxicity intensity helps capture toxicity propagation more accurately.

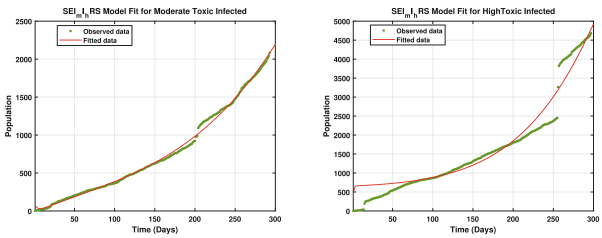
This suggests users in these datasets closely follow the SEI_mI_hRS model's progression, effectively corresponding to its structure. Figure 4 shows examples of how accurately the SEI_mI_hRS model fits the data for both high and moderate toxicity levels.

Table 3. Error Rates Obtained from SEI_mI_hRS Model across Six Datasets.

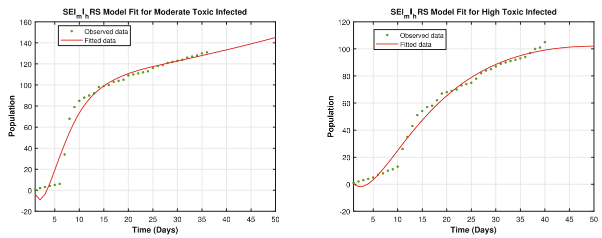
Error rate	F*Covid	F*Mask	F*Vaccine	F*Lockdown	Brazil Anti Gov	Brazil Pro Gov
Moderate	0.009	0.013	0.05	0.06	0.03	0.01
High	0.02	0.019	0.06	0.08	0.09	0.05



(a) F*Mask data



(b) F*Covid data



(c) Brazil Pro Gov

Fig. 4. SEI_mI_hRS Model Fit to the three datasets for high and moderate toxicity intensity. (The rest of the Plots are not shown due to the page limitation.)

5 Conclusion and Future Works

This research addressed the critical issue of the spread of toxicity on social media. It aims to determine if dividing the infected compartment by toxicity intensity improves model accuracy in capturing toxic content dissemination. This is achieved by creating and evaluating the $SEIRS$ model, then introducing the more advanced SEI_mI_hRS model. The SEI_mI_hRS model differentiates between moderate and high infected users to reflect the varying degrees of toxicity spread.

Applying both models to six datasets highlights the effectiveness of the approach. The $SEIRS$ model showed robustness and accuracy with consistently low error rates. However, the SEI_mI_hRS model, which divides the infected compartment into moderate and high toxicity levels, achieved even lower error rates. The key contribution of this work is the SEI_mI_hRS model, which improves toxicity spread modeling accuracy by differentiating between toxicity levels. It offers researchers a precise tool for analyzing online behavior. Policymakers and online platforms gain a detailed view for targeted interventions, allowing them to optimize resource allocation and effectively foster healthier online environments. Future work explores further refinements to the SEI_mI_hRS model, such as incorporating additional compartments to capture other user behaviors or extending the model to different social media platforms and contexts. Additionally, we will look into other error metrics to obtain more comprehensive results and apply our model for prediction analysis and simulation.

References

1. Shajari, S., Alassad, M., Agarwal, N.: Characterizing suspicious commenter behaviors. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, pp. 631–635 (2023). https://doi.org/10.1145/3625007.3627309
2. Shaik, M., Yousefi, N., Agarwal, N., Spann, B.: Evaluating role of instagram's multimedia in connective action leveraging diffusion of innovation and cognitive mobilization theories: Brazilian and peruvian social unrest case studies. In: Proceedings of 2023 10th International Conference on Behavioural and Social Computing (BESC), pp. 1–6 (2023). IEEE (2023). https://doi.org/10.1109/BESC59560.2023.10386436
3. Shajari, S., Agarwal, N., Alassad, M.: Commenter behavior characterization on youtube channels. In: Proceedings of the eKNOW International Conference on Information, Process, and Knowledge Management, Venice Italy, pp. 59–64 (2023)
4. DiCicco, K., Noor, N.B., Yousefi, N., Maleki, M., Spann, B., Agarwal, N.: Toxicity and networks of COVID-19 discourse communities: a tale of two social media platforms. In Proceedings of the 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval 2023, co-located with The 45th European Conference on Information Retrieval (ECIR 2023), Dublin, Ireland, 2–6 April 2023, pp 30–42 (2023). ISSN (1613), p. 0073. http://ceur-ws/

5. Noor, N.B., Yousefi, N., Spann, B., Agarwal, N.: Comparing toxicity across social media platforms for COVID-19 discourse. In: Proceedings of the Ninth International Conference on Human and Social Analytics, pp. 21–26 (2023)
6. Yousefi, N., Cakmak, M.C., Agarwal, N.: Examining multimodal emotion assessment and resonance with audience on youtube. In: Proceedings of the 9th International Conference on Multimedia and Image Processing (2024)
7. Yousefi, N., Noor, N.B., Spann, B., Agarwal, N.: Towards developing a measure to assess contagiousness of toxic tweets. In: Proceedings of the International Workshop on Combating Health Misinformation for Social Wellbeing, p. 43 (2023). https://doi.org/10.36190/2023.43.
8. Yousefi, N., Noor, N.B., Spann, B., Agarwal, N.: Examining toxicity's impact on reddit conversations. In: Proceedings of the 12th International Conference on Complex Network, pp 401–411 (2023). https://doi.org/10.1007/978-3-031-53503-1_33
9. Falade, T.C., Yousefi, N., Agarwal, N.: Toxicity prediction in reddit. In: Proceedings of the 30th Americas Conferences on Information Systems (2024)
10. DeMarsico, D., Bounoua, N., Miglin, R., Sadeh, N.: Aggression in the digital era: assessing the validity of the cyber motivations for aggression and deviance scale. Assessment **29**, 764–781 (2022)
11. Jiang, J.A., Scheuerman, M.K., Fiesler, C., Brubaker, J.R.: Understanding international perceptions of the severity of harmful content online. PLoS ONE **16**, e0256762 (2021)
12. Obadimu, A., Khaund, T., Mead, E., Marcoux, T., Agarwal, N.: Developing a socio-computational approach to examine toxicity propagation and regulation in COVID-19 discourse on YouTube. Inf. Process. Manag. **58**, 102660 (2021). https://doi.org/10.1016/j.ipm.2021.102660
13. Kauk, J., Kreysa, H., Schweinberger, S.R.: Understanding and countering the spread of conspiracy theories in social networks: evidence from epidemiological models of Twitter data. PLoS ONE **16**, e0256179 (2021)
14. Maleki, M., Arani, M., Mead, E., Kready, J., Agarwal, N.: Applying an epidemiological model to evaluate the propagation of toxicity related to COVID-19 on twitter. In: The 55th Hawai'i International Conference on System Sciences (HICSS) (2022)
15. Maleki, M., Mead, E., Arani, M., Agarwal, N.: Using an epidemiological model to study the spread of misinformation during the black lives matter movement. In: International Conferences on Fake News, Social Media Manipulation and Misinformation (2021)
16. Yousefi, N., Agarwal, N.: Studying the influence of toxicity intensity on its propagation using epidemiological models. In: Proceedings of the 30th Americas Conference on Information Systems (2024)
17. Jin, F., Dougherty, E., Saraf, P., Cao, Y., Ramakrishnan, N.: Epidemiological modeling of news and rumors on twitter. In: Proceedings of the 7th Workshop on Social Network Mining and Analysis, pp. 1–9 (2013)
18. Nie, Q., Liu, Y., Zhang, D., Jiang, H.: Dynamical SEIR model with information entropy using COVID-19 as a case study. IEEE Trans. Comput. Social Syst. **8**, 946–954 (2021)
19. Dong, S., Deng, Y.-B., Huang, Y.-C.: SEIR model of rumor spreading in online social network with varying total population size. Commun. Theor. Phys. **68**, 545 (2017)
20. Bae, S., Sung, E., Kwon, O.: Accounting for social media effects to improve the accuracy of infection models: combatting the COVID-19 pandemic and infodemic. Eur. J. Inf. Syst. **30**, 342–355 (2021)

21. Liu, X., Li, T., Tian, M.: Rumor spreading of a *SEIR* model in complex social networks with hesitating mechanism. Adv. Differ. Equ. **2018**(1), 1–24 (2018). https://doi.org/10.1186/s13662-018-1852-z
22. Zhang, Q., Fang, H., Yao, P., Zhang, H.: Research on the online spread of rumors about the quality and safety of agricultural products in the new media era using an improved SEIRS model. Disc. Dyn. Nature Soc. **2021**(1), 5526775 (2021)
23. Hosseini, S., Zandvakili, A.: The SEIRS-C model of information diffusion based on rumour spreading with fuzzy logic in social networks. Int. J. Comput. Math. **99**, 1918–1940 (2022)
24. Tong, Q., Wang, H., Zhang, J., Li, L., Huang, Q.: The fractional SEIRS epidemic model for information dissemination in social networks. In: Liu, Y., Wang, L., Zhao, L., Yu, Z. (eds.) ICNC-FSKD 2019. AISC, vol. 1075, pp. 284–291. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-32591-6_30
25. Gardner, M.T., Beard, C., Medhi, D.: Using SEIRS epidemic models for IoT botnets attacks. In: 13th International Conference DRCN 2017-Design of Reliable Communication Networks, VDE, pp 1–8 (2017)
26. Arisman, A., Simbolon, H.F.S.: Modeling the dynamics of misinformation spread on social media platforms. Jurnal Teknik Informatika CIT Medicom **15**(297–305), 297–305 (2024)
27. Hanu, L.: Unitary, team: Detoxif (2020). https://github.com/unitaryai/detoxify,
28. Saveski, M., Roy, B., Roy, D.: The structure of toxic conversations on Twitter. In: Proceedings of the Web Conference 2021, pp. 1086–1097 (2021)

VSM-ACT-R: Toward Using Cognitive Architecture For Manufacturing Solutions

Siyu Wu[1(✉)], Alessandro Oltramari[2], and Frank E. Ritter[1]

[1] College of Information Sciences and Technology, Penn State University, University Park, USA
sfw5621@psu.edu
[2] Bosch Research and Technology Center, Pittsburgh, USA

Abstract. The era of Industry 4.0 demands innovative solutions to produce high-quality products within tight lead times. This paper explores the integration of cognitive architectures (CAs) into manufacturing solutions, with a focus on using VSM-ACT-R, a cognitive architecture model built upon the ACT-R architecture. VSM-ACT-R aids in making informed decisions in smart scheduling that boosts productivity while ensuring consistent quality. The model stands out in three key aspects of decision-making in manufacturing: First, it executes tasks using decision-making algorithms and knowledge representations observed in human subjects, supported by declarative memories that reflect intuitive and domain-specific knowledge. Second, it mimics various levels of decision-making-from novice through to expert—using production rules and retrieval mechanisms that replicate variations of human behavior. Third, it simulates the learning processes of decision-makers, managed by a decision-choice control center that is driven by utility learning and reinforcement reward. We conclude by discussing an evaluation of this model, its applications, and its implications.

Keywords: Cognitive Architecture · Manufacturing · ACT-R

1 Introduction

Industry 4.0 aims to create "intelligent factories" where advanced manufacturing technologies enable smart decision-making through real-time communication and cooperation among humans, machines, and sensors [13]. Smart scheduling, which leverages advanced models and algorithms using sensor data, exemplifies one such solution [10].

A value stream map (VSM) is an essential tool in smart scheduling. It serves as a sophisticated flowchart that visualizes and controls the production line [8]. VSM meticulously tracks metrics like inputs, outputs, processes, overall equipment effectiveness (OEE), and cycle times-all crucial for quality and efficiency analysis in production control. However, plant managers face significant challenges in using VSM in production management. These challenges include difficulty applying VSM concepts to complex, real-world scenarios characterized by

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 69–79, 2024.
https://doi.org/10.1007/978-3-031-72241-7_7

a high number of intertwined variables. This complexity consistently impedes plant decision-makers from making timely and optimal decisions regarding both time reduction and maintaining stable quality on the production lines.

This paper proposes a novel approach to address these challenges by integrating cognitive architectures into decision-making processes for manufacturing. Specifically, it employs a cognitive architecture to build models representing decisions and their process related to boosting productivity and ensuring consistent quality. This model leverages data derived from the VSM and decision-makers at Bosch plants.

Cognitive architectures (CAs) aim to create a unified model of the mind using invariant mechanisms to simulate and explain human behavior [1,5,7]. CAs use task-specific knowledge to generate behavior. They represent various types of knowledge, including declarative (factual), procedural (how-to), and in recent advancements, perception and motor skills. This knowledge allows CAs to not only simulate behavior but also explain it, both through direct examination and by tracing the reasoning steps involved in real-time (concurrent protocol).

This reports starts from prototypical decision processes distilled by plant managers of Bosch. Their insights, combined with a VSM tailored to their specific plant system, inform the build of our VSM-ACT-R model to enhance decision-making. It then introduces the developed VSM-ACT-R model[1], which stands out in decision-making tasks with three key strengths. First, the model can execute tasks using decision-making behaviors observed in humans and retrieve knowledge representations similarly. This capability is achieved through incorporating declarative memories that cater to intuition and professional knowledge from human subjects.

Second, the model integrates personas ranging from novice to intermediate and expert levels. This is achieved through developed sets of production rules that mimic the behavior of decision-makers at various expertise levels, coupled with retrieval mechanisms for full or partial knowledge representation.

Third, the model simulates the learning processes of decision-makers, transitioning from novice to expert. This simulation is facilitated by the decision-choice control center, which manages error-making, learning, and memory through utility learning and reinforcement rewards. This approach creates a realistic and dynamic decision-making simulation, making the VSM-ACT-R model a robust tool in cognitive architecture-facilitated decision-making in manufacturing.

The following sections discuss the task, the model, the model's performance evaluation, its application, and implications.

2 Related Work

In this section, we introduce ACT-R and its strengths to create a model that simulates human decision making behavior with learning.

There are currently primarily two kinds of knowledge representations in ACT-R: declarative and procedural knowledge. Declarative knowledge consists

[1] https://github.com/SiyuWu528/VSM-ACT-R.

of chunks of declarative memory (e.g., apple is a kind of fruit), while procedural knowledge performs basic operations, moves data among buffers, and identifies the next instructions to be executed (e.g., to submit your answer, you have to click the submit button). ACT-R has extensive applications across psychology and computer science, including professional development [4], military simulations [2], and autonomous driving simulations [11].

ACT-R is effective in developing models to simulate human learning. Three key features distinguish the use of ACT-R in creating models that perform decision-making tasks with learning:

Self-configuration: ACT-R efficiently translates instructions into structured rules, forming the basis for task-specific production rules that enhance the efficiency of task execution.

Modular Design Mirroring Human Cognition: ACT-R's modules emulate human cognitive functions: perceptual modules update the system's view of the environment, a goal module tracks progress towards objectives, a declarative module uses past experiences for contextual understanding, and a central buffer system enables communication between modules. Additionally, the central production system recognizes patterns to initiate coordinated actions.

Subsymbolic Processes for Decision-making: ACT-R excels in its ability to reliably retrieve relevant memories and activate appropriate rules, ensuring both efficient and adaptive performance in decision-making tasks, such as skills training. It does so at a pace that mirrors human performance and offers the opportunity to model learning during this process.

3 The Task

This section details formulating a domain-specific decision problem for optimal production efficiency, leveraging VSM to define efficiency sectors and then abstracting the problem for mathematical modeling.

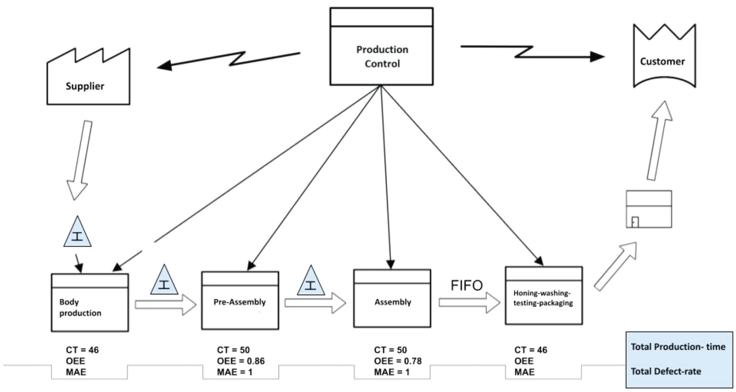


Fig. 1. An Example of Value Stream Map in a Plant Floor

The VSM (Fig. 1) depicts a prototypical manufacturing production line workflow from supplier to customer. Key components include Body Production, Pre-Assembly, Assembly, Honing, Washing, Testing, and Packaging. Later stages are interconnected via First-In-First-Out (FIFO) processes. Metrics displayed for each stage include Cycle Time (CT), Overall Equipment Effectiveness (OEE), and Mean Absolute Error (MAE). The flow progresses through each stage, aiming for efficient operation, performance monitoring, and error minimization to ensure high-quality production output and timely customer delivery.

Focusing on maintaining stable output for the plant, we consider the plant managers' feedback alongside the Value Stream Map (VSM) structure to develop a decision-making problem that aims to reduce total assembly time while minimizing the increase in defect rate. The task: Our manufacturing line has two sections with potential defect sources: pre-assembly and assembly. Pre-assembly takes 40 s with an OEE rate of 88%, while assembly takes 44 s with an OEE rate of 80.1%. To reduce total assembly time by 4 s, we need to identify which section can be shortened with minimal defect increase. There are two options: reduce pre-assembly time or reduce assembly time.

4 The Model

This section starts with capturing intuition and domain knowledge from decision makers, followed by the model structure and learning mechanism, and concludes by examining a model output snippet from one run of our VSM model.

4.1 Model Design

The model, built upon the prototypical decision process distilled by Bosch plant managers, incorporates how cognitive models are designed for different levels of expertise [3,6]. For novices, the model utilizes intuitive deliberative chunks to make decisions. For intermediates, it understands key metrics such as cycle time (CT) and Overall Equipment Effectiveness (OEE). However, intermediates often lack the ability to systematically analyze how these metrics interrelate and cumulatively impact efficiency and quality. Experts, on the other hand, make well-informed judgments based on a comprehensive view of all relevant metrics, obtained through Value Stream Mapping (VSM).

4.2 Declarative Chunks

We created chunks representing knowledge from intuitions to professional expertise. These representations are divided into three chunk types: decisions, decision merits, and goals. Decision chunk encodes six slots: reduction time, decision-making state (e.g., novice, intermediate, expert), OEE, and CT. The decision merits chunk holds knowledge on weights for sectors, defect increase for sectors, and the difference in defect rate increase between the two. The goal chunk encodes the initial production conditions and the ultimate goal of making the optimal decision.

4.3 Production Rules

Three sets of production rules represent the decision-making behaviors of novice, intermediate, and expert decision-makers. These sets comprise a total of 17 rules, each driven by goal-focused objectives across 14 states.

We use the expert production rule set as an example, as shown in Fig. 2. Once the decision-choice center decides to activate this set of expert decision productions, it starts by perceiving the problem and retrieving related decision-making metrics from chunks. The imaginal buffer then acts as a temporary workspace, holding and manipulating relevant information during decision-making. It allows the model to build new mental representations or modify existing ones based on incoming data or problem-solving needs. This involves using the imaginal buffer to assess the relationships between the decision target and decision metrics, particularly considering the impact of each sector's weight on the defect rate change, and determining the final defect rate increase for each sector. These results are stored in the imaginal buffer and later retrieved for comparison. This then allows the model to select the sector with the lowest defect increase.

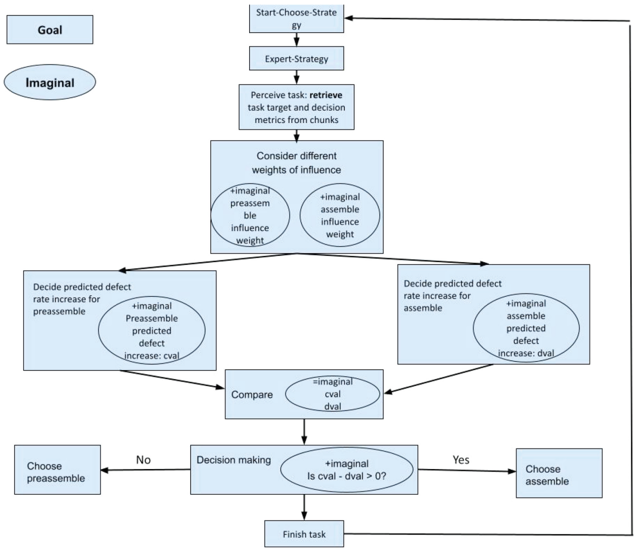


Fig. 2. Production rules control structure for expert decision making and their use of the ACT-R Goal and Imaginal buffers

4.4 Level of Expertise Mechanism

The model can learn while performing tasks through two mechanisms leading to varying levels of expertise, as shown in Fig. 3.

The model mimics human decision-making behavior through differentiating knowledge representations. **Declarative Memories**: These memories store

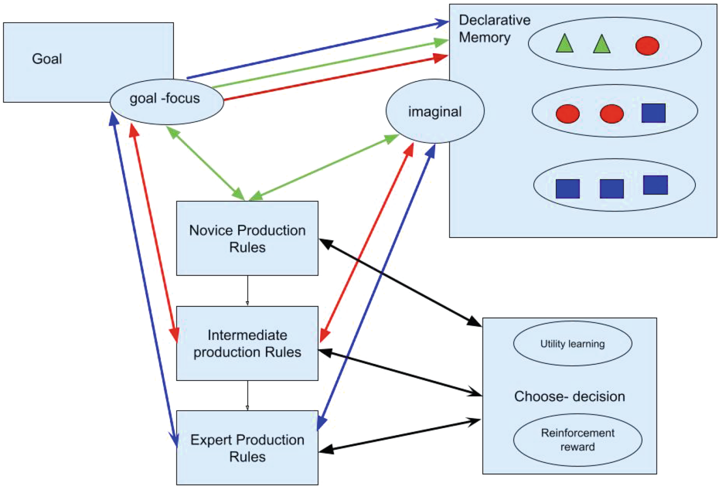


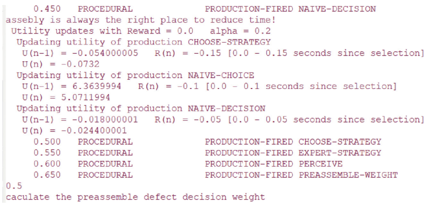
Fig. 3. Level of expertise mechanism in VSM-ACT-R

knowledge that aligns with human intuition and expertise gained from the VSM. For example, the green triangles in the figure represents a portion of the intuition used by novice decision-makers. **Production Rules**: These rules capture the rational decision-making processes observed in human subjects. The green (lighter) lines illustrate how the imaginal buffer retrieves relevant portions of the novice declarative memory and feeds them to the novice production rule set. Intermediate and expert decision-making levels follow the same principle. Red and blue shapes represent their respective declarative memory chunks, and the corresponding (darker) colored arrows show the flow of information through their production rule sets. Finally, the goal buffer utilizes the "goal focus" command to manipulate the different phases of the task.

Beyond mimicking human behavior, the model also simulates the learning progress achieved by the **Decision-Choice Control**, which manages errors, learning, and memory through utility learning and reinforcement rewards. Novice decision-making starts with a utility base and includes a noise setting. The intermediate and expert production rules receive rewards when the corresponding decision-making results are achieved. The utility of these production rules updates is based on the rewards received and the retention of memory, which depends on the time passed since the rule last fired.

4.5 Model Output

The partial trace in Fig. 4 shows how the model transitions from naive to more expert-like behaviors. Each production rule's utility is updated based on the reward received and the time since the last selection. For example, the NAIVE-CHOICE rule's utility decreased from 6.36 to 5.07 due to a reward of –0.1 for the time passed since the last selection. As the utility of naive strategies decreases, the likelihood of the EXPERT-STRATEGY being fired increases.



```
    0.450   PROCEDURAL              PRODUCTION-FIRED NAIVE-DECISION
assebly is always the right place to reduce time!
 Utility updates with Reward = 0.0   alpha = 0.2
  Updating utility of production CHOOSE-STRATEGY
   U(n-1) = -0.054000005   R(n) = -0.15 [0.0 - 0.15 seconds since selection]
   U(n) = -0.0732
  Updating utility of production NAIVE-CHOICE
   U(n-1) = 6.3639994   R(n) = -0.1 [0.0 - 0.1 seconds since selection]
   U(n) = 5.0711994
  Updating utility of production NAIVE-DECISION
   U(n-1) = -0.018000001   R(n) = -0.05 [0.0 - 0.05 seconds since selection]
   U(n) = -0.024400001
    0.500   PROCEDURAL              PRODUCTION-FIRED CHOOSE-STRATEGY
    0.550   PROCEDURAL              PRODUCTION-FIRED EXPERT-STRATEGY
    0.600   PROCEDURAL              PRODUCTION-FIRED PERCEIVE
    0.650   PROCEDURAL              PRODUCTION-FIRED PREASSEMBLE-WEIGHT
0.5
caculate the preassemble defect decision weight
```

Fig. 4. VSM-ACT-R Model Output

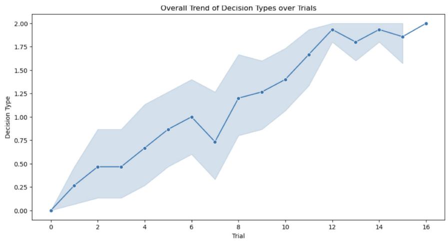


Fig. 5. Trend of Decision Types over Trials with SD shown as gray fill (Color figure online)

5 Model Evaluation

To answer the question of whether this model learns and how it simulates learning progression and captures individual differences, we first use descriptive statistics and linear regression to show the average progression of decision types across 16 trials. We then use a mixed linear model to assess and illustrate the effects of trials on decision types across ACT-R model personas, with repeated measures of trials, and random effects to account for individual differences. Last but not least, we use an ordered logistic regression to analyze and understand the relationship between the number of trials and an ordinal dependent variable of learning progress from novice to expert.[2]

5.1 Analyzing Learning Rate

We ran the ACT-R model 15 times to understood its behavior [9]. Each time, we asked it to run 15–16 trials until the model achieved stable expert behavior. We collected data with decision types encoded as 0, 1, and 2 for novice, intermediate, and expert strategies.

The decision-making data for the runs, acting as ACT-R personas, are shown in Fig. 5. the average progression of decision types from novice (0) to expert (2) across 16 trials. Starting at approximately 0 in trial 0, the mean decision type rises to about 0.75 by trial 4 and reaches around 1.25 by trial 8. Despite slight fluctuations, the trend continues upward, with the mean decision type approaching 1.75 by trial 12 and around 1.9 by trial 16. The narrowing 95% confidence intervals, ranging from approximately 0.5 to 2.0 initially to 1.5 to 2.0 in later trials, indicate increasing consistency among participants' decision-making abilities.

The learning rate, defined as the rate at which decision type progresses from novice (0) to expert (2) across trials, is modeled using a linear regression. This

[2] Notebook and data can be accessed at https://github.com/SiyuWu528/VSM-ACT-R/tree/main/Brims_data_analysis.

model assumes a constant learning rate across all trials shown in Eqn. 1.

$$Eqn.1:\ y = \beta \times x + \alpha$$

where y is the mean of decision type, x is the trial number, and β (the slope) represents the learning rate. The learning rate for the ACT-R personas is 0.111 with variance of the residuals of 0.026.

5.2 Analyzing Individual Differences

We then use a mixed linear model that includes both fixed and random effects, to assess the effects of trials on decision types, and random effects to account for individual differences. This analysis allows to handling data with nested structures (e.g., multiple trials per personas). In addition, it accounts for the correlation of responses within the same participant and allows for the inclusion of random effects due to individual differences (Table 1).

Table 1. Mixed Linear Model Regression Results

Dependent Variable:			decision_type			
No. Observations:	227		**Method:**		REML	
No. Groups:	15		**Scale:**		0.4014	
Min. group size:	15		**Log-Likelihood:**		-232.9159	
Max. group size:	16		**Converged:**		Yes	
Mean group size:	15.1					
	Coef.	**Std.Err.**	**z**	**P>|z|**	**[.025**	**.975]**
Intercept	0.151	0.112	1.340	.180	-0.070	0.371
Trial	0.127	0.010	13.198	.000	0.108	0.146
Group Var	0.076	0.063				

Significant Effect of Trial on Decision Type. The coefficient for the trial is 0.127 with a p-value of $<.05$, indicating a highly significant positive effect of trial on decision type. This suggests that experience or exposure to more trials positively influences the decision-making process, resulting in higher decision type scores. Participants learn or adapt their decision-making strategies over time, becoming more proficient or confident with each subsequent trial.

Individual Differences Among Participants. The random effects component of the model shows a variance of 0.076 for participants, indicating variability in the intercepts across different participants. This variability suggests that while the overall trend shows an increase in decision-type scores with more trials, individual participants start from different baseline levels. In humans, some participants may naturally have higher or lower decision-type scores due to personal characteristics, prior experience, or other unmeasured factors.

5.3 Analyzing Learning Progress

We then use an ordered logistic regression model without considering individual differences, to analyze the relationship between the number of trials and an ordinal dependent variable of learning progress from novice to expert. This aims to look deeper into how changes in the predictor influence the likelihood of different levels of the ordered outcome in decision-making.

Table 2. Ordered Model Regression Results

Dep. Variable:	decision_type			**Log-Likelihood:**		-182.40
Model:	OrderedModel			**AIC:**		370.8
Method:	Maximum Likelihood			**BIC:**		381.1
No. Observations:				227		
Df Residuals:				224		
Df Model:				1		
	coef	**std err**	**z**	**P>|z|**	**[.025**	**.975]**
Trial	0.3545	0.040	8.802	.000	0.276	0.433
0/1	1.6906	0.310	5.447	.000	1.082	2.299
1/2	0.2262	0.139	1.631	.103	-0.046	0.498

Table 2 shows that the threshold 0/1 (1.69) with p-value < 0.05 indicates a significant cut-off between novice and intermediate categories. The threshold 1/2 (0.23) is not statistically significant (p-value = .103), suggesting that the model does not provide strong evidence for a clear separation between intermediate and expert decision types over just 15 trials.

ACT-R personas tend to move to higher decision categories as they undergo more trials, with a significant transition between novice and intermediate, but not as clear a transition between intermediate and expert. The initial learning curve is steep, however, once personas reach an intermediate level, further improvements become subtler.

6 Conclusion and Discussion

This study towards using cognitive architecture to enhance manufacturing efficiencies by creating VSM-ACT-R, created a model that incorporates learning and behavior differentiation in a decision-making task aimed at optimizing a manufacturing production line. The model simulates three types of behavior-novice, intermediate, and expert-mirroring human decision-making rationales. The model learns over the course of trials and exhibits individual differences. It demonstrates a human-like learning progression, showing a steep learning curve at the beginning and gradual improvements later on.

It is worth noting that the subtle and gradual progression of the model from intermediate to expert levels may be inherent to the model itself. It is possible that the differentiation between the expert and intermediate decision-making levels is not distinct enough. Additionally, the incentives provided to the expert strategy may be significantly higher than those given to the intermediate strategy and could lead the model to have a smooth learning curve later on. Therefore, it could be worthwhile to adjust the model based on real-world scenarios.

VSM-ACT-R could be extended to teach novice decision-makers not only optimized strategies but also highlight common mistakes they might make, guiding them through a learning trajectory. It will be able to serve as a tutor in manufacturing decision-making, not just by providing the right answers but by guiding them on how to achieve those answers, akin to a peer.

We are particularly excited about the model's potential for further deployment with open-source large language models [12]. In manufacturing decision-making, off-the-shelf generative models often struggle to deliver accurate results and learning behavior exhibited by cognitive models. We can leverage VSM-ACT-R's ability to simulate tens of thousands of ACT-R participants in decision-making tasks to generate target data that incorporate learning and optimized decisions. This generated data will be used as the target to fine-tune large language models, aiming to align their decision-making with the ACT-R agents we will develop, The fine-tuned language model not only predicts human decisions for new problems but also provides important insights into the learning and correction rates in these tasks.

Acknowledgments. This article was completed during Siyu Wu's internship at the Bosch Technology and Research Center in Pittsburgh, working in a project on Neuro-Symbolic AI for manufacturing.

References

1. Anderson, J.R.: How Can the Human Mind Occur in the Physical Universe? Oxford University Press, New York, NY (2007)
2. Best, B.J., Lebiere, C.: Teamwork, communication, and planning in ACT-R agents engaging in urban combat in virtual environments. In: Proceedings of the 2003 IJCAI Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions, pp. 64–72 (2003)
3. Blessing, S.B., Anderson, J.R.: How people learn to skip steps. J. Exp. Psychol. Learn. Mem. Cogn. **22**(3), 576–598 (1996)
4. Chen, H., Liu, S., Pang, L., Wanyan, X., Fang, Y.: Developing an improved ACT-R model for pilot situation awareness measurement. IEEE Access **9**, 122113–122124 (2021)
5. Laird, J.E.: The Soar Cognitive Architecture. MIT Press, Cambridge (2012)
6. Nerb, J., Ritter, F.E., Krems, J.F.: Knowledge level learning and the power law: a soar model of skill acquisition in scheduling. Kognitionswissenschaft **8**(1), 20–29 (1999)
7. Newell, A.: Unified Theories of Cognition. Harvard University Press, Cambridge (1994)

8. Rahani, A.R., Al-Ashraf, M.: Production flow analysis through value stream mapping: a lean manufacturing process case study. Procedia Eng. **41**, 1727–1734 (2012)
9. Ritter, F.E., Schoelles, M.J., Quigley, K.S., Klein, L.C.: Determining the number of simulation runs: treating simulations as theories by not sampling their behavior. In: Rothrock, L., Narayanan, S. (eds.) Human-in-the-Loop Simulations, LNCS, pp. 97–116. Springer, London (2011). https://doi.org/10.1007/978-0-85729-883-6_5
10. Serrano-Ruiz, J.C., Mula, J., Poler, R.: Smart manufacturing scheduling: a literature review. J. Manuf. Syst. **61**, 265–287 (2021)
11. Wu, S., Bagherzadeh, A., Ritter, F.E., Tehranchi, F.: Long road ahead: lessons learned from the (soon to be) longest running cognitive model. In: Proceedings of the 21st International Conference on Cognitive Modeling (ICCM), pp. 281–287 (2023)
12. Wu, S., Giles, C.L., Ritter, F.E.: LLAMA-ACT-R, a neuro-symbolic architecture (ACT-R) for LLM decision making. University of Michigan Institute for Data Science, Poster presented at Annual Ethical AI Symposium (2024)
13. Zhong, R.Y., Xu, X., Klotz, E., Newman, S.T.: Intelligent manufacturing in the context of industry 4.0: a review. Engineering **3**(5), 616–630 (2017)

User-Donated Screenshots Analysis: Feasibility of a New Approach to Collect Objective Social Media App Usage in Adolescents

Yuning Liu[1](✉), Geoff Klassen[2], Jenna Mee[2], Justin Pointer[2], Marvi Baloch[2], Laura Marciano[1], and Nathaniel Osgood[2]

[1] Harvard T.H. Chan School of Public Health, Cambridge, MA 02138, USA
{yuning_liu,lmarciano}@hsph.harvard.edu

[2] University of Saskatchewan, Saskatoon, SK S7N 5A2, Canada
{geoff.klassen,Jen917,justin.pointer,marvi.baloch,nathaniel.osgood}@usask.ca

Abstract. Objective data on social media use is now urgently needed for understanding its impact on adolescent well-being. Traditional objective social media data collection methods, such as data donation and passive sensing, face challenges including intrusiveness, privacy concerns, and limitations in adolescent—a critical demographic in this research area. In our study, we introduced a novel, less intrusive method using user-donated screenshots within an ecological momentary assessment (EMA) framework. We recruited 374 adolescents from Switzerland, who were instructed to capture and share three daily screenshots detailing their total and app-specific usage across screentime, activations, and notifications. From this, we collected 6,819 screenshots, with 25% of participants failing to submit any screenshots, 14% submitted incorrect or incomplete ones, while 64% provided complete data for more than five days. To process this data, we developed an image-to-text pipeline using Tesseract OCR that achieved a 96% average accuracy rate. This user-donated screenshot method proved to be less burdensome than traditional data donation, capable of capturing detailed app-specific usage across smartphone operating systems, and applicable among adolescents. Nonetheless, success of the user-donated screenshot approach hinges on user compliance. We analyze attrition sources and suggest six strategies to enhance future research, such as incentivizing participation, implementing pre-upload image checks, and improving participant onboarding and education.

Keywords: screenshot analysis · social media use · optical character recognition · ecological momentary assessment · adolescents

Y. Liu, G. Klassen—Co-first author.

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 80–89, 2024.
https://doi.org/10.1007/978-3-031-72241-7_8

1 Introduction

1.1 The Need for Objective Social Media Use Data

Social media and smartphone usage have surged over the last decade, coinciding with a marked decline in adolescent well-being [1–3]. This parallel trend has rung the alarms among researchers and policymakers. The U.S. Surgeon General's 2023 advisory on Social Media and Youth Mental Health emphasizes the need to closely examine social media behaviors to connect specific usage patterns with well-being outcomes [4]. While the link between social media use (SMU) and adolescent well-being is still contentious, understanding nuanced social media activities is central to this debate [5, 6].

The complexity of individual SMU patterns makes it challenging to assess their impact on well-being. Simply tracking total screentime is insufficient considering that identical usage durations can involve different apps and behaviors, leading to a variety of effects on well-being [7, 8]. For instance, two adolescents with equal screentime might engage differently—one frequently checking their phone, the other spending prolonged periods browsing. Such discrepancies highlight the limitations of self-reported data, which often underestimates actual usage or shows small correlation with objective usage data [9]. Therefore, objective data on SMU are now crucially needed for exploring the relationship between SMU and adolescent well-being, to inform the policy making agenda and develop public health interventions.

1.2 Limitations of Current Objective Social Media Use Data Collection Methods

Scholars have developed several methods to collect objective SMU data, known as digital trace or phenotyping. These methods capture user activities saved in online environments [10]. Objective SMU data can be collected via Application Programming Interfaces (APIs), app data donation, and passive sensing, though each method faces challenges [11]. API-based collection is limited by content restrictions, rate limits, and policy changes. Additionally, it is challenging to correlate SMU data obtained from APIs with well-being outcomes, which are typically assessed via surveys or professional diagnoses [11, 12]. App data donation packages consist of archives from social media platforms that each user can request to download. This approach is constrained by low compliance rates, delays in receiving data packages, and the complexity of extracting information from these packages, with different social media platforms presenting varied and complex data formats, including compliance with privacy issues [13].

Passive sensing apps generally only measure overall screen time and, except a few on Android, fail to track screen states by app due to iOS restrictions [14]. To address these challenges, the Human Screenome Project introduces Screenomic analysis, capturing smartphone screenshots every five seconds during active use [15, 16]. This approach represents a significant advancement in passive social media tracking, yet it raises concerns over its intrusiveness and the privacy risks [17]. Additionally, high-frequency data collection produces a vast amount of image data, which requires substantial computational resources to process and derive meaningful insights. In summary, effectively gathering granular and objective data on SMU, while adhering to international regulations and minimizing privacy concerns, poses a significant challenge.

1.3 The Present Study

This study introduces a novel method for collecting objective social media data through user-donated screenshots on app-specific usage, attempting to address several limitations of existing methods. This innovative approach utilizes an ecological momentary assessment (EMA) design, requesting participants to share screenshots of app usage from their smartphone Settings alongside survey responses in each EMA. This method ensures privacy protection and non-intrusiveness and is adaptable across various populations and settings. We developed an image-to-text pipeline to extract data from these screenshots. This paper presents our new approach, the user-donated screenshots analysis, and demonstrates the pipeline's performance through evaluation results. Additionally, we summarize the attrition patterns observed and discuss strategies to mitigate participant drop-off in future research. Ultimately, our study seeks to answer the following three main research questions: (i) How feasible is it to derive app-specific usage features from user-donated screenshots? (ii) How practical is it to collect such screenshots from participants in an EMA study? (iii) And how can future studies applying user-donated screenshot approach be designed to minimize attrition?

2 Method

2.1 Data Collection

The HappyB study studies social media use and well-being by collecting intensive longitudinal data from ecological momentary assessments (EMAs) among 374 Swiss adolescents (mean age = 15.71, SD = 0.82; 235 females, 62.8%) across 14 days in spring 2022. Participants were recruited from four high schools in Switzerland, encompassing students in the first and second high school grades, occurring between March 7th and March 23rd, 2022. Participants initially completed a baseline survey and engaged in EMAs through the Ethica app (now known as Avicenna), with assessments conducted three times daily at 12 pm, 6 pm, and 9 pm from Day 1 to Day 14. Self-reported questions on social media activity, experience, and well-being are included in each EMA. Furthermore, during the 12 pm EMAs, participants were directed to access their smartphone's Settings, capture three screenshots illustrating the total usage and the top three app usage statistics from the previous day—specifically, daily usage time, screen activations, and notification counts—and subsequently upload these screenshots to Ethica. A gamification system was developed inside the app to track participants' adherence. Each adolescent received a medal labeled in different ways (e.g., "conqueror", "hero") depending on their total submitted EMAs. At the end of the study, depending on adherence rates and then medal obtained, each participant received a gift card up to 20 CHF. Participants were then asked to upload the three screenshots to Ethica. Example screenshots are presented in Fig. 1. The study received the ethical approval by the IRB of USI, Università della Svizzera italiana, Lugano, Switzerland, and was supported by the Department of Education, Culture, and Sport of canton Ticino.

2.2 Image-to-Text Pipeline

We extract data on overall and app-specific usage for the three features—screentime, number of activations, and number of notifications—from the user-donated screenshots. Specifically, we extract the total use, identify the names of the top three most-used apps, and obtain the usage statistics for these apps from each screenshot. As shown in Fig. 1(C), the extracted information includes a total of 347 notifications received on the previous day, with WhatsApp accounting for the highest number at 267 notifications. TikTok and Instagram follow, with 24 and 17 notifications respectively.

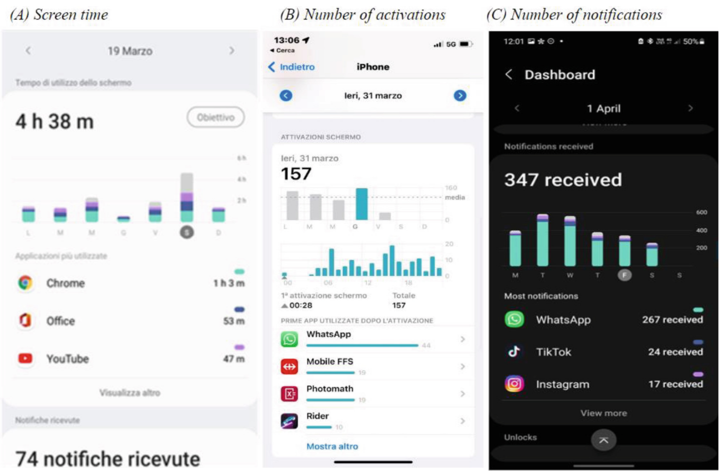


Fig. 1. Example of user-donated screenshots on the screen time of App use (A), activation by App (B), and notification by App (C)

We developed an image-to-text pipeline using Tesseract Optical Character Recognition (OCR) to extract total usage and app-specific data from screenshots. The workflow of this pipeline is depicted in Fig. 2. To assess the accuracy of our text extraction, we randomly selected 600 screenshots, divided equally among iOS and Android devices for screen time, activation, and notification data, and annotated them to establish ground truth. For each screenshot, we generated seven metrics: total screentime/count, the app name with the highest screentime/count, screentime/count for this top app, names and screentime/counts for the second and third highest usage apps. The pipeline's performance was evaluated by comparing the accuracy rate—the proportion of screenshots where extracted text precisely matched the ground truth (Table 1).

2.3 Attrition Pattern Analysis

To evaluate the feasibility of using user-donated screenshots, we assessed not only the performance of the image-to-text pipeline but also analyzed the attrition patterns. This analysis aimed to identify how errors could arise from the user-donated screenshots.

Understanding the sources of these errors is crucial for assessing feasibility. We identified that issues occur at two levels: the user level (loss to follow-up or failure to submit screenshots) and the screenshot level (submission of incorrect or incomplete information). Consequently, we analyzed and summarized the sources of these problems at both levels, with the results depicted in the flowchart in Fig. 3.

3 Result

3.1 Image-to-Text Pipeline Achieves High Accuracy

Our image-to-text pipeline (Fig. 2) contains a series of steps to process screenshots from smartphones, utilizing Tesseract OCR for text recognition.

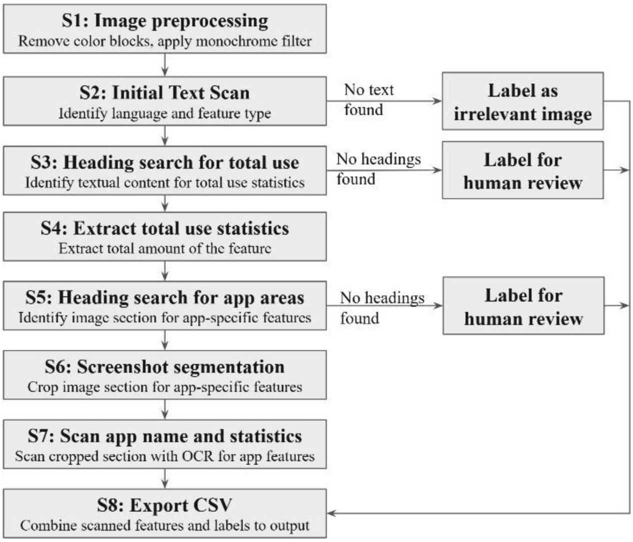


Fig. 2. Image-to-text pipeline in user-donated screenshot analysis in the HappyB study

First, it conducts image preprocessing by applying monochrome filters and removing color blocks. Second, the pipeline conducts a comprehensive scan of the textual content within the screenshot. From the textual content, the pipeline identifies the language, and the type of data displayed (screentime, activation, or notifications). It is worth noting that screenshots are submitted after participants self-identify their device OS (e.g. iOS or Android), so the pipeline does not include OS detection. Third, the pipeline conducts language-specific, type-based, and version-oriented searches for headings of total use statistics within the textual content from screenshots. The total use statistics are the total screentime, activations, and notifications from a day. For example, headings like "pickups" and "yesterday" are used to ascertain that the textual content after the heading is the total number of activations from yesterday. To reduce the errors from the OCR process and enhance the accuracy of heading detection, we implement an edit distance algorithm that matches the OCR-generated text against a pre-defined list of expected headings.

Fourth, with the identified headings for total use statistics, the pipeline extracts the total use statistics for screentime, activation, and notification from the given areas of the text. Fifth, the pipeline searches for headings for app-specific features. Headings like "most used" indicate that the textual content below is app-specific information. Edit distance-based search is also applied here. Sixth, with the identified headings for app-specific features, the pipeline segments the screenshot to isolate the specific areas containing app usage data. This segmentation is guided by spatial coordinates: the upper boundary is set by the heading for app-specific data (e.g., "most used"), the lower boundary is set by key text that appears below the app-specific data (e.g. "view more" or a subsequent heading), and the lateral boundaries are defined by the edges of the display or icons that appear on either side of the app data. Seventh, a second OCR process is then applied to segments of app-specific usage area to extract app names and associated usage statistics. To ensure the extracted app names are correctly aligned with their corresponding usage statistics, the pipeline conducts location-based logic checks to identify and address any discrepancies arising from the second OCR pass. Finally, the pipeline labels images that need human checkups, including the images with the absence of relevant headings, low OCR confidence scores, and identification of data as representing a weekly rather than daily overview. Data extracted from the images are systematically recorded in a CSV file for further analysis.

Table 1. The accuracy rate of the image-to-text pipeline in HappyB screenshot analysis

	IOS			Android		
	Screentime	Activation	Notification	Screentime	Activation	Notification
Total stats	0.93	0.91	0.99	0.95	0.98	0.99
APP #1 name	0.98	0.99	0.99	0.93	0.99	0.97
APP #1 stats	0.98	0.99	0.99	0.88	0.99	0.96
APP #2 name	0.96	0.99	0.99	0.92	0.99	0.95
APP #2 stats	0.97	0.99	0.97	0.93	1.00	0.96
APP #3 name	0.94	0.94	0.96	0.92	0.99	0.93
APP #3 stats	0.93	0.96	0.94	0.91	0.99	0.93

The accuracy rate of our pipeline, as presented in Table 1, achieved an average of 96.07% in extracting text across the seven metrics. Specifically, the pipeline demonstrated over 91% accuracy in identifying total usage statistics. For app-specific names, screentime, and activation/notification counts, it achieved an average accuracy rate of 96.11%. The notably high accuracy in the Android activation category is largely due to many screenshots from an older Android version in our dataset, which only records total activation counts without app-specific details.

3.2 Attrition Pattern at the User and Screenshot Level

The attrition pattern of the HappyB study is illustrated in Fig. 3. Initially, 374 participants were recruited for the baseline survey. On the subsequent day, 309 of these participants completed the first EMA. If each of the 309 participants had remained in the study and submitted screenshots for each of the 14 days, we would expect to receive 4,326 sets of screenshots covering screentime, activation, and notification. However, we received 2,273 sets of screenshots (equating to 6,819 screenshots) from 277.

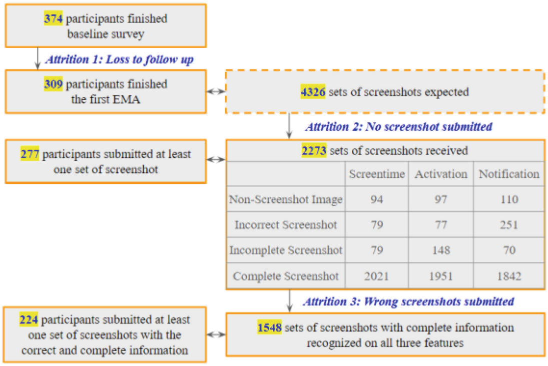


Fig. 3. Flowchart showing the attrition pattern in the HappyB EMA study with user-donated screenshots.

participants. Upon applying the image-to-text pipeline, we found that 301 submitted images (94 screentime, 97 activations, and 110 notifications) were not relevant screenshots but rather unrelated images such as personal photos. Additionally, 407 images (79 screentime, 77 activations, and 251 notifications) were screenshots but failed to include the total usage or the app-specific information we required, such as screenshots depicting only the battery state of the smartphone. Furthermore, 297 images (79 screentime, 148 activations, and 70 notifications) were incomplete, often due to participants cropping out key app-specific areas or usage details. Consequently, we obtained 1,548 sets of complete screenshots, covering all seven metrics across three screenshot types, with 224 participants providing complete data for all three features in at least one EMA.

We also visualize the user attrition pattern in screenshot donation in Fig. 4(A). On day 1, 277 participants submitted screenshots. From day 2 to 5, we observed a daily loss of 18 (6%) to 33 (12%) users. Between day 6 and 13, the attrition slowed to between 5 (2%) and 15 (5%) users per day. However, on the final day, we saw a significant drop with 60 users (21%) not submitting their screenshots. As shown in Fig. 4(B), among the 277 users who submitted screenshots, 59 (21%) submitted fewer than 3 sets, while 179 (65%) submitted more than 5 sets and 129 (47%) submitted more than 10 sets.

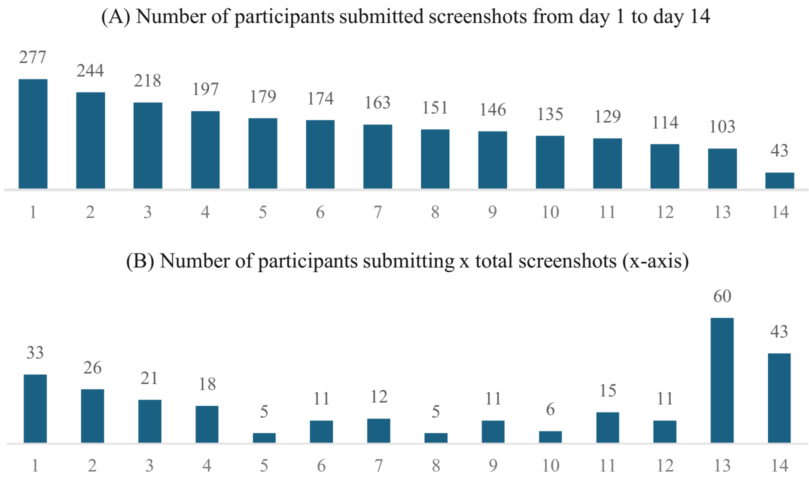


Fig. 4. Pattern of participants attrition in screenshot submission from day 1 to day 14 (A) and the distribution of the total number of screenshots submitted by participants (B)

4 Discussion

4.1 Summary of the User-Donated Screenshot Approach

In our study, we introduced a novel method for collecting social media data using user-donated screenshots. The user-donated screenshot method presents several advantages over traditional objective data collection methods. It is less intrusive than passive screen tracking apps and less burdensome for users than data donation packages. This approach allows for the collection of app-specific usage data across different smartphone operating systems. Notably, most passive sensing apps are limited to Android and cannot track on iOS. Furthermore, many data collection methods like API requests and some passive apps are restricted to adults, excluding adolescents—a key demographic for studying SMU and well-being. The user-donated screenshot approach allows for objective data collection in adolescents.

Success depended on two key factors: the accuracy of text extraction pipeline and user compliance. On the technical side, our image-to-text pipeline achieved an average accuracy rate of 96%, confirming the effectiveness of user-donated screenshots as a reliable source for app usage analysis. At the behavioral level, 26% of participants did not submit any screenshots, 14% submitted incorrect or incomplete ones, and 48% consistently provided complete screenshots for over five days. Our compliance rates align with other EMA study [18]. However, this approach has two main limitations. First, the user-donated screenshot method only captures basic metrics such as screen time, activations, and notifications, lacking depth in insights into social media activities and content. Second, its success heavily relies on user compliance. In the next section, we discuss and propose strategies to reduce attrition in future screenshot donation studies.

4.2 Solutions for Reducing Attrition in Future Screenshot Donation Studies

As far as we know, multiple research teams are now initiating plans for user-donated screenshot data collection. Therefore, it is crucial to summarize our learnings regarding the sources of attrition and strategies for reducing it in future studies. The taxonomy of attrition sources in user-donated screenshot analysis is detailed in Table 2. Of the 374 study participants, 277 submitted at least one set of screenshots, indicating a 26% non-submission rate. This may be attributed to practical burdens (participants feeling overwhelmed by the task), privacy concerns, and other barriers such as misunderstanding the screenshot donation process. To engage the 26% of participants who either lost to follow-up or did not submit any screenshots, future studies could: simplify the EMA design to lessen participant burden; enhance consent information and support to address privacy concerns; and incorporate gamification or rewards to incentivize participation.

Beyond cases of non-submission, 53 out of 374 participants (14%) submitted either non-screenshots or images lacking the requested information. This may result from accidentally uploading incorrect images or a lack of understanding of the required information. To address these issues, future studies could implement a built-in pre-check function in the EMA app that allows users to review images before submission. Additionally, integrating an asynchronous image-to-text pipeline that flags submissions lacking text could prompt users to confirm their uploads. Enhancing education during participant onboarding can also ensure a clearer understanding of the study's objectives.

Table 2. The taxonomy of the source of attrition in user-donated screenshot analysis

<table>
<tr><th>Source of attrition</th><th>Solution in future studies</th></tr>
<tr><td>1. Loss to follow-up from EMA
Due to practical burden, privacy concerns, or barriers</td><td rowspan="2">a. Simplify EMA design
b. Address privacy concerns
c. Incentivize participants</td></tr>
<tr><td>2. Stayed in EMA but no screenshots submitted
Due to privacy concerns, barriers, or knowledge gap</td></tr>
<tr><td>3. Non-screenshot images submitted
Due to uploading the wrong image from album</td><td rowspan="3">d. Implement screenshot pre-check
e. Async verification via OCR
f. Enhance participant onboarding and education</td></tr>
<tr><td>4. Screenshots lack requested information
Due to lack of understanding of the study instruction</td></tr>
<tr><td>5. Screenshots with incomplete information
Due to lack of understanding of the study instruction</td></tr>
</table>

Acknowledgments. This study was funded by the Swiss National Science Foundation (grant P500PS_202974) and the NIH/NIMH (grant 1R21HD115354-01).

References

1. Massarat, E.A.V.: Risa Gelles-Watnick and Navid: Teens, Social Media and Technology (2022)

2. NW, 1615 L. St, Washington, S. 800, Inquiries, D. 20036 U.-419-4300 | M.-857-8562 | F.-419-4372 | M.: Social Media Fact Sheet (2024). https://www.pewresearch.org/internet/fact-sheet/social-media/
3. CDC: Youth Risk Behavior Surveillance Data Summary and Trends Report: 2011–2021 (2021). https://www.cdc.gov/healthyyouth/data/yrbs/pdf/YRBS_Data-Summary-Trends_Report2023_508.pdf
4. Advisory of U.S. Surgeon General: Social Media and Youth Mental Health — Current Priorities of the U.S. Surgeon General. https://www.hhs.gov/surgeongeneral/priorities/youth-mental-health/social-media/index.html
5. High, A.C., Ruppel, E.K., McEwan, B., Caughlin, J.P.: Computer-mediated communication and well-being in the age of social media: a systematic review. J. Soc. Pers. Relationsh. **40**, 420–458 (2023). https://doi.org/10.1177/02654075221106449
6. Meier, A., Reinecke, L.: Computer-mediated communication, social media, and men-tal health: a conceptual and empirical meta-review. Commun. Res. **48**, 1182–1209 (2021). https://doi.org/10.1177/0093650220958224
7. Kross, E., Verduyn, P., Sheppes, G., Costello, C.K., Jonides, J., Ybarra, O.: Social media and well-being: pitfalls, progress, and next steps. Trends Cogn. Sci. **25**, 55–66 (2021). https://doi.org/10.1016/j.tics.2020.10.005
8. Orben, A.: Teenagers, screens and social media: a narrative review of reviews and key studies. Soc. Psychiatry Psychiatr. Epidemiol. **55**, 407–414 (2020). https://doi.org/10.1007/s00127-019-01825-4
9. Ohme, J., Araujo, T., de Vreese, C.H., Piotrowski, J.T.: Mobile data donations: as-sessing self-report accuracy and sample biases with the iOS Screen Time function. Mob. Media Commun. **9**, 293–313 (2021). https://doi.org/10.1177/2050157920959106
10. Sultan, M., Scholz, C., van den Bos, W.: Leaving traces behind: using social media digital trace data to study adolescent wellbeing. Comput. Hum. Behav. Rep. **10**, 100281 (2023). https://doi.org/10.1016/j.chbr.2023.100281
11. Ohme, J., et al.: Digital trace data collection for social media effects research: APIs, data dona-tion, and (Screen) tracking. Commun. Methods Meas. 1–18 (2023). https://doi.org/10.1080/19312458.2023.2181319
12. Davidson, B.I., et al.: Platform-controlled social media APIs threaten Open Science (2023). https://osf.io/ps32z
13. van Driel, I.I., Giachanou, A., Pouwels, J.L., Boeschoten, L., Beyens, I., Valkenburg, P.M.: Promises and pitfalls of social media data donations. Commun. Methods Meas. **16**, 266–282 (2022). https://doi.org/10.1080/19312458.2022.2109608
14. Stier, S., Breuer, J., Siegers, P., Thorson, K.: Integrating survey data and digital trace data: key issues in developing an emerging field. Soc. Sci. Comput. Rev. **38**, 503–516 (2020). https://doi.org/10.1177/0894439319843669
15. Reeves, B., Robinson, T., Ram, N.: Time for the human screenome project. Nature **577**, 314–317 (2020). https://doi.org/10.1038/d41586-020-00032-5
16. Reeves, B., et al.: Screenomics: a framework to capture and analyze personal life experiences and the ways that technology shapes them. Hum. Comput. Interact. **36**, 150–201 (2021). https://doi.org/10.1080/07370024.2019.1578652
17. Ram, N., Haber, N., Robinson, T.N., Reeves, B.: Binding the person-specific approach to modern AI in the human screenome project: moving past generalizability to transferability. Multivar. Behav. Res. 1–9 (2023). https://doi.org/10.1080/00273171.2023.2229305
18. Tonkin, S., et al.: Evaluating declines in compliance with ecological momentary assessment in longitu-dinal health behavior research: analyses from a clinical trial. J. Med. Internet Res. **25**, e43826 (2023). https://doi.org/10.2196/43826

Extending VRAT: From 3D Eye Tracking Visualization to Enabling ACT-R to Interact with Virtual Reality Environments

Amir Bagherzadeh(✉) and Farnaz Tehranchi

Pennsylvania State University, University Park, PA 16802, USA
abb6024@psu.edu

Abstract. Cognitive architectures have been used to understand learning new tasks, forgetting, error making or navigation and mental map development. The insight cognitive architectures provide can be utilized to understand human behavior at a cognitive level. In this paper, we report on the recent developments of our toolbox VRAT which will provide a framework for designing experiments, collecting and analyzing data, and then developing cognitive models that can see and interact with the environment similar to users. The differentiating factor of our toolbox from previously developed tools is that its abilities are extended to Virtual Reality (VR). The ability to create three-dimensional visual scenes and to measure responses (i.e., gaze data, head and hand movements data) to the visual stimuli enables behavioral researchers to test hypotheses in a way and scale that were previously unfeasible. The difficulty facing the researcher is that sophisticated 3D graphics engines (e.g., Unity) have been created for game designers rather than behavioral scientists. To overcome this barrier VRAT provides a plug-and-go design to help researchers to convert their 2D experiments into VR. It also enables eye tracking and eye tracking visualization in all VR experiments. Enabling researchers to collect and analyze data more efficiently. Additionally, our tool enables (a) a straight forward transition from 2D environment design to 3D, (b) an efficient way to use data collection and visualization framework and (c) an interaction method for cognitive models to extend their capabilities and see and interact with VR environments similar to users.

Keywords: Virtual Reality · Cognitive Architectures · Eye tracking · Study Design

1 Introduction

1.1 Cognitive Architectures for Human Performance Simulation

Cognitive architectures are computational frameworks that can be used to develop computational models of human cognitive processes (Newell, 1994). These architectures have been instrumental in advancing our understanding of human cognition in specific task environments and have been applied to the development of various intelligent systems and agents, such as cognitive robots. This work focuses on extending the abilities

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 90–99, 2024.
https://doi.org/10.1007/978-3-031-72241-7_9

of cognitive architectures, such as Soar (Laird et al., 2012), ACT-R (Anderson, 2007), and CLARION (Sun, 2006), PyIBL (Gonzalez et al., 2003) to interact with virtual environments. Cognitive architectures do not have a built-in capability to interact with any environment. This enables cognitive modelers to make predictions about how users are making decisions. In this work, we aim to enable researchers to design experiments, collect data, analyze the collected data, and develop cognitive models that mirror users' behavior in Virtual Reality (VR). VR allows researchers to create highly controlled and immersive environments that facilitate more accurate statistical analysis of the results and detailed observations of cognitive processes. This capability enables the exploration of complex and dynamic scenarios that are difficult to replicate in traditional 2D experiments. Additionally, VR can integrate various sensory modalities and interactions, providing a richer context for studying human behavior and improving the precision of cognitive models while opening the door to factors that modifying them is not feasible in 2D (such as visual distractions or sudden changes in the environment). This approach opens new avenues for understanding cognitive phenomena and developing robust theories in cognitive science, ultimately leading to more comprehensive insights and applications in fields like neuropsychology, affective cognition, and clinical research (Katona, 2021; Oker et al., 2022).

1.2 Virtual Reality, the Future of Behavioral Research

The potential advantages of VR in behavioral research have been recognized for at least two decades (Loomis et al., 1999). Recent advancements in technology and the availability of hardware and software have made VR a feasible tool for a wide range of behavioral researchers, rather than being limited to a small number of specialist VR labs. For instance, researchers now have access to powerful software engines, such as Unity (also known as Unity3D), which allow the creation of rich 3D environments. Unity is a widely used 3D game engine for developing video games, animations, and other 3D applications, and its popularity continues to grow.

Unity has well-developed systems in place for rich graphics, realistic physics simulation, particles, animations, and more. However, it does not contain any features specifically designed for the needs of human behavior researchers. We set out to produce an open-source software resource that would empower researchers to exploit the power of Unity for behavioral studies.

1.3 Cognitive Modeling in VR

Cognitive model development in VR benefits virtual character design, as demonstrated by Best and Lebiere (2006) using ACT-R in military training simulations. VR enables precise monitoring of human actions, allowing for adjustment of virtual agents' behavior. Cognitive architectures, VR, and character design converge in various contexts, including modeling game player behavior (Moon & Anderson, 2012) and studying embodied cognition (Smart & Sycara, 2015). Other applications include simulating social behaviors, developing believable game characters (Arrabales et al., 2009; Bringsjord et al., 2005), and digital modeling (Lawson & Burnett, 2015).

VR systems offer new opportunities for behavioral research by displaying 3D stimuli without adhering to Newtonian mechanics (Wann & Mon-Williams, 1996). They support natural interactions and precise movement measurements (Felnhofer et al., 2015), while their affordability facilitates non-laboratory research.

1.4 Tools to Enable Interaction for Cognitive Architectures

Cognitive models, while proficient at simulating internal cognitive processes, often lack the capability to perceive external environments or execute actions that impact the physical world, presenting a boundary between simulated cognition and tangible user interactions (Tehranchi & Ritter, 2018). Nearly all existing cognitive models fail to directly interact with uninstrumented external task environments (Tehranchi & Ritter, 2017), significantly limiting their ability to accurately model and predict human behavior in real-world interactive tasks.

Several approaches have been used by modelers to provide some level of interaction between cognitive models and task environments, such as ACT-R/PM (Fleetwood & Byrne, 2002), Cognitive Code (Salvucci, 2013), MONGSU (Ong & Ritter, 1994), JSON ACT-R (Hope et al., 2014), and ACT-CV (Halbrügge, 2013). However, these approaches suffer from limitations such as forcing the use of a single programming language, requiring interfaces to be re-implemented for each task, adding new modules that must be re-configured, or only working with special software (Ritter et al., 2007).

To enable cognitive models to perform interactive tasks in the same way humans do, the models need visual perception and motor action capabilities - in other words, simulated eyes and hands (Kieras, 2009; Tehranchi, 2020). What is needed is a general approach that allows cognitive models to interact with the same unmodified interfaces used by human subjects (Ritter et al., 2002). Tehranchi and Ritter (2017, 2018) proposed an Eyes and Hands model that aims to provide this universal interaction capability to facilitate the development and application of more complete and accurate cognitive models.

However, existing tools lack modularity capabilities and are not expandable. They are either application-specific or rely on traditional template-matching algorithms, making them incapable of finding visual stimuli with minimal modifications. Overcoming the interaction limitations of current models is necessary to enable them to more accurately capture and simulate human interactive behavior (Byrne, 2001; Kieras, 2009). VisiTor is one of the latest interaction tools that has aimed to bring modularity, superior performance in terms of accuracy in finding visual objects and computation time and ease of use to cognitive models with the intention to interact with external environments (Bagherzadeh & Tehranchi, 2022). It was first used in simulating probability learning in a dynamic environment (Bagherzadehkhorasani & Tehranchi, 2023a, b). Ever since it has been used in the longest cognitive model of driving (Wu et al., 2023), simulating error making in spreadsheets (Bagherzadehkhorasani & Tehranchi, 2023a, b). This tool has been developed in Python and is an open-source toolbox[1]. Researchers can add new modules to VisiTor by adding new functionality to the code. Here, to enable the interaction with VR environments, we extended VisiTor through a new module that can be

[1] https://github.com/HCAI-Lab/VisiTor.

added to the VisiTor directory and enabled it to simulate a VR headset and controllers. Using VisiTor, researchers can change the location and the orientation of the headset and controllers. This feature enables new cognitive model development opportunities in VR that can extend our understanding of human behavior.

2 VRAT: A Toolkit Towards VR Research

2.1 Environment Design and Data Collection

Designing VR environments demands advanced programming skills and familiarity with game engines (e.g., Unity or Unreal Engine). Besides, there are a large number of experiments that are already designed for 2D displays, making it financially and logistically challenging to redesign them for VR.

VRAT (Bagherzadeh & Tehranchi, 2024) addresses these issues by providing an interface to deploy 2D experiments into VR, along with accurate eye tracking ability for large displays, which traditional screen-based eye trackers struggle with. Traditional screen-based eye trackers fail to track users' gazes effectively across multiple screens and are cost-prohibitive for large displays. Wearable eye trackers, though suitable for these purposes and have extensively been utilized by researchers in this area (Kocejko et al., 2015; MacInnes et al., 2018), pose challenges such as difficulty in mapping eye gazes to display pixel coordinates and requiring special lenses for eyeglass wearers, further increasing costs. These issues motivated us to develop a more cost-effective and accurate method using VR. Thus, (a) the complexity of mapping eye gazes into pixel coordinates and (b) the heavy financial investment that is required for wearable eye trackers, motivated us to find a cheaper and more accurate method.

Our previous research (Bagherzadeh; & Tehranchi, 2024) demonstrated that the VR headset (Oculus Quest Pro) provided consistent eye tracking accuracy across users, often matching or surpassing traditional screen-based eye trackers, especially for large displays (e.g., 35 inches or above). Essentially, VR brings the advantages of wearable eye trackers while providing easier analysis capability as the researcher has control over the VR scene and can map users' eye gazes to accurate hit points to visual stimuli in the scene. Originally, VRAT was introduced with two distinctive environment designs (Fig. 1 shows VRAT different predefined environments: left column is the experiment room simulation, and the right column is the display broadcast into a black space).

We initially designed VRAT to cast 2D display content into a VR environment and record users' gaze data on the 2D display in VR (Fig. 1.b). Also, VRAT made visualizing gaze heatmaps on 2D videos of the display content possible. In this paper, we extend VRAT eye-tracking capabilities and gaze data visualization. We introduce a new Unity package that enables eye tracking and data collection. VRAT can now record users' gaze data including eye locations and gaze direction in 3D in any VR environment in which is beyond casting 2D display in VR environment. Collected data can be exported in a CSV file. Researchers can visualize users' gaze data overlayed on different 3D objects within the VR environment and analyze fixation times on areas of interest (AOIs). Along with the visualization, the fixation times on the objects and AOIs in the environment will be saved in another file for post analysis. Furthermore, VRAT can track dynamic AOIs that is an improvement in comparison to other VR eye gaze visualization tools such as Salient360! (David et al., 2024), which works best with flat static 360 images. VRAT integrates eye tracking with dynamic 3D environments, enhancing post-analysis efficiency and insight into user behavior.

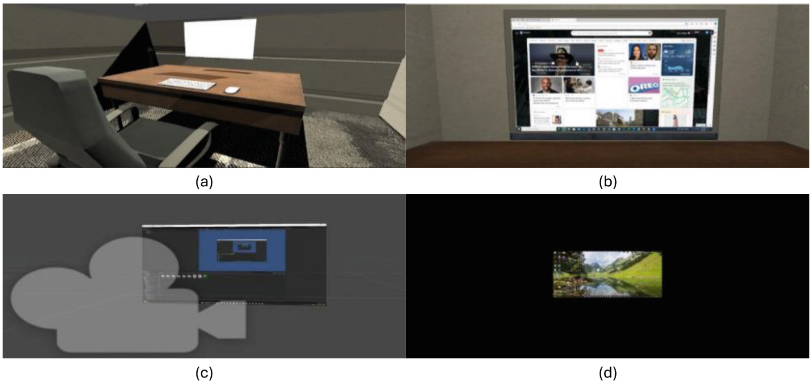


Fig. 1. Screenshots of the environments in VRAT. The left column (a and c) is the Unity game engine editor, and the right column (b and d) is the user's point of view of the VRAT environment. The top row (a and b) is the 2D display broadcast into a VR simulation.

2.2 VisiTor for VR Interactions

Experiment design, data collection, and analysis are essential intermediate steps toward cognitive modelers' ultimate goal: developing cognitive models that simulate user behavior. Without these models, our work lacks purpose. Therefore, we present a significant update to VisiTor, now featuring an interface for direct interaction with VR environments by emulating HTC VIVE headsets and controllers.

VisiTor establishes a TCP connection, allowing users to connect and execute actions via any programming language, such as Python or LISP. This tool facilitates cognitive models and their interaction with VR environments without requiring backend access to the game engine, making it compatible with VR applications in the Steam library, such as driving Desert Bus VR or ER simulator (i.e., Surgeon Simulator: Experience Reality).

To demonstrate VRAT's potential, we utilized it for data collection, visualization, and developing a PyIBL agent to navigate a maze in VR. The environment, designed in Unity, examines the effects of different visual stimuli on visual search. There are two maze versions: one with colored walls and another with home appliances as visual cues (See Fig. 2).

Users interact with the environment using the headset and controllers, with teleportation as the locomotion method. At each tile, instructions are provided on one of the walls, asking the user to look for a specific visual stimulus and move toward it. Once the user reaches the destination point, the game terminates, and the user is placed back at the origin point and is asked to reach the destination without instructions to test visual search and memory implications of different types of visual stimuli.

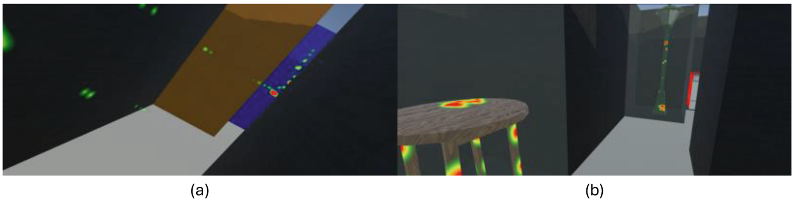


Fig. 2. Illustration of the 3D heatmap in two mazes with (a) color (i.e., brown and blue) and (b) object visual cues (i.e., table and lamp). (Colour figure online)

We added the VRAT eye tracking package to collect the eye gaze positions and rotations. In the next step, the environment was duplicated to add the gaze visualization package of VRAT for posttest gaze analysis of the users. Figure 2 shows the environment and the 3D heatmap overlay using VRAT. A PyIBL model was developed to interact with the maze using VisiTor and determine if an instance-based learning cognitive model could learn to follow instructions. The PyIBL model reaches the destination point following the state of visual cues similar to the model's point of view that is extracted from the maze' map into a structured format (See an example in Fig. 3). The model uses visual cues in each direction alongside the direction given by the instruction. Then the model decides which direction to move to.

If the direction matches the direction suggested by the instruction, the model will be rewarded for the action and penalized otherwise. The model does not yet simulate users' behavior but showcases how the VR emulation enables researchers to utilize VisiTor to train a model and play a game similar to how users play the game. Figure 4 illustrates the learning process of the PyIBL model over time by showing the proportion of correct answers. Initially, the model makes random decisions. Hence, the proportion of correct answers fluctuates around the 50% threshold. However, after the 500th Decision Point, the PyIBL model learns to always make the correct decision. The proportion of the correct answers has a positive slope at almost all the Decision points after 500th point and the proportion is converging to 1, meaning the model is always making the correct decision in any state that it is placed in.

For this work, the model was trained using structured data extracted from the maze map and after being trained, the model was tested in the VR environment. The model successfully moved its head and controllers, performed required locomotion actions, and navigated to the destination point.

3 Conclusion and Future Work

In this paper, we presented the latest developments and extensions to our Virtual Reality Analysis Toolkit (VRAT). VRAT provides a framework for all the steps that are required in developing cognitive models in VR (i.e., designing experiments, collecting and analyzing data, and developing cognitive models that can perceive and interact with VR environments in the same way as users.)

The key advantages of VRAT are:

1. It simplifies the process of converting 2D experiments into VR, making VR research more accessible.
2. It enables efficient and accurate eye tracking and gaze visualization in dynamic 3D virtual environments.
3. The integration with VisiTor allows cognitive architectures like ACT-R to directly interact with the same VR interfaces utilized by users. This opens up new opportunities for studying a broader range of cognition and behavior in realistic virtual environments.

To demonstrate these capabilities, we designed a maze study and used VRAT to collect data, visualize gaze patterns, and develop a PyIBL model integrated with VisiTor that could navigate the VR maze by following instructions and cues given, similar to users. The model's behavior was made more realistic by incorporating timing data from human subjects.

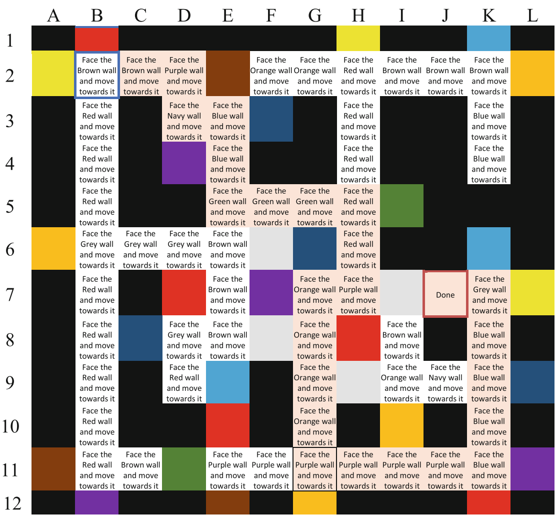


Fig. 3. A holistic view of the maze. Users and PyIBL model start from point A2 and need to follow the instructions given at each time to reach destination J7. At each tile, the PyIBL model makes actions based on the visible visual cues and the instruction. For example, if the model is at point E5, the model makes a decision based on the following instances: {North: Brown, South: Blue, West: Black, East: Green, Instruction: Green} (Colour figure online)

VRAT provides the structure to move to VR. However, there is a long way ahead. Currently, cognitive architectures do not support the types of motor movements that are required for interacting with VR such as head and body movements, and also moving controllers in the space. Hence, research is needed to understand these movements and develop devices to accommodate these actions.

Moreover, there is a need for more data collection in the maze to make the PyIBL model more realistic, GOMS time should be extracted from users who completed the test to set pause times within the PyIBL code, simulating the completion time of each action[2].

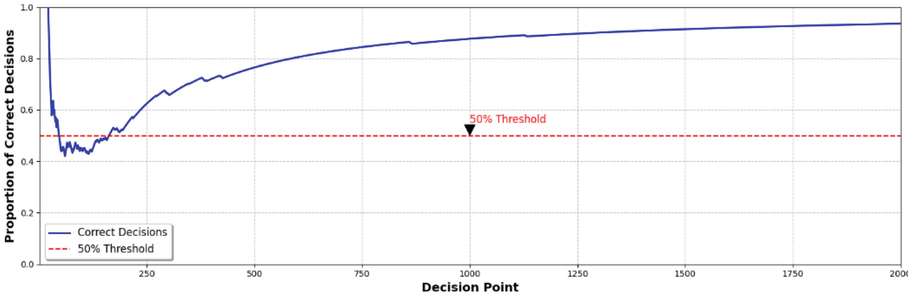


Fig. 4. The blue line shows the proportion of correct decisions over time. The positive slope of the line (after approximately the 500th decision point) shows that the model almost always makes the correct decision at any decision point after training for enough iterations. (Color figure online)

In summary, VRAT aims to empower behavioral researchers to harness the potential of VR for studying human cognition and behavior in more naturalistic, interactive settings. By bridging the gap between powerful game engines like Unity and the needs of cognitive modelers, our toolkit enables the development of artificial models that can more closely approximate human-like perception, action, and learning in virtual environments. These advances will lead to richer insights and more ecologically valid models in the cognitive and behavioral sciences.

References

Anderson, J.R.: How Can the Human Mind Occur in the Physical Universe? Oxford University Press, Oxford (2007)

Arrabales, R., Ledezma, A., Sanchis, A.: Towards conscious-like behavior in computer game characters. In: 2009 IEEE Symposium on Computational Intelligence and Games (2009)

Bagherzadeh, A., Tehranchi, F.: Comparing cognitive, cognitive instance-based, and reinforcement learning models in an interactive task. In: Proceedings of ICCM-2022–20th International Conference on Cognitive Modeling (2022)

Bagherzadeh, A., Tehranchi, F.: Computer-Based Experiments in VR: A Virtual Reality Environment to Conduct Experiments, Collect Participants' Data and Cognitive Modeling in VR Proceedings of ICCM-2024–22nd International Conference on Cognitive Modeling (2024)

Experiments, Collect Participants' Data and Cognitive Modeling in VR Proceedings of the 22th International Conference on Cognitive Modelling (Accepted), Tilburg, Netherlands

Bagherzadehkhorasani, A., Tehranchi, F.: Automatic error model (AEM) for user interface design: a new approach to include errors and error corrections in a cognitive user model. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting (2023a)

Bagherzadehkhorasani, A., Tehranchi, F.: A pipeline for analyzing decision-making processes in a binary choice task. In: Proceedings of the 21th International Conference on Cognitive Modelling, Amsterdam, Netherland (2023b)

[2] A demo of the PyIBL model is available at the GitHub page of VRAT (https://github.com/HCAI-Lab/Virtual-Reality-Analysis-Tool-VRAT--).

Best, B.J., Lebiere, C.: Cognitive agents interacting in real and virtual worlds. Cognition and multi-agent interaction: from cognitive modeling to social simulation, pp. 186–218 (2006)

Bringsjord, S., Khemlani, S., Arkoudas, K., McEvoy, C., Destefano, M., Daigle, M.: Advanced synthetic characters, evil, and E. Game-On (2005)

Byrne, M.D.: ACT-R/PM and menu selection: applying a cognitive architecture to HCI. Int. J. Hum. Comput. Stud. **55**(1), 41–84 (2001)

David, E., Gutiérrez, J., Võ, M.L.-H., Coutrot, A., Da Silva, M.P., Le Callet, P.: The Salient360! toolbox: handling gaze data in 3D made easy. Comput. Graph. 103890 (2024)

Felnhofer, A., et al.: Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios. Int. J. Hum. Comput. Stud. **82**, 48–56 (2015)

Fleetwood, M.D., Byrne, M.D.: Modeling icon search in ACT-R/PM. Cogn. Syst. Res. **3**(1), 25–33 (2002)

Gonzalez, C., Lerch, J.F., Lebiere, C.: Instance-based learning in dynamic decision making. Cogn. Sci. **27**(4), 591–635 (2003)

Halbrügge, M.: ACT-CV: bridging the gap between cognitive models and the outer world. Grundlagen und anwendungen der mensch-maschine-interaktion **10**, 205–210 (2013)

Hope, R.M., Schoelles, M.J., Gray, W.D.: Simplifying the interaction between cognitive models and task environments with the JSON network interface. Behav. Res. Methods **46**, 1007–1012 (2014)

Katona, J.: A review of human–computer interaction and virtual reality research fields in cognitive InfoCommunications. Appl. Sci. **11**(6), 2646 (2021)

Kieras, D.: Model-based evaluation. In: Human-Computer Interaction, pp. 309–326. CRC Press (2009)

Kocejko, T., Ruminski, J., Wtorek, J., Martin, B.: Eye tracking within near-to-eye display. In: 2015 8th International Conference on Human System Interaction (HSI) (2015)

Laird, J.E., Kinkade, K.R., Mohan, S., Xu, J.Z.: Cognitive robotics using the soar cognitive architecture. In: Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (2012)

Lawson, G., Burnett, G.: Simulation and Digital Human Modelling, pp. 201–218. CRC Press, Boca Raton (2015)

Loomis, J.M., Blascovich, J.J., Beall, A.C.: Immersive virtual environment technology as a basic research tool in psychology. Behav. Res. Methods Instrum. Comput. **31**(4), 557–564 (1999)

MacInnes, J.J., Iqbal, S., Pearson, J., Johnson, E.N.: Wearable eye-tracking for research: automated dynamic gaze mapping and accuracy/precision comparisons across devices (2018). BioRxiv, 299925

Moon, J., Anderson, J.: Modeling millisecond time interval estimation in space fortress game. In: Proceedings of the Annual Meeting of the Cognitive Science Society (2012)

Newell, A.: Unified Theories of Cognition. Harvard University Press, Cambridge (1994)

Oker, A., Pecune, F., Vallverdu, J.: Virtual reality for neuropsychology and affective cognitive sciences: theoretical and methodological avenues for studying human cognition. Front. Virtual Real. **3**, 1100387 (2022)

Ong, R., Ritter, F.: Mechanisms for routinely tying cognitive models to interactive simulations. In: HCI International'95: Poster Sessions Abridged Proceedings (1994)

Ritter, F.E., Kukreja, U., Amant, R.S.: Including a model of visual processing with a cognitive architecture to model a simple teleoperation task. J. Cogn. Eng. Decis. Mak. **1**(2), 121–147 (2007)

Ritter, F.E., Van Rooy, D., St. Amant, R.: A user modeling design tool based on a cognitive architecture for comparing interfaces. In: Kolski, C., Vanderdonckt, J. (eds.) Computer-Aided Design of User Interfaces III, pp. 111–118. Springer, Dordrecht (2002). https://doi.org/10.1007/978-94-010-0421-3_10

Salvucci, D.D.: Integration and reuse in cognitive skill acquisition. Cogn. Sci. **37**(5), 829–860 (2013)

Smart, P.R., Sycara, K.: Using a cognitive architecture to control the behaviour of virtual robots (2015)

Sun, R.: Cognition and Multi-agent Interaction. Cambridge University Press, Cambridge (2006)

Tehranchi, F.: An eyes and hands model: Extending visual and motor modules for cognitive architectures The Pennsylvania State University] (2020)

Tehranchi, F., Ritter, F.E.: An eyes and hands model for cognitive architectures to interact with user interfaces. In: MAICS (2017)

Tehranchi, F., Ritter, F.E.: Modeling visual search in interactive graphic interfaces: adding visual pattern matching algorithms to ACT-R. In: Proceedings of ICCM-2018-16th International Conference on Cognitive Modeling (2018)

Wann, J., Mon-Williams, M.: What does virtual reality NEED?: human factors issues in the design of three-dimensional computer environments. Int. J. Hum. Comput. Stud. **44**(6), 829–847 (1996)

Wu, S., Bagherzadeh, A., Ritter, F., Tehranchi, F.: Long road ahead: lessons learned from the (soon to be) longest running cognitive model. In: 21st International Conference on Cognitive Modeling (ICCM) at the University of Amsterdam, The Netherlands (2023)

A Tool for Distributed Collaborative Causal Discovery

Alexey Tregubov(✉), Jeremy Abramson, Stephen Schwab, and Jim Blythe

USC Information Sciences Institute, Marina del Rey, CA, USA
{tregubov,abramson,schwab,blythe}@isi.edu

Abstract. The development of accurate causal models is crucial for achieving and explaining desired outcomes that require interventions. Building these models efficiently requires combining available data with expert causal knowledge. Often experts have unique data and model insights, but sharing them is challenging due to privacy or security concerns. Federated machine learning addresses similar issues by allowing multiple sites to collaborate on a common model without sharing private datasets. This paper introduces CCaT, a distributed causal discovery tool enabling collaborative development of a shared causal model while preserving local models and data privacy. CCaT allows each site to evaluate and refine the shared model using its private dataset, sharing only summary statistics or suggested new causal relations. The tool supports maintaining distinct local causal models, as analysts can choose to adopt or change parts of the shared model. CCaT enhances the accuracy of causal models by leveraging diverse expertise and data, achieving a generality and accuracy unattainable by individual sites. We present several common scenarios with the CCaT to demonstrate its effectiveness.

1 Introduction

An accurate causal model is essential when action must be taken to achieve some desired outcome [2,8]. However, many domains lack a formal causal model. These can be built most efficiently by combining available data with elements of causal knowledge from human experts. Tools such as UPREVE [10] and CauseWorks [4] allow users to create and edit graphical causal models and provide an interface to evaluate models against data and suggest modifications. In many domains, several experts may each possesses a distinct subset of the larger causal model and each may have access to a unique dataset. Each expert is motivated to share aspects of their causal model with others in order to improve all of their models, but may be unable to share their data in detail. In medical settings, for example, experts from different groups may not be able to share data due to privacy laws. In the intelligence community, security concerns may prohibit data sharing.

In this paper, we describe a distributed causal discovery tool that supports a collaboration between sites that pool their expertise to develop a shared causal model of their domain, without sharing any other information about their private data. This approach is analogous to sharing weights or gradients on a deep

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 100–109, 2024.
https://doi.org/10.1007/978-3-031-72241-7_10

network in the federated learning approach [3], which has received considerable attention in recent years, but does not support sharing causal knowledge. In our approach, each site maintains a private dataset that it uses to evaluate the shared model and optionally to mine new causal relationships. Each site shares summary statistics on a developing shared model based on their local dataset, without revealing that dataset, and may also introduce new variables and new causal links into the model with associated summary statistics. In contrast to most federated learning approaches, since causal discovery typically requires human input, the tool supports maintaining a local causal model at each site which may be different from the shared causal model, since analysts may choose whether to adopt or ignore any part of the shared model.

We present CCaT, (Collaborative Causal discovery Tool), an implemented collaborative causal discovery tool that supports groups in contributing to a shared causal model while maintaining a distinct local model, and evaluating both models with private data. We describe several scenarios in which CCaT supports several groups building a shared causal model with an accuracy that may not be achievable at any one site individually, under assumptions that we detail about the distribution of observational data and expert knowledge among the groups. In some cases, the shared model can achieve a level of generality that would not be possible at each individual site, allowing more rapid sharing of causal knowledge.

In the next section we define the collaborative causal discovery task that our tool is designed to solve. The following section introduces CCaT and its UI and analytical capabilities that support the task. We then describe several scenarios that illustrate the power of this tool, and finally discuss privacy considerations and future work.

2 Related Work

UPREVE offers both a GUI for causal discovery and a set of integrated algorithms and metrics [10]. However, it does not support combined human-machine discovery or collaborative causal discovery with private data. Stefano et al. introduce a paradigm for multi-agent collaborative learning of causal networks in cases of partial observability, in which agents may ask others to perform experiments on their behalf [6]. Meganck et al. propose an algorithm for distributed learning of multi-agent causal models, an extension of causal bayesian networks to a distributed domain [7]. However, neither of these approaches integrates human insight with algorithmic results via a GUI. Causeworks is a GUI that supports collaborative construction of causal models, overlaying analytic results on a shared model built by experts [4,5]. Compared to CCaT, it does not support private data or private models; instead, all participants have access to all relevant data. CCaT is designed for a collaboration mode where local models and data remain private unless they are shared (partially or in full).

3 Problem Description

There are many domains (healthcare, financial institutions, education) where information is collected in a distributed manner and cannot be shared outside the local information system (for privacy, security and other reasons). When training data cannot be shared directly (or at all), we seek to share model parts to the extent it is feasible and gain common benefits for all distributed sites without breaking security and privacy constraints.

In this paper, we focus on an approach for collaborative causal discovery, designed for settings where distributed sites cannot share their private local data but can choose what model aspects to share. These aspects include: individual rules learned by each site, parts of a local causal model (e.g. partial DAGs), model parameters and aggregated data properties (e.g. best thresholds discovered by each site and/or summary evaluation statistics computed from their local data). Shared aspects (models, rules, thresholds) can be tested and incorporated by other sites and a global evaluation of the model can be approximated from individual site summary statistics. New findings can then again be shared with everybody in a continuous, iterative model improvement process. This approach allows distributed sites jointly to discover confounding bias and incorporate new features or more relevant features into the model, benefiting from the experience of each site while maintaining data privacy. We do not yet support results from explicit interventions by individual sites into our model but intend to do so in a future extension.

4 CCaT Overview

CCaT is a distributed causal discovery tool with a web interface. Each distributed site runs its own instance of a tool and shares selected models and model parameters with others. Shared models and parameters are distributed across all sites. The CCaT web interface provides access to local and shared causal models. These models can be edited, then fitted and tested on local historical data. CCaT allows sharing causal models as a whole or just their parts (fit data such as thresholds, weights, etc.).

In this paper we use a social media domain example to illustrate CCaT, in which causal models are developed to explain bans on a social media platform (e.g. censorship policy valuations on the platform or effective bans that are triggered by internet trolls or other actors). Each site has its own historical observational data on what actions social media influencers/users took in the past (posts, comments, likes etc.) and observed ban events. This historical data has information about several influencer/user accounts and their history of bans. In CCaT each causal model is called a policy, and consists of one or more rules combined together. Causal relationships among features and with the outcome variable (ban or no ban label) are reflected in the model's DAGs.

The home screen of the CCaT user interface provides access to three tabs: (1) local and shared policies viewer to view and compare policies; (2) rules library with all available rules; (3) editor for DAG editing and evaluation.

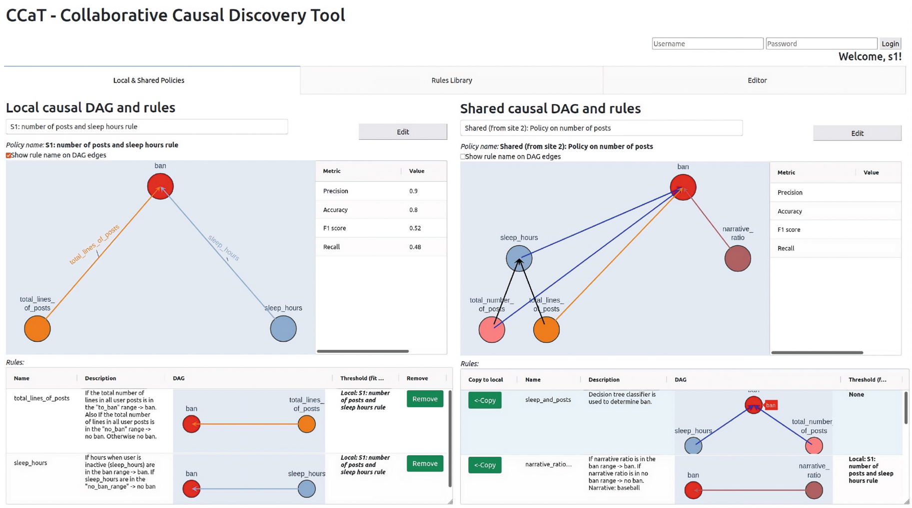


Fig. 1. CCaT Local and shared policies viewer tab of the CCaT.

The local and shared policies viewer tab (Fig. 1) allows experts to see details (DAGs, constituent rules, evaluation summary) about local (left panel) and shared policies (right panel). Model fragments from shared policies can be copied to local models. Policies can be opened in the editor tab.

Model fragments in the rules library (Fig. 2) are collected from both shared and local policies. For each fragment, all available thresholds (or model fit data) are listed in the "Available thresholds" column. In this social media example, there are three sets of general (or global) thresholds: local historical data fit from the current site and shared thresholds from two other sites. The list of thresholds also includes thresholds that were shared as part of some rule. Both of these groups of thresholds are displayed in the "Available thresholds" column.

Once the historical data with labels is loaded in CCaT, it generates a set of simple model fragments (feature-to-outcome causal links). These fragments, which we refer to as rules, can be found in the rules library tab (Fig. 2). Each rule is associated with some classifier model (e.g. a simple threshold or decision tree classifier). These rules serve as building blocks for policy models. Each rule can be added to the model currently opened in the DAG editor tab.

Figure 3 shows DAG editor and policy/model evaluation interface. Rules can be added or removed from the policy currently in the editor. For each rule, the expert can choose what parameters to use (thresholds, fit data) in the Thresholds column. Once all rules have thresholds/fit data selected, the evaluation panel will immediately show how the current model scores on local historical data.

The evaluation section of the DAG editor tab shows precision, accuracy, recall and F1 score. The histogram on the right shows how many times each rule in the policy was triggered (multiple rules can be triggered simultaneously). If policy uses rules with decision tree classifiers, the bar chart on the right also shows the importance of each feature using Gini impurity.

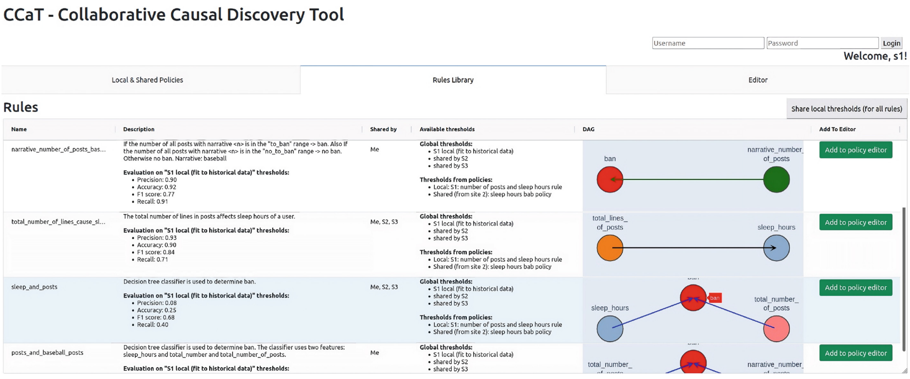


Fig. 2. CCaT Rules library tab of the CCaT.

5 Scenarios

In this section we show several common scenarios with CCaT. For demonstration purposes we use a social media domain where a modeling expert develops causal models explaining user bans on the platform. We assume that the modeler has observational data on the activities of several social media influencers/users. Some of these activities triggered bans (or ban warnings) from the social media platform. Reasons for these bans could be very different depending on the circumstances (platform policy violation, censorship, targeted troll attacks) but all bans are caused by influencer/user actions.

In the scenarios below we assume that the initial feature engineering has been done, and each distributed site has a set of already precomputed features for its local data set. Each distributed site has observational historical local data for multiple users and features of these users and their activities. Ban labels are also included in the data set. Each site may have different feature sets, and CCaT can help modeling experts discover appropriate features and model parameters (e.g. thresholds for rules in a ban policy) as described in the subsections below.

5.1 Adopting Causal Variables with a Common Effect

Our first example shows how a site may change the thresholds in its existing model when incorporating a new causal relationship from the shared model (and therefore proposed by a different site). This can happen when the new relationship shares a common effect variable with the existing model, under the

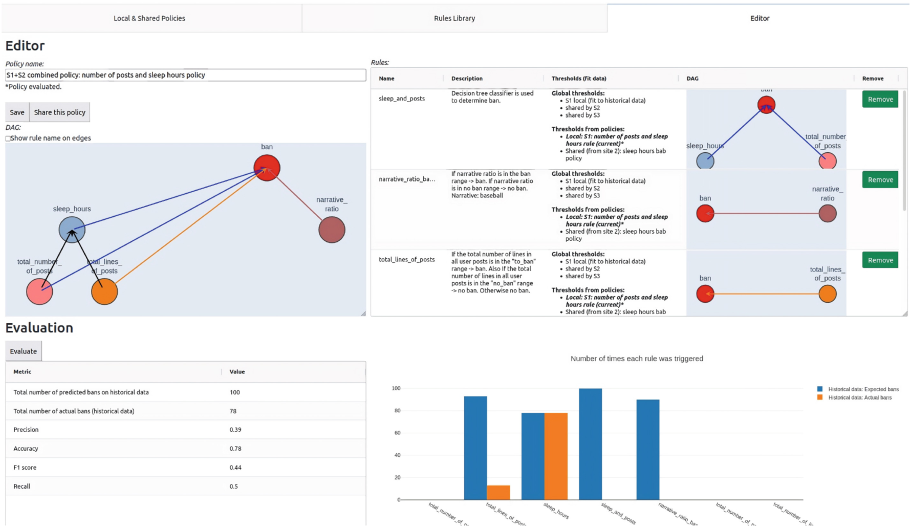


Fig. 3. CCaT DAG editor and policy evaluation UI of the CCaT.

assumption that the site chooses thresholds for the best fit to its dataset both before and after incorporating the new model fragment.

In this case, shown in Fig. 4, site 1 begins with a rule determining if a user will be attacked based solely on the total number of posts made, with a recall of 0.48 and accuracy of 0.57. This is shown in the top of the figure, along with a shared rule that uses only the number of hours the user has been asleep. The lower window shows the site's local model after incorporating the rule based on sleep into its existing model. On fitting to local data, the new model has a recall of 0.84 and accuracy of 0.81.

Since the new rule shared an effect variable, which signifies whether the user will be attacked, this variable now has two incoming links in the model. In general, the rule that best fits the data might combine several thresholds for each variable, for example if it is generated by a decision tree algorithm [9]. Our tool supports this approach but allows the user to use any of a set of approaches to fit a rule to its dataset. One such example is a disjunctive set of single-variable rules that might be found by a rule-learning approach such as Ripper [1]. Even in this case, we note that the thresholds for each variable may be changed from the original thresholds on each individual rule, because each may have over-compensated for cases better handled by the other rule.

5.2 Jointly Discovering Confounding Bias

In our second scenario, adopting a variable from the shared model uncovers a confounding bias that leads a variable to be dropped from the model. In confounding bias, the new variable affects both the outcome of interest and the

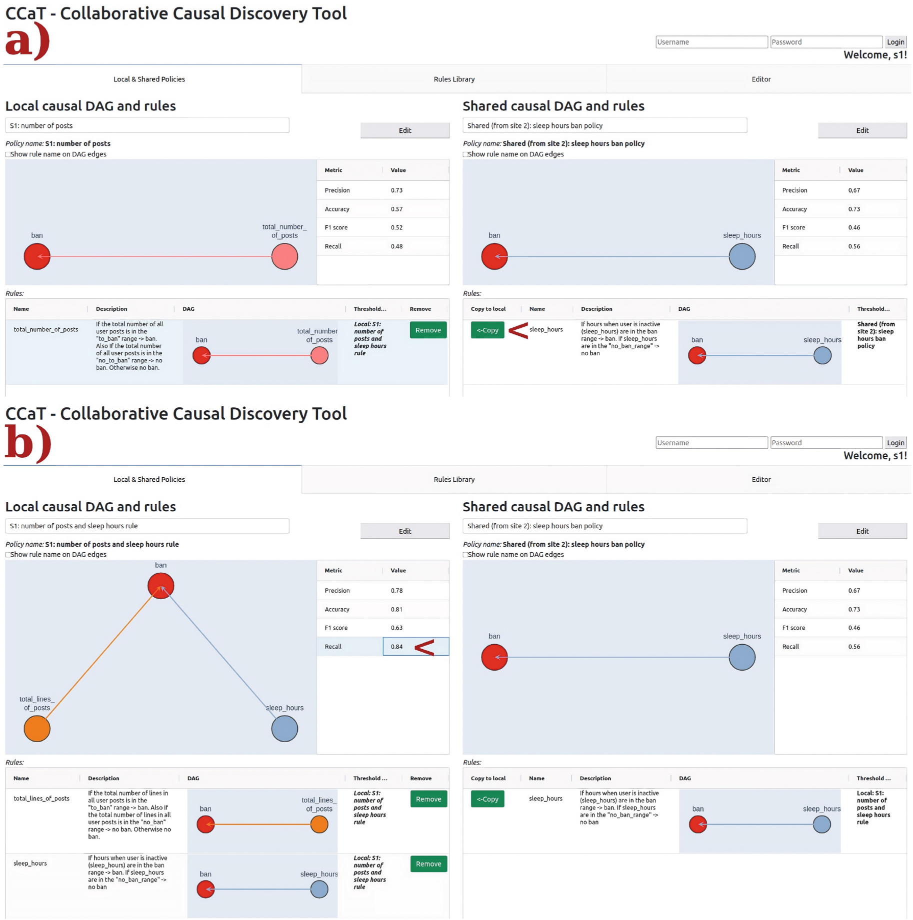


Fig. 4. Scenario 1: a) Site 1 observes that site 2 uses a different rule (the number of sleep hours model on the right) and copies that rule and threshold to local model (left). b) Site 1 uses a combined model (left), leading to improved recall and accuracy.

current treatment variable. One way to mitigate confounding bias is to stratify the model on the confounding variable. In this example, doing so would uncover the fact that a variable in the local site's original model is effectively subsumed by the newly adopted variable, and has little to no effect on the outcome when conditioned on the new variable. Here, however, this situation is more effectively uncovered when the new model is fitted to the local data, and the local site's original, subsumed variable can be safely dropped from the model.

Specifically, the local site has the initial model that being attacked is dependent on the number of posts made in a particular narrative, while the shared model has the (correct) rule that it is caused by the total number of posts in all narratives, as shown in Fig. 5. The site incorporates the shared rule into its local model, then finds a best fit to its local data, increasing its F1 score from 0.47 to 1. The decision tree used to fit the data leans heavily on the total number of posts, as can be seen by the feature importance graph in the lower right corner of Fig. 5. On seeing this, the user at the local site chooses to delete the rule based on the number of posts in a particular narrative from the local model.

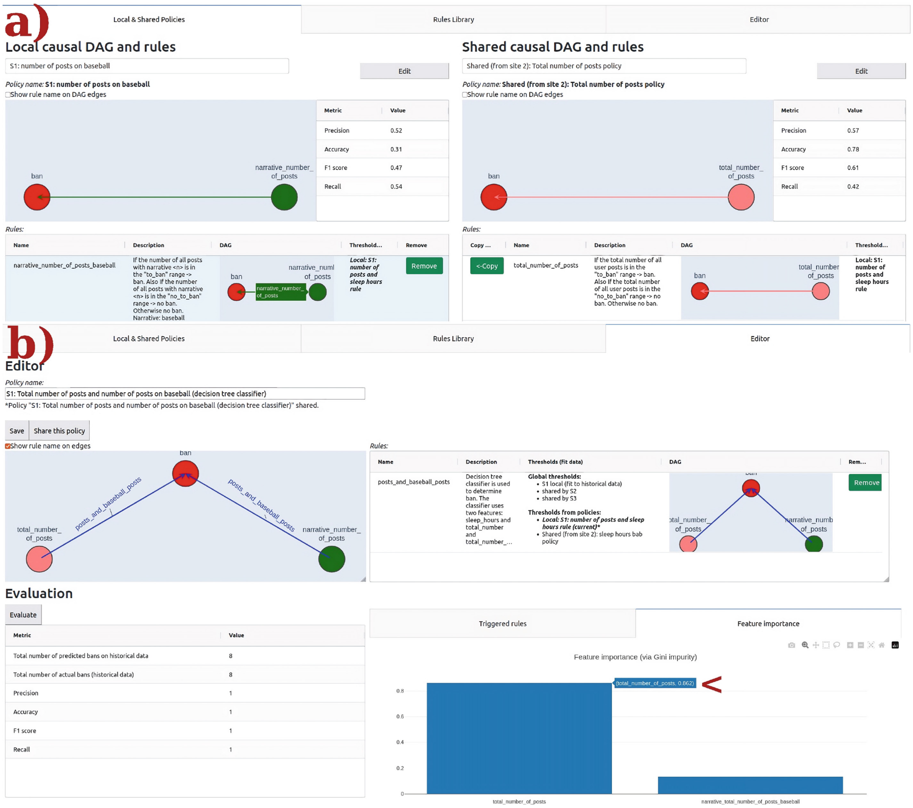


Fig. 5. Scenario 2: a) Site 1 observes that site 2 uses a different rule (the total number of posts, on the right) b) Site 1 uses a combined decision tree model with both features. A correlation analysis would reveal a confounding bias when both variables are included, however an analysis of feature importance shows that the original feature should be dropped from the model.

5.3 Collaborative Support for the Model

The previous scenarios focused on sharing new variables and causal links between sites, and updates made by local sites when fitting the emerging model against their private data. An important and complementary component of building a shared model is aggregating statistics about the shared model's support from each of the component sites. Since the variance of the mean of a set of n independent variables decreases in proportion to $1/n$, each site can combine summary statistics from all the sites to evaluate its model with lower variance than possible using only local data. The approach can also mitigate overfitting that may occur using only local data, and does not share private data from any site. This approach is analogous to sharing weights or gradients on a deep network in federated learning. The complementary component of sharing causal variables and links is typically not done in that setting.

We illustrate the use of sharing statistics from each site within the first scenario above. 10-fold cross-validation may yield a mean and variance for the model accuracy of μ_1 and σ_1^2, while the same method applied by site 2 may yield μ_2 and σ_2^2 respectively. By combining these we can assign a mean and variance to samples drawn from the combined dataset. The mean is the mean of μ_1 and μ_2 while the variance is given by: $\sigma_{1,2}^2 = 1/2(\sigma_1^2 + \sigma_2^2 + 2(\mu_1 - \mu_2)^2)$. Thus, when μ_1 and μ_2 are relatively close, the sample mean of this combined dataset has roughly half the variance $(\sigma_{1,2}^2/20)$ of the sample mean for either site $(\sigma_1^2, 2/10)$.

Improving model estimates via shared summary statistics is most straightfoward when all the sites adopt the shared model and find a good fit with a single shared threshold or decision rule at each link in the model. In this case, meaningful statistics can be shared for the entire model as well as individual links, allowing each site to estimate the global performance of the current model as well as regions where potential improvements according to local data can be shared to update the model. However, CCaT does not enforce this constraint in order to provide support even when individual sites may wish to deviate from the shared model. This may either be due to the local site having different goals from other sites, leading to different tradeoffs on the model, or because the sites have datasets that are not identically distributed due to the nature of the sites. For example, in our social media domain, different users might be working with a different mix of social media platforms, or might be engaging in topics or geographic locations that lead to systematic differences in the data they observe.

6 Discussion

We discussed several scenarios where CCaT addresses some of the challenges of collaborative causal discovery. In situations when sharing data or entire models is impossible, collaboration via partial sharing of models, rules, and summary statistics for models can help discover appropriate features, confounding biases and new causal links.

Our CCaT provides basic functionality for distributed collaborative causal discovery. It addresses some of the limitations of existing tools by providing a web user interface for distributed model development and sharing for human experts with algorithmic support for testing against local data and sharing summary statistics with collaborators. Modeling experts can choose what knowledge to share (parts of the model, individual rules and their parameters, statistics).

The current version of the CCaT was developed as a prototype and has several limitations. It does not support experimental interventions, which is an important tool in causal discovery. We plan to extend our tool to interventions in the future, though we also note it is not appropriate for some domains, for example where they may be too expensive or infeasible.

The user interface is currently limited to sharing small model fragments that use fixed relations between cause and effect (threshold rules, decision tree and disjunctive normal form), and model sharing is automatic between all sites (users cannot choose with whom to share). These limitations will be addressed in future versions of the CCaT with a decentralized peer-to-peer sharing system where pairwise model sharing is possible and users can choose with whom to share their models. We will also test CCaT on larger models from a broader set of domains for scalability and user experience, and incorporate more summary statistics, such as propensity-based methods [11].

The tool can be downloaded for testing by contacting the authors.

Acknowledgements. The authors are grateful to the Defense Advanced Research Projects Agency for their support and to Daryl Best and George Cooper for valuable discussion.

References

1. Cohen, W.W., et al.: Learning rules that classify e-mail. In: AAAI Spring Symposium on Machine Learning in Information Access, vol. 18. Stanford, CA (1996)
2. Imbens, G.W., Rubin, D.B.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge (2015)
3. Kairouz, P., et al.: Advances and open problems in federated learning. Found. Trends Mach. Learn. **14**(1–2) (2021)
4. Kapler, T., Gray, D., Vasquez, H., Wright, W.: Causeworks collaboration: simultaneous causal model construction and analysis. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. CHI EA '21 (2021)
5. Kapler, T., Gray, D.W.S., Vasquez, H., Wright, W.: Causeworks: a mixed initiative framework for causal modeling. SN Comput. Sci. **4**, 1–19 (2022)
6. Mariani, S., Roseti, P., Zambonelli, F.: Towards multi-agent learning of causal networks. In: International Conference on Autonomous Agents and Multiagent Systems (2023)
7. Meganck, S., Maes, S., Manderick, B., Leray, P.: Distributed learning of multi-agent causal models. In: International Conference on Intelligent Agent Technology (2005)
8. Pearl, J.: Causal inference in statistics: an overview. Statistics Surveys (2009)
9. Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE Trans. Syst. Man Cybern. **21**(3), 660–674 (1991)
10. Unni, S.J., Sheth, P., Ding, K., Liu, H., Candan, K.S.: UPREVE: an end-to-end causal discovery benchmarking system. In: SBP-BRiMS Demo (2023)
11. Xiong, R., Koenecke, A., Powell, M., Shen, Z., Vogelstein, J.T., Athey, S.: Federated causal inference in heterogeneous observational data. Stat. Med. **42**(24) (2023)

A Cognitive Slack Approach to Organizational Design

David Mortimore[1,2](✉), Raymond R. Buettner, Jr.[2], and Kathryn J. Aten[2]

[1] Naval Undersea Warfare Center Division, Keyport, Keyport, WA 98345, USA
[2] Naval Postgraduate School, Monterey, CA 93943, USA
{dbmortim1,rrbuettn,kjaten}@nps.edu

Abstract. Organizational performance and goal attainment depend upon robust decision-making and group employment of relevant information and expertise. However, the effects of structural interference can both inhibit and undermine the ability of organizations to locate, access, and employ needed information and cognitive specializations. Importantly, structural interference is a system property: its effects cannot be eliminated, only mitigated. Contending with the effects of structural interference also consumes scarce organizational resources, particularly cognitive slack, negatively impacting performance and stakeholder value generation. Purposeful use of contributor-bias, receiver-bias, point of origin curation (PoOC), and non-human knowledge worker (NHKW) techniques can lessen the cognitive slack lost to structural interference, enabling groups to make more gainful use of relevant information and expertise in choosing how to best accomplish organizational strategies. Computational experiments indicated that receiver-bias, PoOC, and NHKW techniques can effectively return meaningful amounts of cognitive slack, which can then generate more robust group decision-making and performance, ultimately.

Keywords: Attention · Transactive Memory · Cognitive Slack · Decision-Making · Team Performance

1 Introduction

Decision-making characterizes organizational performance and goal attainment: more robust decision-making reasonably generates stronger group performance and greater stakeholder value. Group decision-making reifies an organization's strategy, which expresses how a group intends to accomplish its mission [1–3]. Choosing between options regarding task designs and the shared employment of resources, therefore, predicates organizational performance and goal attainment [4–9]. However, gathering relevant information is costly: identifying the location of and retrieving needed expertise and other information consume scarce resources that could be used for more gainful purposes [4, 6, 10, 11]. Moreover, greater degrees of ambiguity and uncertainty associated with both an organization's external environment and task performance generally necessitate gathering and processing greater volumes of information, increasing the costliness of decision-making [1, 12, 13].

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 110–120, 2024.
https://doi.org/10.1007/978-3-031-72241-7_11

Structural interference, an organizational system property, further increases the costliness of employing relevant information, diminishing performance and goal attainment. Structural interference generally increases the challenges group members encounter in identifying who possesses, employing, and maintaining relevant information [14, 15].[1] Such difficulties inhibit the ability of group members to employ key information in making decisions that, in turn, can result in less optimal outcomes [8, 15]. Simply, structural interference impairs organizational decision-making and performance. Problematically, the nature of structural interference, as a system property, means that its impacts cannot be eliminated, only lessened [12, 15, 17]. Generating stronger group performance, therefore, necessitates organizations purposefully design and incorporate approaches to mitigate the effects of structural interference.

This study investigated the conjecture that techniques reifying four novel constructs—contributor bias, receiver bias, point of origin curation (PoOC), and non-human knowledge workers (NHKWs)—can mitigate the effects of structural interference on organizational performance. This paper describes the relevancy of cognitive slack, a fifth proposed construct, to organizational performance and impacts of notional techniques on mitigating structural interference effects. Then, the research design, including the experiment scenario, hypotheses, and results, are summarized. Finally, study limitations are addressed and recommendations for future work offered. Consistent with organization science studies, in general, this paper situates both the theoretical and empirical work in continuing research trajectories, including seminal studies that remain relevant.

2 Cognitive Slack as Performance Driver

To fully appreciate the relationship between cognitive slack and organizational performance, it is necessary to characterize organizations. Consistent with the modern school of organization theory, organizations are primarily information processing and communication systems that provide members the means to overcome limitations associated with individual performance [12, 18–22]. Among the set of temporal, cognitive, physiological, and institutional limitations on individual performance, the constraints imposed by cognitive performance are particularly impactful because they limit decision-making robustness [4, 5, 11, 18–23].[2] Organizations are effectively decision-making systems [4, 12, 19, 21], and decision-making, itself, is inherently a search for relevant expertise and information [4, 12, 19, 21, 24].

Cognitive slack is essential to robust organizational decision-making and performance and represents the likely performance-limiting resource in groups—the capability to process relevant information. "The scarce [organizational] resource is not information; it is processing capacity to attend to information" [11, p. 270], leading to a need

[1] In the context of transactive memory, *structural interference* is the disruption to the performance of cognitively interdependent systems resulting from organizational network phenomena that increase the costliness of or otherwise discourage organizational members from employing relevant expertise and information [16].

[2] Decision-making *robustness* encapsulates the rationality of decisions [8, 25]. Cognitive limitations and other factors affecting the cognitive performance of individuals effectively bound the rationality of decisions, limiting group performance and goal attainment [8, 20, 22].

to organizationally manage attention for stronger group decision-making and, by extension, performance [5, 10, 11, 26]. Cognitive slack represents the processing capacity described by [11], which is crucial to group decision-making performance, and provides organizational managers and theorists a parameter for managing attention [5, 10, 11].[3] The increasing complexity of decision-making in modern organizations results in processing greater volumes of information [5, 11, 12], amplifying the importance of cognitive slack as an organizational design and management parameter.[4] When inadequate cognitive slack exists, work backlogs result, negatively impacting decision-making and performance [13, 20, 22, 27].

2.1 Structural Interference Reduces Cognitive Slack

Structural interference is particularly pernicious to organizational decision-making and performance. Importantly, structural interference always has a negative impact on organizational performance [14, 15]. Addressing the effects of structural interference generally consumes cognitive slack—wasting an organizational resource that conditions decision-making and, as a result, performance. Structural interference generates impediments to employing relevant expertise, such as cumbersome information retrieval and sharing processes, that necessitate group members use additional amounts of organizationally scarce amounts for task performance [14, 15].

Such impediments effectively prevent group members from employing cognitive slack for more consequential work [14, 15, 18, 20, 22, 28]. The net effect is that the robustness of decision-making generally suffers, leading to less optimal organizational performance and stakeholder value generation [4, 5, 10–12, 15].

2.2 Mitigating Structural Interference Effects

Techniques that reify contributor bias, receiver bias, PoOC, and NHKW constructs—four newly developed constructs, in the context of transactive memory—can mitigate the effects of structural interference on organizational performance. *Contributor bias* describes the intentional design and use of organizational structures, including information sharing and reward mechanisms, to incentivize making relevant expertise and information available to group members [16]. In comparison, *receiver bias* describes the purposeful design and use of organizational structures to encourage group members to retrieve and employ relevant expertise and information. In each construct, the bias is positive: it promotes design and management approaches that encourage making relevant expertise and information available to other group members for use in decision-making and other tasks. *PoOC* describes organizational designs and practices that curate group information stores, automatically, maintaining their relevancy. *NHKWs* are artificial computational agents that perform tasks in the technical core of organizations, alongside human group members [29].

[3] *Cognitive slack* is the volumetric difference between information processing capacity and demands [16].

[4] Notably, cognitive slack is relevant to organizations of human and artificial agents, strengthening its importance in the design and management of modern organizations [18, 20, 22, 23].

Contributor-bias, receiver-bias, PoOC, and NHKW techniques effectively return cognitive slack to organizations, when appropriately applied, allowing use of crucial and limited information processing capacity for more consequential work. Three constructs generally lessen the costliness of employing organizational expertise and information, and the fourth mainly reduces the volume of cognitive workload that human knowledge workers (HKWs) perform. Contributor-bias, receiver-bias, and PoOC techniques usually emphasize the ease of accessing and using information by addressing phenomena that discourage group members from sharing, maintaining, and employing organizationally possessed expertise and information [16, 28]. Unwieldly information sharing mechanisms (ISMs) and inadequate reward processes can dissuade individuals from sharing knowledge, skills, and abilities needed by others [14, 28, 30]. In comparison, NHKWs perform tasks in lieu of HKWs, allowing HKWs to use cognitive slack for more productive work [29, 30]. Employing these constructs to mitigate structural interference impacts on cognitive slack effectively provides organizations the ability to process greater volumes of information, leading to more robust decisions and organizational performance.

3 Research Design

This study employed an organizational engineering modeling application, a community-accepted experiment design, and a generalized call for proposals (CfP) process to explore the conjecture that contributor-bias, receiver-bias, PoOC, and NHKW techniques positively impact group performance.

3.1 Project Organization and Workflow for Edge Research

Project Organization and Workflow for Edge Research (*POWer*) is a computational organization theory (COT) application designed to model group performance of knowledge work, making it particularly appropriate for this study. COT and its tools are grounded in the view of organizations as mainly information processing and communication systems [12, 21, 31, 32], and that decision-making constitutes the primary task performed by organizations [4, 5, 11, 19]. In other words, organizational performance usually represents decision-making performance. COT frameworks and tools facilitate both theory development and testing and enable scholars and practitioners to perform a broader range of experiments, while controlling specific parameters, than laboratory and field experimentation typically allow [31, 32].

POWer is an enhanced version of a validated organizational engineering COT tool [13, 27, 33]. Organizational engineering models use groups as the unit of analysis, instead of individual agents, in detailed representations of the task, coordination, communication, and reporting networks through which organizations reify their strategies to accomplish goals [13, 32, 33]. *POWer* uses discrete event simulation, coupled with both numeric and symbolic reasoning, to stochastically emulate organizational work volumes and impacts on group performance from task uncertainty, project and non-project related communications, cultural norms, and workflows [13, 32, 33]. Moreover, *POWer* is intentionally designed to emulate knowledge work-driven organizations, which typically generate larger volumes of exceptions and, as a result, decisions [13, 27, 33]. Thus, *POWer* provides a robust capability to emulate the effects of contributor-bias, receiver-bias, PoOC, and NHKW techniques on group performance.

3.2 Experiment Design

Strengthening the research design, this study used a COT community-accepted experimentation framework. Summarized below is the design detailed by [33].

- Select the organizational design factors to vary.
- Set other model parameters to average/medium values.
- Using organizational information processing theory (OIPT) [12], predict the directional change in outcomes.
- Run three sets of simulations, with 1,000 trials in each set, for the varied organizational design factors.
- Calculate the mean values of the simulation results.
- Statistically compare experiment results with predicted outcomes.

This research design includes several key assumptions. First, the total amount of organizational work performed is represented by the *total work volume*, which is the aggregate of production work, coordination work, and waiting time volumes [13, 27, 33]. Second, the absence of work backlogs represents adequate cognitive slack, whereas the presence of work backlogs represents inadequate cognitive slack.[5] Third, other factors that can impact organizational performance are modeled by: an individual's experience and skillset; the complexity of tasks and solutions to problems; exception handling at the individual and group levels; and the assigned project and task priorities [13, 27, 33]. A relatively large number of empirical studies used to both develop and test *POWer* and the *Virtual Design Team*, on which *POWer* is based, justify the validity of these assumptions [13, 27, 33]. Table 1 summarizes additional assumptions, which are further described in subsequent sections.

3.3 Scenario Description and Hypotheses

The notional development and dissemination of an annual CfP for an organizational investment program provides a strong scenario for exploring likely impacts of contributor-bias, receiver-bias, PoOC, and NHKW techniques on group performance. This annual task is costly to the organization: it consumes a fairly significant amount of cognitive slack and competes for the attention of group members—that the organization could use for more significant work. Group members must gather, synthesize, and interpret investment inputs from both internal and external sources to draft the CfP,

[5] This study mainly focuses on organizations characterized by greater degrees of uncertainty and ambiguity in both task performance and their operating environments, which usually result in larger exception-handling and decision volumes [1, 11–13]. The performance of such organizations tends to be more sensitive to the availability of cognitive slack.

solicit feedback from stakeholders, make needed revisions, and then obtain organizational approval of the CfP, so it can be disseminated. After the CfP is published, group members continue collecting and reviewing investment inputs for the remainder of the year, in case a significant shift in stakeholder priorities warrants revising the CfP.

However, other organizational priorities compete for the attention of group members, reducing the resources—including cognitive slack—available to publishing the annual CfP. Meetings, communications, and other projects can interrupt developing and circulating the CfP. Such disruptions can also cause group members to wait on others to provide what is needed, generate rework, and increase coordination workload, especially when non-CfP projects have higher priority.

Table 1. Additional model assumptions by technique.

Technique	Contributor bias	Receiver bias	PoOC	NHKW
Design parameters	Transmit call for research proposals using a customer relationship management-like approach	Visualize investment inputs using graph network algorithms (e.g., bibliometrics visualization)	Gather new external investment inputs using information harvesting and natural language processing (NLP) algorithms	Read and synthesize external investment inputs using graph network and NLP algorithms
Dependent variable	Total work volume			
Unit of analysis	Organization			
Significance level	Five percent			

Figure 1 depicts the baseline computational model used in the experiments. An investment program manager (PM) is responsible for soliciting proposals from the workforce for an annual science and technology (S&T) investment program. In addition to overall coordination of the S&T investment program, the PM and their staff work with internal and external stakeholders, S&T Focus Area representatives, and the Executive Director to develop, approve, and publish a CfP, annually. After gathering inputs from internal and external investment sources, the PM and their team read, synthesize, and interpret the aggregated inputs. Normally, the PM drafts an updated CfP and solicits stakeholder feedback, before obtaining the Executive Director's approval. The PM and their staff then send the CfP to the workforce, typically through multiple ISMs, such as electronic mail, intranet sites, and presentations. Once the CfP is published, the PM and their team continue monitoring stakeholder investment inputs for significant changes that might warrant publishing a revised CfP.

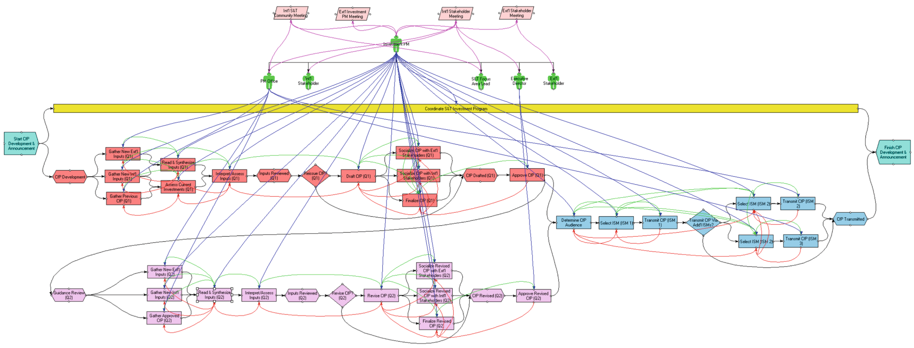


Fig. 1. Baseline S&T investment program CfP computational model.

3.4 Hypotheses

It is hypothesized, in the context of OIPT [12, 33], that:

- contributor-bias, receiver-bias, PoOC, and NHKW techniques will each result in performing a lesser total work volume, compared to the baseline model; and
- an NHKW technique will impact the total work volume the organization performs more significantly than contributor-bias, receiver-bias, and PoOC techniques.

4 Results

Experiment results mostly agree with expectations—three of the four techniques appear to have mitigated the effects of structural interference on cognitive slack and, by extension, organizational performance, and a second technique produced results similar to the NHKW technique. Table 2 compares the mean simulated total work volumes, total work volume deviations, durations, and duration deviations for the CfP development and announcement project [13, 27, 33, 34]. Statistical tests indicated that the differences in mean total work volumes between the baseline model and each of the receiver-bias, PoOC, and NHKW technique models were meaningful, at a five-percent significance level. Confidence interval analyses for the differences in mean total work volumes, again at a five-percent significance level, corroborated these findings. However, statistical tests indicated that there was no meaningful difference between the mean total work volume for the baseline and contributor-bias technique models, at a five-percent significance level. This result was further corroborated by confidence level analysis. Thus, computational results led to rejecting one of the four hypotheses regarding structural interference mitigation techniques.

Surprisingly, the receiver-bias and NHKW techniques generated the same simulated mean total work volumes and total work volume deviations, which led to rejecting the hypothesis that the NHKW technique would outperform other structural interference mitigation techniques. Furthermore, the receiver-bias and NHKW techniques generated equivalent simulated outcomes for the mean duration and mean duration deviation for the CfP development and announcement project (see Table 2). These results suggest that the receiver-bias and NHKW techniques might have completely offset the cognitive slack consumed by the effects of structural interference interpreting/assessing investment inputs and reading and synthesizing investment inputs, respectively.

Table 2. CfP project experiment simulated outcomes (in days).

	Mean total work volume	Mean total work volume deviation	Mean duration	Mean duration deviation
Baseline	641.3	78.9	1,827	238.9
Contributor-bias technique	636.3	81	1,817.3	245.9
Receiver-bias technique	621.9	117.3	1,365.4	213.7
PoOC technique	612.8	79	1,800	236
NHKW technique	621.9	117.3	1,365.4	213.7

5 Discussion

Paradoxically, the mixed results underscore the importance of this and future studies regarding designing techniques to mitigate the effects of structural interference—and how to purposefully incorporate them in engineering organizations. The notion that slack typically contributes positively to organizational performance is fairly well-established [12, 35, 36]. However, it is important to consider in organizational designs that overly zealous efforts to reduce the effects of structural interference might, themselves, generate additional structural interference that consumes more cognitive slack—thereby diminishing group performance—than doing nothing.

Of the limitations associated with this study, three stand out. The first and perhaps most significant limitation is that each experiment considered a single construct: no attempts were made to explore possible interactions between constructs. This conservative approach is in keeping with the community-accepted experiment design used and seems appropriate at this relatively early stage of work [33]. Second, experiment outcomes do not fully describe the amount of cognitive slack available to an organization. *POWer* version 3.4a, used in this study, indicates when organizational information processing capacity is exhausted (i.e., the presence of backlogged work) and the date and maximum amount of backlogged work [34]. Absent is a depiction, or readily attainable calculation, of the volume of information processing capacity remaining when cognitive slack is not exhausted. Lastly, *POWer* generally assumes when individuals perform at 100% capacity when working on project-related tasking, except when group members work longer than the standard workday duration, which can result in greater amounts of rework and other issues [13, 33]. Notwithstanding these limitations, the use of an enhanced version of a community-validated COT tool, specifically designed to emulate organizational behaviors associated with knowledge work, and community-accepted experiment design strengthen confidence in study findings [13, 27, 32, 33].

In modern organizations, it seems increasingly important to treat cognitive slack as an organizational design parameter and purposefully design and incorporate techniques to mitigate the effects of structural interference. Decision-making continues to become increasingly complex; therefore, groups need to gather, process, and communicate even greater volumes of information [4, 10–12, 37], making the organizational management of attention more crucial to performance. Future studies should continue to explore mitigating structural interference effects using contributor-bias, receiver-bias, PoOC, and NHKW techniques—with an eye towards identifying thresholds when the cure could be worse than the disease.

6 Conclusion

Stronger performance of modern organizations likely depends upon treating cognitive slack as a key group-design parameter and using novel contributor-bias, receiver-bias, PoOC, and NHKW techniques to mitigate the effects of structural interference. Computational experiments indicated that intentionally employing already available technical capabilities, such as those in customer relationship management software, can enable groups to use cognitive slack more productively and generate stronger organizational outcomes. In our information-rich world, the largely unmanaged group employment of cognitive slack—which represents the scarce organizational resource—might limit performance and goal attainment more than anything else.

Acknowledgments. The authors thank Dr. Robert (Bob) Iannucci for permitting us to use the expression point of origin curation, in lieu of the original terms used. This study was funded by the Office of Naval Research Cooperative Autonomous Swarm Technology, Cognitive Science and Human & Machine Teaming, and In-House Laboratory Independent Research programs.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Burton, R., Obel, B.: Strategic Organizational Diagnosis and Design: The Dynamics of Fit, 3rd edn. Springer Science Business Media, New York (2004). https://doi.org/10.1007/978-1-4419-9114-0
2. Burton, R., Obel, B., Håkonsson, D.: Organizational Design: A Step-By-Step Approach, 4th edn. Cambridge University Press, Cambridge (2021)
3. Nadler, D., Tushman, M.: Strategic Organization Design: Concepts, Tools, & Processes. Scott, Foresman and Co., Glenview (1988)
4. Arrow, K.: The Limits of Organization. W. W. Norton & Co., New York (1974)
5. Arrow, K.: The economics of information: an exposition. Empirica **23**, 119–128 (1996). https://doi.org/10.1007/BF00925335
6. Cohen, M., March, J., Olsen, J.: A garbage can model of organizational choice. Adm. Sci. Q. **17**(1), 1–25 (1972). https://doi.org/10.2307/2392088
7. Malone, T., Crowston, K.: Toward an interdisciplinary theory of coordination [MIT Sloan Working Paper #3294–91-MSA], pp. 1–48 (1991). http://hdl.handle.net/1721.1/2356

8. Simon, H.: A behavioral model of rational choice. Q. J. Econ. **69**(1), 99–118 (1955). https://doi.org/10.2307/1884852
9. Simon, H.: On the concept of organizational goal. Admin. Sci. Q. **9**(1), 1–22 (1964). https://doi.org/10.2307/2391519
10. Ocasio, W.: Attention to attention. Organ. Sci. **22**(5), 1286–1296 (2011). https://doi.org/10.1287/orsc.1100.0602
11. Simon, H.: Applying information technology to organization design. Public Admin. Rev. **33**(3), 268–278 (1973). https://doi.org/10.2307/974804
12. Galbraith, J.: Organization Design. Addison-Wesley Publishing Co., Menlo Park (1977)
13. Levitt, R.: The virtual design team: designing project organizations as engineers design bridges. J. Organ. Des. **1**(2), 14–41 (2012). https://doi.org/10.7146/jod.6345
14. Hollingshead, A.: Communication, learning, and retrieval in transactive memory systems. J. Exp. Soc. Psychol. **34**(5), 423–442 (1998). https://doi.org/10.1006/jesp.1998.1358
15. Wegner, D., Raymond, P., Erber, R.: Transactive memory in close relationships. J. Personal. Soc. Psychol. **61**(6), 923–929 (1991). https://doi.org/10.1037/0022-3514.61.6.923
16. Mortimore, D., Buettner, Jr., R., Carley, K., Boger, D.: Purposefully mitigating the effects of structural interference on organizational decision-making. Department of Information Sciences, Naval Postgraduate School (2024)
17. Gell-Mann, M.: The Quark and the Jaguar: Adventures in the Simple and Complex. W. H. Freeman & Co., London (1994)
18. Carley, K., Gasser, L.: Computational organization theory. In: Weiss, G. (ed.) Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, pp. 299–330. MIT Press, Cambridge (1999)
19. Cyert, R., March, J.: A Behavioral Theory of the Firm. Cambridge University Press, Cambridge (1963)
20. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus and Giroux, New York (2013)
21. March, J., Simon, H.: Organizations. John Wiley & Sons, New York (1958)
22. Sloman, S., Fernbach, P.: The Knowledge Illusion: Why We Never Think Alone. Riverhead Books, New York (2017)
23. Schroder, H., Driver, M., Streufert, S.: Human Information Processing: Individuals and Groups Functioning in Complex Social Situations. Holt, Rinehart and Winston, New York (1967)
24. Drucker, P.: The rise of the knowledge society. Wilson Q. **17**(2), 52–71 (1993). http://archive.wilsonquarterly.com/sites/default/files/articles/WQ_VOL17_SP_1993_Article_02_1.pdf
25. Edwards, W.: The theory of decision making. Psychol. Bull. **51**(4), 380–417 (1954). https://doi.org/10.1037/h0053870
26. Woolley, A., Gupta, P.: Understanding collective intelligence: investigating the role of collective memory, attention, and reasoning processes. Perspect. Psychol. Sci. **19**(2), 344–354 (2024). https://doi.org/10.1177/17456916231191534
27. Jin, Y., Levitt, R.: The virtual design team: a computational model of project organizations. Comput. Math. Organ. Theory **2**(3), 171–196 (1996). https://doi.org/10.1007/BF00127273
28. Hollingshead, A., Fulk, J., Monge, P.: Fostering intranet knowledge sharing: An integration of transactive memory and public goods approaches. In: Hinds, P., Kiesler, S. (eds.) Distributed Work, pp. 335–355. MIT Press, Cambridge (2002)
29. Mortimore, D.: Moving beyond human-centric organizational designs. J. Organ. Des. **13**(2), 65–75 (2024). https://doi.org/10.1007/s41469-024-00167-z
30. McWilliams, D., Randolph, A.: Transactive memory systems in superteams: the effect of an intelligent assistant in virtual teams. Inf. Technol. People (2024). https://doi.org/10.1108/ITP-12-2022-091
31. Carley, K., Prietula, M.: Computational Organization Theory. Lawrence Erlbaum Assoc., Hillsdale (1994)

32. Frantz, T., Carley, K., Wallace, W.: Computational organization theory. In: Gass, S., Fu, M. (eds.) Encyclopedia of Operations Research and Management Science, LNCS. Springer, New York (2016). https://doi.org/10.1007/978-1-4419-1153-7_143
33. Levitt, R., Cohen, G., Kunz, J., Nass, C., Christiansen, T., Jin Y.: The virtual design team: simulating how organization structure and information processing tools affect team performance. In: Carley, K., Prietula, M. (eds.) Computational Organization Theory, pp. 1–18. Lawrence Erlbaum Assoc., Hillsdale (1994)
34. Collaboratory for Research on Global Projects: Project Organization and Workflow for Edge Research (POWer) (Version 3.4a) [Computer software] (n.d.)
35. Mitroff, I., Mohrman, S.: The slack is gone: how the United States lost its competitive edge in the world economy. Acad. Manag. Exec. **1**(1), 65–70 (1987). https://doi.org/10.5465/ame.1987.4275909
36. Peters, T., Waterman, R.: In Search of Excellence: Lessons from America's Best-Run Companies. Harper & Row, New York (1982)
37. Heavey, C., Simsek, Z.: Transactive memory systems and firm performance: an upper echelons perspective. Organ. Sci. **26**(4), 941–959 (2015). https://doi.org/10.1287/orsc.2015.0979

Data-Driven Approaches

Moderating Democratic Discourse with LLMs

Aaditya Bhatia[1,2] and Gita Sukthankar[2(✉)]

[1] United States Military Academy, West Point, NY 10996, USA
aaaditya.bhatia@westpoint.edu
[2] University of Central Florida, Orlando, FL 32816, USA
{aadityabhati,gita.sukthankar}@ucf.edu

Abstract. Many social media platforms promote political polarization by creating online echo chambers where people are only exposed to information confirming their beliefs. Newer systems such as Polis and Kialo aim to foster constructive conversations and teach critical reasoning skills. However, these platforms rely heavily on human moderators to manage discussions effectively. This paper examines the effectiveness of large language models (LLMs) as moderators on Polis, an open-source, real-time system designed for democratic discourse. We evaluate the F1 score of various prompting techniques at classifying five Polis datasets labeled by human moderators. Our findings indicate that LLMs are robust to different prompting strategies and produce minimal false positives. While LLMs come with certain risks, we argue that they can be valuable tools to support human moderators, enabling broader participation in democratic discourse.

Keywords: Content moderation · LLMs · Computational democracy

1 Introduction

Strong democracies thrive on active citizen engagement in constructive debate. *Collective intelligence*, the combined power of diverse perspectives working together, offers solutions to complex issues [9]. Public participation allows policymakers to tap into the "wisdom of the crowds" for better decision-making [10]. Polis is one of the platforms, developed by the Computational Democracy Project[1] to harness this collective wisdom. This open-source tool facilitates real-time online discussions, gathering and analyzing citizen viewpoints directly.

Public deliberation tools face hurdles in handling the sheer volume of public opinion data. This data can be messy, filled with personal views, and may lack strong evidence. Additionally, these platforms are susceptible to manipulation by those spreading misinformation. Thus, most platforms rely heavily on human moderators to ensure that political discourse remains productive, making it more difficult to scale policy discussions to include a significant percentage of the

[1] https://compdemocracy.org/.

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 123–132, 2024.
https://doi.org/10.1007/978-3-031-72241-7_12

citizenry. Moreover, there is a hidden cost—the potential for mental health issues like PTSD and depression in moderators who are regularly exposed to nefarious content [2].

This paper evaluates the usage of LLMs for content moderation on Polis and examines the role of different prompting strategies on F1 score, false positive rate, and confidence level. Our experiments on five different Polis datasets show that content moderation is fairly robust to different prompting strategies, exhibits a low false positive rate, but is not reliable enough to entirely supplant human moderators. This is unsurprising since moderation on Polis also involves removing redundant ideas from discussions. To enhance Polis's content moderation capabilities, we advocate for a human-machine collaborative approach that integrates large language models (LLMs) into the moderation pipeline.

2 Related Work

2.1 Deliberative Democracy

constitutes discourse that focuses on evidence and reasoning, which encourages participants to reflect on various perspectives and form an informed opinion. It assumes that through rational discourse, participants can arrive at decisions that are more legitimate and informed [18]. According to Fishkin [4], the practical realization of deliberative democracy faces several challenges. While effective deliberation can be facilitated in small groups, scaling this to a larger population becomes difficult. Maintaining inclusivity and equality in discussions is especially critical to ensure that all voices are heard and considered. Most importantly, the moderation efforts needed to structure and manage complex discussions are essential to keep them focused and productive, and this time commitment grows significantly with the number of participants. Hadfi et al. [8], proposed the usage of conversational agents to promote constructive discussion in democratic forums; however their agent was proactively driving discussion towards consensus, rather than performing content moderation.

Technological Platforms. Rather than using existing social media platforms, Klein [10–12] championed the creation of technological platforms that facilitate the aggregation, organization, and analysis of collective inputs, ensuring that the deliberative process is efficient, scalable, and inclusive. The Polis platform was specifically designed to host public deliberation and has been used by several organizations and governments in Austria, Taiwan, New Zealand, and the US [16]. Discourse on Polis is structured around conversations, created by an owner who sets the discussion topic. Participants submit comments on the topic, which are then routed to other participants to vote on. The current version of Polis relies more heavily on voting rather than natural language processing and uses matrix factorization to understand group opinions based on participants' votes on comments. The authors of the Polis platform themselves highlighted some of the ways that LLMs could promote scalable deliberation and also potential risks related to LLMs [17]. Their report was cautiously optimistic about the usage of LLMs on Polis and identified summarization, topic modeling, reporting, vote prediction, and content moderation as potential applications. However, they remained concerned about the problem of hidden biases inherited from training data.

Algorithmic Content Moderation. Algorithmic content moderation includes "systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome." [5] Grimmelman presented a taxonomy of moderation practices for building online communities, including exclusion, pricing, organization, and norm-setting [6]. Our research is an example of moderation through *organization*, shaping the information flow between content producers and consumers. Much of the research in this area has focused on the detection of hate speech [14], removal of copyrighted content [20], and age-appropriate content moderation [1]. However, Polis does not contain the same type of inappropriate content found on platforms like Twitter and YouTube, since there is no pathway for content monetization. Instead, it is more similar to content moderation on Reddit, where the aim is to remove material that does not adhere to specific guidelines determined by the moderator. However, some of the Polis content moderation guidelines are mainly meant to facilitate voting on policy issues, which is not a consideration for most platforms.

3 Method

3.1 Comment Moderation

For our set of experiments on comment moderation, we use the Polis moderation guidelines to design prompts for various language models.[2] This method aims to identify and label irrelevant and overly complex statements. Specifically, each statement undergoes individual analysis by the language model for classification purposes. The efficacy of this strategy is evaluated by comparing the outcomes of the language model's spam detection against a gold standard-moderation data previously labeled by the moderators of each Polis study dataset. This comparison seeks to ascertain the spam detection accuracy across various language models.

To manage text generation from our language models, we use the *guidance* framework originally developed by Microsoft. This library represents a unique programming paradigm that enhances control and efficiency for a language model by constraining generation through regular expressions and context-free grammar. Developers can freely add text to the context window at any point between text generations, effectively interleaving control and generation seamlessly using traditional programming paradigms such as conditionals and loops.

The Polis project has proposed the following moderation guidelines. Each organization conducting a user study ultimately decides its policy for moderation but generally follows these guidelines.

[2] Our code is available at: https://github.com/aadityabhatia/polis-argmap.

– Spam: Comments devoid of relevance to the discussion.
– Duplicative: Comments restating a previously made point.
– Complex: Comments articulating multiple ideas or problems.

Our experiment considers the effect of several variables on moderation outcomes, including:

1. Class labels used for classification by the language model, varying across experiments between a simple set (ACCEPT, UNSURE, REJECT) and a more detailed set (ACCEPT, UNSURE, SPAM, IRRELEVANT, UNPROFESSIONAL, SCOPE, COMPLEX).
2. A set of examples provided to the language model to help it better understand the task.
3. Chain-of-thought reasoning techniques that ask the model to state a pre-decision "thought" statement and potentially enhance the model's reasoning abilities, as demonstrated by the ReAct strategy [19].
4. Application of additional reasoning techniques, like statement deconstruction, to dissect and understand statements in depth.

We took steps to ensure that the model is not too strict in its moderation, since there is more harm done from falsely rejecting a statement than from falsely accepting it. A statement that is allowed into discussion will still be voted upon and moderated by the community. For example, a statement that adds no value to the discussion is likely to be treated as noise and see a strong disagreement. However, a statement that is incorrectly rejected is less likely to be seen again. Moderating legitimate statements discourages community participation and damages the trust people place in the system. This is also a place where inherent biases in language models can surface. Therefore, we prefer false positives to false negatives.

We structured our methodological approach to leverage the capabilities of transformer-based language models, specifically focusing on their application in comment moderation tasks. Using the *transformers* library and specifically chat-tuned versions of Mixtral model, we defined the task using instructional prompts and user statements for the model to respond to in a chat paradigm. We carefully crafted our inputs to align with our experimental objectives by providing context, task and expected output format. The use of the *guidance* framework allowed us to retain control over most aspects of the model's output, ensuring adherence to our structured generation format.

To uncover the latent structures within the discussion statements, we employed the BERTopic library [7], leveraging an advanced pipeline based on transformer embeddings and class-based Term Frequency-Inverse Document Frequency (TF-IDF) statistics. This process starts with generating statements embeddings, reducing their dimensionality using Uniform Manifold Approximation and Projection (UMAP) [15], and identification of clusters using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). The resulting clusters were then analyzed for their most representative keywords and assigned descriptive labels for easy identification and interpretation using an LLM. Figure 2 shows an example topic distribution projected on a 2D plane using PCA (Fig. 1).

user	List the 8 most important areas of improvements from these statements, each on a new line.
assistant	- Cooperation between WKU and Bowling Green for community involvement and internships - Accessible mentorship programs for young adults and teenagers - Community enrichment classes and activities, including woodworking, welding, gardening, and home beautification - More after-school activity centers for youth with extended hours of operation - Increased number of concerts held in Bowling Green - Better incorporation of university resources in community, including Kentucky Museum - Public adult swim facilities for year-round physical use - More activities and places to go for teenagers
user	AREA OF IMPROVEMENT: Cooperation between WKU and Bowling Green for community involvement and internships
assistant	PROBLEMS IDENTIFIED: Lack of cooperation and partnership between WKU and Bowling Green, limited opportunities for community involvement and internships for students ACTIONABLE SOLUTIONS: Strengthening mutually beneficial partnerships between WKU and Bowling Green, increasing cooperation for community involvement and internships ARGUMENT: Strengthening WKU-Bowling Green partnerships will create more community involvement and internship opportunities, benefiting both students and the local community. ARGUMENT LABEL: Boosting student opportunities and community involvement
user	AREA OF IMPROVEMENT: Community enrichment classes and activities, including woodworking, welding, gardening, and home beautification
assistant	PROBLEMS IDENTIFIED: Limited community enrichment classes and activities in Bowling Green, particularly in woodworking, welding, gardening, and home beautification ACTIONABLE SOLUTIONS: Establishing and promoting community enrichment classes and activities that cater to various interests and hobbies ARGUMENT: Introducing diverse community enrichment classes and activities will cater to various interests, enhancing the overall quality of life in Bowling Green. ARGUMENT LABEL: Diversifying community enrichment
user	AREA OF IMPROVEMENT: More after-school activity centers for youth with extended hours of operation
assistant	PROBLEMS IDENTIFIED: Insufficient after-school activity centers for youth with limited hours of operation in Bowling Green ACTIONABLE SOLUTIONS: Developing and expanding after-school activity centers with extended hours of operation ARGUMENT: Providing more after-school activity centers with extended hours will offer youth safe, engaging, and educational spaces beyond traditional school hours. ARGUMENT LABEL: Expanding youth opportunities

Fig. 1. The set of instructions and prompts used to generate insights using *Guidance* framework. Text highlighted in green was produced by LLM, while everything else was programmatically inserted into the LLM context. Guidance enables developers to intertwine text and generation, allowing precise control of language model contexts.

Datasets. Our research focuses on two sets of events from the Polis dataset: *american-assembly.bowling-green* and *scoop-hivemind*. The first one originates from a conversation facilitated by the American Assembly in Bowling Green, Kentucky. This dataset explores local and regional perspectives regarding the city's well-being and probes the narrative of a divided America. It comprises 896 comments, of which 289 were moderated out, leaving a substantial 607 comments for analysis. This dataset is particularly valuable for understanding community priorities and perceptions at a local level.

The second set, *scoop-hivemind*, stems from multiple conversations conducted by New Zealand's Public Engagement Projects (PEP) in partnership with the news outlet Scoop regarding issues of national significance. It consists of a total of 752 comments submitted by 96 people, with 294 comments moderated out, resulting in 458 accepted comments. Within this group, the *biodiversity* and *freshwater* datasets provide insights into protecting and restoring New Zealand's biodiversity and preserving freshwater resources, an area of global environmental concern that drives significant policy decisions, while *taxes* and *affordable-housing* address the socio-economic challenges faced by the public. These datasets include detailed voting data accounting for each *agree* or *disagree* vote cast by the participants.

Fig. 2. Distribution of statements in the *american-assembly.bowling-green* dataset colored by topic projected in a 2D plane using PCA.

4 Experiments

Table 1 shows eight experimental configurations tested for comment moderation. For the baseline, we used a simple instructional prompt to categorize comments into three labels: ACCEPT, UNSURE, or REJECT. When the model rejected a statement, we further classified it as either SPAM or COMPLEX. Subsequently, the configurations were expanded to a more complex seven-class system, providing detailed instructions for each category to enhance the model's decision-making precision and explainability. The *semantic extraction* technique focused on the content of each comment to identify the problem being addressed, the proposed solution and the number of ideas introduced. We require each comment to mention at least one problem or solution and no more than one unique idea. Combining it with chain-of-thought reasoning, this approach deviated from standard moderation guidelines and instead judged comments based on their relevance to the ongoing conversation. *Chain-of-thought* reasoning allowed the model to articulate its thought before each decision, aiming for a higher moderation accuracy and transparency. We made extracted ideas and thought statements available as a part of our results for increased explainability.

Evaluation Metrics. To evaluate these experiments, we selected datasets with high-quality statements that align closely with Polis moderation guidelines. Raw statements were used with no pre-processing, mirroring real-world conditions where moderators make quick decisions without access to the full dataset. Effectiveness was measured using the F1 score, false positive rate, and the rate at which the model selected *UNSURE* over *ACCEPT* or *REJECT*. The F1 score is a balanced measure that considers both the precision and recall of the classification process, particularly useful when the costs of false positives and false

Table 1. Summary of experimental configurations for comment moderation. **E** = semantic extraction and **T** = chain of thought reasoning.

Configuration	Target Classes	Chain-of-Thought	Semantic Extraction
3	3	No	No
7	7	No	No
3T	3	Yes	No
7T	7	Yes	No
3E	3	No	Yes
7E	7	No	Yes
3ET	3	Yes	Yes
7ET	7	Yes	Yes

negatives differ significantly. It is especially useful when dealing with imbalanced datasets where positive cases, which in our case are the comments to be rejected, are significantly less common than the negative ones. It is calculated using the formula

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Throughout these experiments, we carefully considered the implications of false positives on the moderation outcome, with a particular emphasis on minimizing false positives to foster inclusive community discussions.

5 Results

Table 2 shows the F1 score, false positive rate and number of comments classified by the LLM as uncertain across the five Polis datasets. The weighted average shows the average across all datasets, normalized by the size of the dataset. Our results show that the simple three-class baseline in which the model is simply asked to classify a statement as ACCEPT, REJECT, or UNSURE outperforms the other prompting strategies. The more sophisticated prompting strategies, chain of thought (T) and semantic extraction (E) did not consistently improve the F1 score over the baseline. The overall F1 score was not sufficiently high to make the LLM alone a convincingly good replacement for a human Polis moderator.

A second question is whether an LLM can work well in tandem with a human moderator. All the model variants exhibit low false positive rates, making them well suited for this application. Both chain of thought and semantic extraction tend to make the model more unsure, which is less problematic when the LLM is not expected to be the final arbiter.

The LLM struggles with duplicate detection, which is more important for Polis since the moderators don't want to keep issuing redundant votes. Detecting duplicates requires a large context window and increases the memory footprint linearly with the number of comments, often causing the model to run out of GPU memory or overflow past its maximum context window length. A more effective approach would involve clustering comments using their text embeddings and detecting semantically identical statements, which is a promising technique for implementing a moderation system.

Kolla et al. recently published a paper on the usage of GPT-3.5 for moderating Reddit content [13]. Although they employed different prompting strategies, they reported similar performance trends in terms of false negative and true positive rates. One issue that they noted is that the LLM cannot be easily queried about its confidence level and sometimes reversed its decision upon further queries. Hence, we believe that our simpler strategy of including UNSURE as a possible class produces superior results.

Table 2. Comment moderation results averaged over all datasets. Weighted average normalizes the results by the size of the dataset.

Metric	Dataset	Configuration							
		3	7	3T	7T	3E	7E	3ET	7ET
F1↑	american-assembly.bowling-green	0.26	0.08	0.21	0.19	0.15	0.11	0.18	0.12
	scoop-hivemind.biodiversity	0.33	0.09	0.25	0.21	0.20	0.22	0.25	0.13
	scoop-hivemind.freshwater	0.26	0.07	0.16	0.29	0.16	0.06	0.18	0.12
	scoop-hivemind.taxes	0.10	0.04	0.06	0.07	0.10	0.19	0.26	0.17
	scoop-hivemind.affordable-housing	0.11	0.12	0.15	0.12	0.10	0.12	0.24	0.20
	Weighted Average	**0.24**	0.08	0.19	0.18	0.15	0.14	0.20	0.13
FPR↓	american-assembly.bowling-green	0.06	0.02	0.04	0.05	0.02	0.01	0.02	0.02
	scoop-hivemind.biodiversity	0.05	0.03	0.07	0.04	0.06	0.06	0.07	0.03
	scoop-hivemind.freshwater	0.10	0.04	0.20	0.02	0.16	0.13	0.27	0.09
	scoop-hivemind.taxes	0.02	0.01	0.16	0.04	0.12	0.04	0.10	0.05
	scoop-hivemind.affordable-housing	0.06	0.02	0.15	0.04	0.10	0.01	0.12	0.06
	Weighted Average	0.06	**0.02**	0.08	0.04	0.05	0.03	0.06	0.03
Unsure	american-assembly.bowling-green	0.01	0.05	0.05	0.04	0.02	0.02	0.02	0.03
	scoop-hivemind.biodiversity	0.02	0.08	0.07	0.07	0.04	0.04	0.06	0.11
	scoop-hivemind.freshwater	0.04	0.07	0.12	0.12	0.10	0.10	0.04	0.12
	scoop-hivemind.taxes	0.07	0.11	0.03	0.11	0.09	0.09	0.09	0.13
	scoop-hivemind.affordable-housing	0.04	0.05	0.05	0.09	0.02	0.08	0.01	0.10
	Weighted Average	**0.02**	0.06	0.06	0.06	0.03	0.04	0.03	0.07

6 Conclusion

In this study, we assessed how well large language models (LLMs) perform as content moderators for Polis, a platform designed to facilitate open and democratic discussions. We tested different prompting techniques, such as those that focus on extracting meaning (semantic extraction) and revealing the reasoning process (chain-of-thought reasoning). Our findings indicate that for the Mixtral model, prompts focused on simple classification yielded the best results.

The use of artificial intelligence (AI) tools in public debates and policymaking has the potential to significantly change how we understand and address social issues. LLMs can lighten the load on human moderators by automating some content review tasks, paving the way for democratic discourse to truly flourish at scale. Our research suggests that LLMs are best suited to work alongside human moderators as part of a larger moderation system. This is because they have a low rate of incorrectly flagging comments, making it less likely that valid content will be removed unnecessarily.

Limitations There is a growing emphasis on addressing algorithmic bias and ethical considerations in these methodologies. As LLMs are trained on extensive datasets, there is a risk of inheriting biases present within the data. Contemporary methods are frequently focused on mitigating these biases to ensure that the insights generated for policymaking are equitable and representative of diverse viewpoints [8]. Another critical aspect is the scalability and computational efficiency of these approaches, particularly vital when addressing global-scale issues with large number of participants. The application of LLMs and AI-driven tools needs to be weighed against computational costs and the practicality of implementing these solutions on a large scale [3].

Acknowledgements. This material is based upon work partially supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. W911NF-20-1-0008. The conclusions and opinions expressed in this research paper are those of the authors and do not necessarily reflect the official policy or position of the U.S. Government or Department of Defense.

References

1. Ahmed, S.H., Khan, M.J., Sukthankar, G.: Enhanced multimodal content moderation of children's videos using audiovisual fusion. In: The International FLAIRS Conference Proceedings, vol. 37, no. 1 (2024)
2. Arsht, A., Etcovitch, E.: The human cost of online content moderation. Harv. J. Law Technol. (2018)
3. Ding, S., Ito, T.: Self-Agreement: A Framework for Fine-tuning Language Models to Find Agreement among Diverse Opinions (May 2023). https://doi.org/10.48550/arXiv.2305.11460, arXiv:2305.11460 [cs]
4. Fishkin, J.S.: When the People Speak: Deliberative Democracy and Public Consultation. Oxford University Press, Oxford, August 2011. https://doi.org/10.1093/acprof:osobl/9780199604432.001.0001, https://academic.oup.com/book/12596

5. Gorwa, R., Binns, R., Katzenbach, C.: Algorithmic content moderation: technical and political challenges in the automation of platform governance. Big Data Soc. **7**(1), 2053951719897945 (2020)
6. Grimmelmann, J.: The virtues of moderation. Yale JL Tech. **17**, 42 (2015)
7. Grootendorst, M.: BERTopic: neural topic modeling with a class-based TF-IDF procedure, March 2022. arXiv:2203.05794 [cs]
8. Hadfi, R., Ito, T.: Augmented democratic deliberation: can conversational agents boost deliberation in social media? In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, pp. 1794–1798. AAMAS '22, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, May 2022
9. Kirschner, P.A., Buckingham Shum, S.J., Carr, C.S., Diaper, D., Sanger, C. (eds.): Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making. Computer Supported Cooperative Work, Springer, London (2003). https://doi.org/10.1007/978-1-4471-0037-9
10. Klein, M.: How to Harvest Collective Wisdom on Complex Problems: An Introduction to the MIT Deliberatorium (2012). https://www.semanticscholar.org/paper/How-to-Harvest-Collective-Wisdom-on-Complex-An-to-Klein/79541bad78d11f448f94c2d3fd83d304dcf2f733
11. Klein, M.: Crowd-Scale Deliberation For Complex Problems: A Progress Report, March 2022. https://doi.org/10.2139/ssrn.4049680, https://papers.ssrn.com/abstract=4049680
12. Klein, M., Iandoli, L.: Supporting Collaborative Deliberation Using a Large-Scale Argumentation System: The MIT Collaboratorium, February 2008. https://doi.org/10.2139/ssrn.1099082, https://papers.ssrn.com/abstract=1099082
13. Kolla, M., Salunkhe, S., Chandrasekharan, E., Saha, K.: LLM-Mod: can large language models assist content moderation? In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1–8 (2024)
14. MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: challenges and solutions. PLoS ONE **14**(8), e0221152 (2019)
15. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. https://doi.org/10.48550/arXiv.1802.03426, arXiv:1802.03426 [cs, stat]
16. Small, C.: Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. RECERCA. Revista de Pensament i AnÃ lisi, July 2021. https://doi.org/10.6035/recerca.5516, https://www.e-revistes.uji.es/index.php/recerca/article/view/5516
17. Small, C.T., et al.: Opportunities and Risks of LLMs for Scalable Deliberation with Polis, June 2023. arXiv:2306.11932 [cs]
18. Williams, A.E.: Are wicked problems a lack of general collective intelligence? AI Soc. **38**(1), 343–348 (2023). https://doi.org/10.1007/s00146-021-01297-8
19. Yao, S., et al.: ReAct: synergizing reasoning and acting in language models, March 2023. arXiv:2210.03629 [cs]
20. Zhang, D.Y., Badilla, J., Tong, H., Wang, D.: An end-to-end scalable copyright detection system for online video sharing platforms. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 626–629 (2018)

Deciphering Emotional and Linguistic Patterns in Reddit Suicidal Discourse

Salim Sazzed[1,2(✉)]

[1] Georgia Southern University, Statesboro, GA, USA
ssazzed@georgiasouthern.edu
[2] University of Memphis, Memphis, TN, USA

Abstract. Recognizing and understanding the characteristics of suicidal posts on social media are crucial for early intervention and suicide prevention efforts. This study aims to comprehend various traits associated with suicidal posts on Reddit, a prominent social media platform. By analyzing user posts gathered from the Reddit forum 'r/SuicideWatch', we aim to scrutinize the various emotional and linguistic signals intertwined with suicidal content. Furthermore, we seek to identify some of the key factors contributing to the development of suicidal ideation. Our analysis reveals a substantially higher prevalence of negative emotions, particularly anger and sadness, in suicidal posts. Linguistic analysis uncovers several distinguishing signals associated with suicidal posts, such as a high presence of verbs and a significant frequency of negations. An initial content analysis reveals a diverse array of key factors, including social isolation, health issues, personal failures, and traumatic life events, among others. Overall, our investigation unveils diverse characteristics of suicidal posts and key factors contributing to suicidal ideation that can facilitate the automated detection of suicidal tendencies and support the implementation of effective preventive measures.

1 Introduction

Suicide is a significant public health concern, ranking as one of the leading causes of death worldwide[1]. The World Health Organization (WHO) reports that more than 700,000 individuals succumb to suicide annually[2]. Mental illness is highly associated with suicidal tendencies; therefore, early detection of mental health issues is crucial for preventing suicidal feelings. Traditional procedures for detection include clinical interviews and psychological evaluations using questionnaires [5]. While effective, these approaches are time- and resource-intensive involving direct encounters between parties, and may not always be feasible on a large scale.

The growing popularity of social media platforms such as Twitter, Reddit, and Facebook has made them an alternative sources of data for mental health-related research [11–13]. Users often share their emotions, thoughts, and opinions

[1] https://www.who.int/news-room/fact-sheets/detail/suicide.
[2] https://www.who.int/health-topics/suicide#tab=tab_1.

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 133–143, 2024.
https://doi.org/10.1007/978-3-031-72241-7_13

on these platforms, providing insight into their moods, communication patterns, activities, and social interactions. The tone and language used in social media posts can even indicate the presence of mental health issues, such as depression, feelings of worthlessness, guilt, helplessness, and self-hatred. Previous studies have demonstrated that individuals with mental illness may use social media as a means of coping, revealing their struggles through their posts [6,17].

Over the recent years, a number of studies have been conducted to comprehend social media suicidal posts. However, they primarily focused on classification task using word n-grams and other related features. For instance, Ramirez et al. [10] investigated word n-gram features, behavioral features including sleep time tweets ratio (STTR) and the daytime tweets ratio (DTTR), and image features in Twitter data to create machine learning models for identifying users who may be at risk. Tadess et al. [16] employed an Long Short-Term Memory(LSTM)-Convolutional Neural Network (CNN) combined model with word embedding (based on word n-grams) and compared its performance with various classification models. Sinha et al. [15] developed a sequential neural network model that takes into account the temporal aspect of users' historical tweeting activity and its time-variant impact on final predictions made by a classifier. Meanwhile, Lao et al. [7] analyzed various linguistic features in suicidal posts, focusing specifically on readability scores and word-level emotion using LIWC. Existing studies aiming to identify key factors linked to suicidal tendencies often utilize surveys and questionnaires administered within clinical or health settings. While useful, this method may limit individuals' ability to express themselves fully due to the structured nature of question-and-answer formats and the requirement for identifiable information. In contrast, social media's anonymity allows for greater freedom of expression.

To complement the existing research, this study focuses on dissecting suicidal posts from various perspectives, including emotional, linguistic, and semantic viewpoints, to identify attributes and causes related to suicidal tendencies. The suicidal texts are collected from a Reddit discussion forum *r/SuicideWatch*. As the name suggests, this forum contains postings by individuals who are considering taking their own lives [14]. We investigate various emotion-related attributes, such as the presence of emotions at both word and document levels. Additionally, we perform linguistic analysis on these posts to identify multiple types of syntactic and grammatical attributes, which are then compared with non-suicidal posts. The results reveal that fear and sadness-related emotion words are more prevalent in the posts. Moreover, we observe the dominance of anger emotion at the post level. The linguistic analysis suggests that some attributes show a higher presence in suicidal texts than in non-suicidal posts, which may be linked to suicidal tendencies. Finally, a limited content analysis reveals some of the key determinants of suicidal ideation, including social isolation, traumatic life events, relationship issues, and health concerns, among others.

2 Dataset

The suicidal corpus we analyze here represents posts from a Reddit forum dedicated to suicidal feelings, */r/SuicideWatch*. Reddit is a popular social media platform widely used by individuals to discuss various topics. Due to the anonymity it provides, users often share posts about stigmatized topics [16]. Furthermore, the substantial length of Reddit posts makes them a valuable resource for mental health research, enabling exploration of psychological, linguistic, and other textual features.

Both the suicidal and non-suicidal corpora are compiled from multiple Reddit datasets publicly available on Kaggle[3], a web platform for data scientists and machine learning researchers. To create the suicidal corpus, we utilize two datasets: one containing approximately 10,000 suicidal posts and the other containing 6,600 suicidal posts. After excluding posts with less than 200 characters, the final dataset comprises 10,590 suicidal posts. It's worth noting that very short posts (less than 200 characters) were excluded from our analysis due to their inherent lack of substantial linguistic, affective and semantic information. Additionally, a significant portion of these short posts contains errors, such as the presence of irrelevant words and extraneous characters, among other issues. To construct the non-suicidal corpus, we consider Reddit posts collected from various sub-forums such as aww, jokes, sex, and funny. We apply the same inclusion criteria, resulting in a set of 11,308 non-suicidal posts.

Table 1. Length related statistics in the suicidal and non-suicidal posts

Review Length	Suicidal	Non-suicidal
	Median/Mean/Std.	Median/Mean/Std.
#Words	137/204.53/221.12	87/193.80/324.88
#Sentence	9.0/ 12.81/14.04	6/12.75/22.15
#Words per Sentence	15.5/21.31/29.47	16.67/18.73/12.51

The length statistics (Table 1) reveal variability in the length of suicidal posts at both the word and sentence levels. The mean post length at the word level is approximately 204 words, with a notable standard deviation (std.) indicating significant variability. A similar pattern is observed at the sentence level. We observe that these statistics also vary in non-suicidal posts.

3 Emotion Analysis

We incorporate emotion analysis to extract and comprehend the emotional content present in both suicidal and non-suicidal corpora. This enables us to gain insights into the emotional states of individuals and identify trends and patterns in emotional expression.

3.1 Presence of Emotion Words

To evaluate emotional word usage in both corpora, we employ the NRC Emotion Lexicon [9]. This lexicon categorizes English words into eight emotions- *anger,*

[3] https://www.kaggle.com.

fear, anticipation, trust, surprise, sadness, joy, and *disgust*-determined through crowdsourcing. Analyzing text with this lexicon facilitates the identification of emotional content, offering insights into the writer's emotions and overall sentiment. Given the prevalence of negative emotions in suicidal posts, our analysis focuses on anger, fear, sadness, and disgust. We calculate how frequently each emotional category appears at the word-level within each post. Afterward, we provide statistical summaries including the mean, median, and standard deviation (std.).

Table 2. The presence of four types of negative emotion words in both suicidal and non-suicidal posts, represented as percentages.

Emotion type	Suicidal	Non-suicidal
	Median/Mean/Std.	Median/Mean/Std.
Anger	2.18/2.50/1.81	0.84/ 1.25/1.59
Fear	2.90/3.28/2.22	1.21/1.60/1.84
Sadness	3.25/3.60/2.25	1.08/1.45/1.68
Disgust	1.64/1.92/1.63	0.52/0.98/1.37

Table 2 illustrates the presence of four types of emotion words, *anger, fear, sadness* and *disgust* in the suicidal and non-suicidal posts. We find almost twice the presence of negative emotional words in suicidal posts for all of the negative emotions considered, based on both mean and median values. The most dominant emotion in the suicidal group is sadness, which accounts for around 3.25% when the median value is considered. The second dominating emotion words are related to fear, around 3.3%. For the non-suicidal group, the most emotional words are from the fear group, but still, it is less than half of the suicidal group considering both mean and median.

3.2 Document-Level Emotion Recognition

In addition to finding various types of emotion words, we explore the dominant emotion present at the document level (i.e., for entire post) for the suicidal posts. We use an emotion recognition framework EmoNet, [1] that can recognize the following eight primary types of emotions: *joy, anticipation, surprise, trust, anger, disgust, fear, and sadness* in a text. Since suicidal posts are likely to contain primarily negative emotions, here, we consider the following four types of negative emotions- *anger, disgust, fear, and sadness.* Since in document-level, user post can contain multiple types of emotions at various proportions, we look for the most dominant emotion in the text. We use a threshold of 0.5 to decide most dominant emotions.

Table 3. Frequency and percentages of dominant emotions at the document-level in suicidal and non-suicidal posts

Emotion	Suicidal	Non-suicidal
Anger	2276 (21.49%)	1000 (8.87%)
Disgust	139 (1.31%)	312 (2.77%)
Fear	258 (2.43%)	236 (2.09%)
Sadness	206 (1.95%)	192 (1.70%)

When the dominant emotions are considered for the entire post using the emotion recognition framework (Table 3), we find a significant dominance of the anger class. The *anger* emotion accounts for 22.50%, while the second dominant emotion *sadness* accounts for 2.43%. We notice the following orders of dominance in the suicidal group, $anger > fear > sadness > disgust$. For the non-suicidal group, although, the anger is dominant category, we find it is much less, around 8.9%. Besides, we notice a different order for the non-suicidal group, $anger > disgust > fear > sadness$. Overall, the significantly high dominance of anger suggests its close association with suicidal tendencies.

4 Linguistic Analysis

We conduct a linguistic analysis of suicidal posts, as prior studies have shown its efficacy in analyzing mental health-related content [2]. Linguistic analysis involves examining a range of linguistic features within a text, including grammar, syntax, vocabulary, and semantics, providing insights into the structure and context of the language. For example, syntactic analysis entails identifying grammatical categories such as nouns, verbs, adjectives, and adverbs in sentences. Here, we employ various types of linguistic analysis to uncover specific patterns associated with suicidal posts.

 # Subordinating Conjunction: The presence of subordinating conjunctions, which indicates the complexity of sentences, is compared in suicidal and non-suicidal posts. A subordinating conjunction often connects an independent clause to a dependent clause to create a complex sentence. Complex sentences are more difficult to process than simple sentences; nevertheless, they convey a clear and more informative message. A list of 50 commonly used subordinating conjunctions is considered in this study[4].

#Negative Words: Negation is a concept that is used to indicate the opposing or denial of a statement or idea. In English, negation is typically expressed through the use of negative words such as 'not', 'no', 'never', 'none', and 'nothing', among others. The frequency and use of negative words can provide insights into the structure and meaning of language, as well as, the cultural and social contexts in which language is used. We report the percentages of negative words

[4] https://github.com/sazzadcsedu/50SubordinateConjunctions.git.

in both suicidal and non-suicidal posts. The extended VADER [4] negative word list is used as a reference.

#Adjectives, Verbs and Articles. The percentages of adjectives and verbs with respect to the total words are computed for each post. Based on that, we report the mean, median and standard deviation (std.) in the corpus. The spaCy library [3] is used to identify adjectives and verbs in the posts. In addition, the percentages of articles (i.e., *a, an, the*) in the suicidal post are reported.

#Preposition. The percentages of prepositions respect to the total number of words are provided. The following prepositions are considered: *above, across, after, against, among, around, at, time, out, away, before, behind, below, beneath, beside, between, by, down, up, from, in, inside, into, near, next, off, on, onto, outside, over, through, till, to, toward, under, underneath, until.*

#Deontic Modals. Deontic modals are auxiliary verbs that express some kind of necessity, obligation, or moral recommendation [8]. An example of a sentence with a deontic modal could be- "You must wear a seatbelt". We consider the following four common deontic modals: *must*, *should*, *ought*, and *need* in the text.

Table 4. Statistics of various linguistic features in suicidal and non-suicidal posts

Type	Suicidal	Non-suicidal
	Median/Mean/Std.	Median/Mean/Std.
Subordinate Conjunctions	6.51/6.56/2.79	6.82/7.06/5.14
Negation	4.17/4.58/3.29	0.83/1.26/1.64
Verb	15.0/15.16/2.96	12.0/11.67/3.93
Adjective	5.6/5.73/2.30	5.33/5.50/2.82
Preposition	8.45/8.49/2.67	8.94/8.90/3.42
Articles	3.88/3.96/2.04	5.88/6.18/3.15
Deontic modals	0.0/0.27/0.60	0.0/0.26/0.65

Table 4 provides statistics regarding various grammatical attributes including subordinate conjunctions, negation words, verbs, adjectives, and prepositions in suicidal posts. We observe a much higher presence of verbs compared to adjectives in the text, nearly three times as frequent, which contrasts with non-suicidal text where the ratio is approximately two times. Similar presence are noted for subordinate conjunctions, prepositions, and deontic modals in suicidal and non-suicidal corpora. Social media posts are typically more informal and conversational in nature, unlike formal or legalistic texts, which place greater emphasis on obligation and prohibition. Therefore, the limited presence of deontic modals is expected. However, the presence of articles is much lower in suicidal posts compared to non-suicidal posts. The most significant difference pertains to negation words, which appear nearly five times more frequently in suicidal posts, indicating their strong negative connotations.

5 Key Factors of Suicidal Tendencies

We conduct a limited content analysis to identify key factors prevalent in suicidal posts. Based on a manual analysis of a subset of randomly selected 1000 posts, we found that the following contributing factors are highly prevalent in posts discussing suicidal ideation.

Social Isolation and Loneliness: One of the dominant factors that we observe in the suicidal ideation posts is the presence of social isolation and loneliness.

> *# I feel like my whole life I've been abandoned, my parents, friends, and others have left me in the dust for sure.*

The above post suggests deep-seated feelings of abandonment and isolation, indicating a pervasive sense of being left behind by significant people in the individual's life, including their parents and friends. This perception of consistent abandonment can contribute to profound emotional distress, leading to feelings of loneliness, worthlessness, and a lack of belonging.

> *# I'm sick of being alone. People only give a fuck when they have something to gain - even here on this subreddit, ... You wanna help someone? Go outside and talk to someone who looks lonely. Stop fucking around on here. Talk to someone who looks lonely, don't even bother trying to make it profound and interesting, just talk and let it the conversation go where ever it has to.*

The above post reveals feelings of frustration, cynicism, and disillusionment with social interactions. The individual expresses a sense of being undervalued and used by others, suggesting a lack of genuine connection and support in their life.

Relationship Breakup. Marital and other types of relationship breakups can profoundly impact suicidal tendencies. We observe comments depicting various types of relationship breakups with diverse feelings.

> *# My gf left me and if i'm honest I don't/didn't think anyone does/will love me. We dated for around a year and a half, and then out of the blue she sent me a text saying it was over.*

The above post suggests a profound emotional impact from the breakup, including sadness, hurt, rejection, and doubt. The suddenness of the breakup leaves the individual feeling bewildered and questioning their self-worth, as they express doubt about being loved in the future.

> *# My wife and I are splitting up. Its because of years of me choosing weed over my family like my sons birthday, my anniversary, all of it... It's all my fault. The women that I spent 5 years of my life married too, had a wonderful unbelievably amazing daughter and son with.. Idk how to do this. All I want is to kill myself. I don't want to think about.. splitting up material bullshit.. or ..*

In the above post, the mention of wanting to avoid thinking about the separation and material concerns underscores the overwhelming nature of the emotional turmoil the individual is experiencing which contribute to suicidal ideation as a perceived escape from the pain the person is experiencing.

Health Issue. Mental illnesses such as depression, anxiety disorders, bipolar disorder, and schizophrenia significantly elevate the risk of suicide. These conditions often involve intense emotional distress, distorted thinking patterns, and feelings of worthlessness, all of which can contribute to suicidal ideation. Here are some excerpts from related comments:

> *# I've set a dateI'm going to kill myself on December 30th if nothing gets better by then. My story: I've dealt with anxiety, ocd, and panic since I was about 3 years old. I would line up my toys and if one was missing I would freak out. Fears escalated to diseases, rare illnesses, and cleanliness. There was a good half a year where I washed my hands like 20 times a day*

In the above text, the individual recounts a prolonged struggle with anxiety, OCD, and panic dating back to childhood, indicating significant psychological distress. The early onset of these conditions, as early as three years old, emphasizes the chronic nature of their mental health challenges. The individual's experiences with obsessive-compulsive behaviors, such as arranging toys and excessive handwashing, highlight the severity of their symptoms and the impact on their daily life.

Failure and Struggles in Life: Failure in various aspects of life including study, job, or difficulty forming friendships, can significantly impact an individual's mental health and increase the risk of suicidal tendencies.

> *# No Job. No money. I fill out an application for places and get rejected. I go in for interviews and get rejected. I'm so tired of getting rejected. I just hate my life. I hate everything about it.*

The above comment highlights the profound impact of financial strain on the well-being and suicidal tendencies of individuals. The acute stress caused by financial instability has been amplified by the constant cycle of rejection in job searches, resulting in an overwhelming burden of financial hardship.

 # I don't know what to do with myself now. I'm 32. I'll be 33 within a few weeks. I have no money in my bank, I'm in debt a few thousand dollars, and I just lost my job. I'm completely unskilled, and while I am I high school drop-out I do at least have my GED.

The individual's post highlights the profound impact of financial distress on their mental well-being and suicidal tendencies. Despite persistent efforts, the individual faced repeated rejection in job applications, intensifying feelings of worthlessness and hopelessness. The cycle of rejection exacerbates the vulnerability to suicidal ideation.

Traumatic Life Events: Traumatic life events including family abuse and bullying can leave lasting scars and encompass highly distressing and emotionally impactful experiences and causes individuals to have suicidal thoughts.

 # i'm about to end iti'm so fucking done with life. it's pointless for me to keep trying. i don't have a reason to live. none. my mom isn't the best of people. she's always been religious(trying to save my soul at the age of 7), and extremely strict, but when i was 8, stuff got out of control. she screams at me constantly over small stuff, at one point she starved me and my brothers for being ungrateful. she's made me drink a cup of soap before and turns every-damn-thing about her. my dad lets it happen and says BUT SHE IS YOUR MOTHER when i tell him what's been happening. i'm 16. i can't drive, use the phone, or leave the house. i'm too stupid for just about everything.

The above post depicts profound despair and thoughts of suicide, stemming from the individual's experience of severe family abuse. The person describes emotional and physical abuse inflicted by the mother, including screaming, starvation, and forced ingestion of soap, with no intervention from the father. Feeling trapped and helpless in this situation, exacerbated by the individual's young age and lack of autonomy, the suffering becomes evident. experiences.

 # Then my death could at least raise awareness and maybe (definitely) reduce bullying at my school and the pastoral support would double. Maybe that was my purpose after all.

The above post expresses a general sentiment about bullying, highlighting its negative impact on humanity as a whole.

6 Summary and Conclusion

This study aims to understand the multifaceted attributes of suicidal posts on Reddit social media forums. We analyze an array of emotional, grammatical, and syntactic features in the text. Our findings indicate that, at the word level, fear and sadness are the most prominent emotions. At the document level, we observe the dominance of anger-related emotions. Linguistic analysis reveals distinguishing signals associated with suicidal posts, including a high ratio of verbs to adjectives and a high occurrence of negations. Furthermore, the preliminary analysis identifies several prevalent factors associated with suicidal tendencies.

Given the high suicide rate, especially among adolescents, early detection of suicidal tendencies is crucial. As social media has gained mass popularity in recent years, people of all ages use it to express their feelings and emotions. Recognizing the signals and factors associated with suicidal tendencies-often linked to extreme frustration, depression, and feelings of worthlessness-facilitates a better understanding of suicidal behavior and can assist in taking preventive measures, thereby reducing the overall suicide rate. Future research will aim to provide a more comprehensive analysis of the determinants of suicidal ideation through thematic analysis.

Ethical Statement and Limitation. This study utilizes Reddit data that is openly accessible on Kaggle (https://www.kaggle.com/). The research ensures that no personal user information is gathered, utilized, or revealed throughout the analysis or subsequently. Since this study exclusively analyzes suicidal posts sourced from Reddit, there could be inherent bias in the dataset, as it is derived solely from one social media platform and may not capture the diversity of viewpoints and experiences present across other digital platforms.

References

1. Abdul-Mageed, M., Ungar, L.: EmoNet: fine-grained emotion detection with gated recurrent neural networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 718–728. Association for Computational Linguistics, Vancouver (2017)
2. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 7, pp. 128–137 (2013)
3. Honnibal, M., Montani, I.: spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017)
4. Hutto, C., Gilbert, E.: Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, pp. 216–225 (2014)
5. Im, Y., Oh, W.O., Suk, M.: Risk factors for suicide ideation among adolescents: five-year national data analysis. Arch. Psychiatr. Nurs. **31**(3), 282–286 (2017)
6. Ji, S., Pan, S., Li, X., Cambria, E., Long, G., Huang, Z.: Suicidal ideation detection: a review of machine learning methods and applications. IEEE Trans. Comput. Social Syst. **8**(1), 214–226 (2020)
7. Lao, C., Lane, J., Suominen, H., et al.: Analyzing suicide risk from linguistic features in social media: evaluation study. JMIR Format. Res. **6**(8), e35563 (2022)
8. Levine, L.: The distribution of deontic modals in jane austen's mature novels. In: Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 70–74 (2022)
9. Mohammad, S., Turney, P.: Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 26–34 (2010)
10. Ramírez-Cifuentes, D., et al.: Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis. J. Med. Internet Res. **22**(7), e17758 (2020)
11. Sazzed, S.: Banglabiomed: a biomedical named-entity annotated corpus for bangla (Bengali). In: Proceedings of the 21st Workshop on Biomedical Language Processing, pp. 323–329 (2022)
12. Sazzed, S.: A comparative study of affective and linguistic traits in online depression and suicidal discussion forums. In: Proceedings of the 34th ACM Conference on Hypertext and Social Media, pp. 1–6 (2023)
13. Sazzed, S.: Discourse mode categorization of Bengali social media health text. In: Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pp. 52–57 (2023)
14. Shing, H.C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., Resnik, P.: Expert, crowdsourced, and machine assessment of suicide risk via online postings. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp. 25–36 (2018)

15. Sinha, P.P., Mishra, R., Sawhney, R., Mahata, D., Shah, R.R., Liu, H.: # suicidal-a multipronged approach to identify and explore suicidal ideation in twitter. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 941–950 (2019)
16. Tadesse, M.M., Lin, H., Xu, B., Yang, L.: Detection of depression-related posts in reddit social media forum. IEEE Access **7**, 44883–44893 (2019)
17. Zirikly, A., Resnik, P., Uzuner, O., Hollingshead, K.: Clpsych 2019 shared task: predicting the degree of suicide risk in reddit posts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pp. 24–33 (2019)

Extractive Question Answering for Spanish and Arabic Political Text

Sultan Alsarra[1(✉)], Parker Whitehead[2], Naif Alatrush[2], Luay Abdeljaber[2], Latifur Khan[2], Javier Osorio[3], Patrick T. Brandt[2], and Vito D'Orazio[4(✉)]

[1] King Saud University, Riyadh, Saudi Arabia
salsarra@ksu.edu.sa
[2] The University of Texas at Dallas, Richardson, TX 75080, USA
{parker.whitehead,naif.alatrash,luay.abdeljaber,lkhan,pbrandt}@utdallas.edu
[3] The University of Arizona, Tucson, AZ 85721, USA
josorio1@arizona.edu
[4] West Virginia University, Morgantown, WV 26506, USA
vito.dorazio@mail.wvu.edu

Abstract. This study advances the integration of domain-specific large language models (LLMs) for low-resource languages with applications for question-answering (QA). Leveraging on recent LLMs trained to extract events of political violence and conflict, we introduce ConfliBERT-Arabic and ConfliBERT-Spanish, fine-tuned for extractive QA. Contributions include tailored QA fine-tuning techniques for Arabic and Spanish, curation of five datasets, and a comprehensive performance analysis. These new models provide language and domain-specific enhancements over extant models trained on general corpora. Substantively, these tools allow implementation of high-quality QA about conflict and violence in multiple world regions in their native languages.

Keywords: Large language models · Natural language processing · Question answering · Arabic · Spanish

1 Introduction

The ConfliBERT family of domain-specific large language models (LLMs) are trained on texts about political violence and armed conflict in English [10], Arabic [1], and Spanish [18]. The Spanish and Arabic models have focused on binary classification, named entity recognition (NER), and multi-label classification tasks. This study presents an application about question-answering (QA), and contributes to their growth for low-resource languages.

Researchers [11] have characterized three main types of QA: extractive, open-generative, and closed-generative. Closed-generative and open-generative QA both use the model to generate an answer. In contrast, extractive QA identifies the answer to a question in a given context, without generating new text. For instance, when using a news article about conflict in the Middle East, it can

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 144–153, 2024.
https://doi.org/10.1007/978-3-031-72241-7_14

answer questions like which countries are involved or the number of casualties. The output consists of the character indexes in the text that mark the beginning and end of the answer. This is the kind of QA most relevant to the information extraction tasks used in conflict research.

We introduce the ConfliBERT Arabic and Spanish LLMs, fine-tuned for extractive QA for the political violence and conflict domain. Our goal is to elevate their performance in this domain and foster research in QA for low-resource languages. Our work delivers several contributions. First, we implemented fine-tuning techniques tailored to the unique characteristics of Arabic and Spanish for extractive QA. Second, we curated and prepared five datasets, three in Arabic and two in Spanish, providing support for low-resource language processing and expanding NLP accessibility in Spanish by translating the NewsQA dataset from English. Finally, we conducted a comprehensive performance analysis on extractive QA benchmarks, showcasing the enhancements our models offer when compared to base models trained on general corpora.

2 Background and Challenges

BERT [6] opened a new era of NLP capabilities advancing specialized applications including domain-specific tools and improving performance on complex tasks, such as QA. However, domain-specific and QA advances have mostly evolved in parallel without generating integrated tools for domain-specific QA.

2.1 Domain-Specific Developments

Despite its contributions to NLP development, BERT is trained on generic corpora that provide limited leverage for processing text in specialized fields that use technical or specific language. Recognizing this limitation, researchers developed domain-specific adaptations of BERT based on corpora relevant to particular fields. In general, these domain-specific models show higher performance in specialized tasks than the generic BERT [5,15].

In line with domain-specific advances, our research team developed ConfliBERT [10], a BERT-like model specialized on political violence and armed conflict. Results show that ConfliBERT produces much better results than alternative models [8,10], thus reinforcing the trend of enhanced performance by domain-specific models. Based on the initial ConfliBERT architecture in English, this research team advanced multi-lingual extensions of ConfliBERT for Arabic [1] and Spanish [18]. Despite the domain-specific and multi-lingual advances of the ConfliBERT family of models, these tools have not yet been applied to QA.

2.2 Extractive QA Challenges

Extractive QA is widely acknowledged as a complex downstream task for BERT. In this task, the model uses text related to a specific topic as the context, along with a corresponding question. The task is to extract pertinent information from the text to provide an accurate response to the question.

To fine-tune extractive QA, BERT takes the question and context as input, separated by the [SEP] token. The model's output, to the right of the [SEP] token, contains the tokens identified as the 'answer.' Through multiple iterations and training epochs, BERT learns which tokens are relevant for QA.

A key challenge in extractive QA is the limited availability of question-answering datasets, particularly for low-resource languages like Arabic and Spanish. We needed to search for suitable QA datasets, and then modify and process them to fine-tune our models. Discused in detail below, our efforts resulted in the creation of three datasets for Arabic and two for Spanish. These datasets contain the QA pairs used to fine-tuning and test our domain-specific models. We applied extractive QA to both Arabic and Spanish ConfliBERT, comparing the results with their respective base models.

3 Dataset Preparation

For the extractive QA task, we developed scripts tailored to low-resource languages and constructed datasets for ConfliBERT in Arabic and Spanish. The evaluation aimed to measure the effectiveness of our models in real-world applications within the fields of political science and conflict, so it was important to use datasets in this domain.

3.1 Script Development

We developed all scripts from scratch using the HuggingFace library. Our scripts facilitate the QA fine-tuning and evaluation for any language model using Transformers. They have been optimized to handle Arabic and Spanish characteristics, including lower-casing text, removing punctuation, articles, and extra white space. In Arabic, we also removed diacritics like tashkeel and longation, which are phonetic guides unnecessary for the QA task.

To boost the fine-tuning process, we use parallel processing across multiple GPUs to batch multiple fine-tuning jobs by reading JSON arguments. We rely on the National Center for Supercomputing Applications at the University of Illinois with access to 64 CPUs, 248GB of RAM, and 4 A100 GPUs.

All datasets were split into training–used for fine-tuning–and validation–used to assess the model's performance. Both the training and validation sets had five columns: **ID** a unique string for each question-answer pair; **Title** categorizes the context; **Question** the question being asked; **Context** the context in text form; and **Answers** comprising two keys "Answer_Start" (a list of starting indexes for the answers within the context) and "Text" (the raw answer text).

3.2 Spanish Datasets

We use two domain-specific QA datasets for Spanish: NewsQA and SQAC. NewsQA [17] is an extractive QA dataset containing more than 100,000 QA pairs crowd-sourced from CNN articles, with a significant focus on political conflict and violence. The dataset is only available in English, so we translated to Spanish using the Translate Align Retrieve (TAR) [4] method, a proven method for translating extractive QA datasets such as SQUAD [14].

Prepossessing and Cleaning . We accessed the NewsQA data using code made available on Github by Microsoft (github.com/Maluuba/newsqa). Then, we eliminated redundant rows and poor-quality QA pairs, and structured the data into five essential components for our QA tasks: ID, title, context, question, and answers. We stored the formatted data locally in the 'datasetdict' format, which aligns with the extractive QA standards, ensuring consistency and compatibility with other datasets such as SQuAD.

Translation. TAR first uses machine translation to render the question, answers, and context in the target language (Spanish). The original TAR method translates entire contexts as a whole. However, the NewsQA CNN articles are significantly larger than the contexts in SQuAD. Attempting to translate or align them at their full size could introduce inaccuracies and noise. So, we used the NLTK library to split the CNN articles into separate sentences. Then, we employed `opus-mt-en-es`, an English-to-Spanish neural machine translation model for parallel translations. This resulted in a list of translated sentences for each article. The questions were translated directly, without sentence tokenization.

The second TAR step is word alignment. We used `SimAlign`, a novel BERT multilingual approach that aligns source words to their corresponding target words in a sentence. We used the `ArgMax` matching method, as it demonstrated superior performance. By translating the CNN text sentence by sentence, we ensured an equal number of English and Spanish sentences. This allowed us to align only the sentence containing the answer, leading to improved runtime and more efficient memory use. In cases where the answer spanned across multiple sentences, we concatenated and aligned those sentences.

The final TAR step involves retrieving the text. After completing the alignment process, we replaced all English answer text with their corresponding translations in the Spanish text. For answers with empty spaces or gaps, we filled them by highlighting the word with the lowest index next to the word with the highest index, effectively creating our translated answers. Next, we determined the starting index of the answer relative to the translated context, which comprised the concatenation of all sentences from the CNN story. Finally, we converted the dataset back into the datasetdict format, ready for use.

The Spanish Question Answering Corpus (SQAC) [7] is an extractive QA dataset in Spanish. It comprises 6,247 contexts and 18,817 questions, each with 1 to 5 corresponding answers, all sourced from texts originally written in Spanish. These texts were drawn from encyclopedic articles in Spanish Wikipedia, news articles from Wikinews and Newswire, and literary content from AnCora.

To create this dataset we adapted SQuAD v1.1 extractive QA with the help of Native Spanish speakers. Given the dataset's heavy news sources, it contains a substantial number of political QA pairs, providing a valuable resource for evaluating our models in the political domain.

3.3 Arabic Evaluation Datasets

The XQUAD dataset (Cross-Lingual Question Answering Dataset) [2] was developed by Google Deepmind to assess cross-lingual extractive QA performance. It includes 240 paragraphs and 1,190 QA pairs originally sourced from the SquAD v1.1 dataset [14] and translated into multiple languages, including Arabic. For our evaluation, we used XQUAD Arabic, which was prepared for extractive QA tasks by the XTREME benchmark [9]. The Arabic XQUAD dataset contains numerous questions on political topics and subjects.

The MLQA (MultiLingual Question Answering) dataset [12] was developed by Facebook Research to assess cross-lingual QA performance. It includes over 5,000 QA pairs multiple languages, including Arabic. The data are in SQaAD format, have been machine-translated from Wikipedia paragraphs, and have a substantial number of political topics and questions.

The Arabic Reading Comprehension Dataset (ARCD) [13] comprises 1,395 crowd-sourced QA pairs using Arabic Wikipedia articles. The dataset categorizes the answers into numerical answers, such as dates, and non-numerical answers, such as verbs, nouns, or adjective phrases. In cases of noun phrases, they checked for named entities and conducted manual verification. Additionally, the authors manually labeled the dataset for synonyms, world knowledge, syntactic variation, multiple-sentence reasoning, and ambiguity to facilitate question reasoning. We split the dataset into train/test sets, with 702 questions drawn from 78 articles in the test set. Many of these articles focus on political subjects and topics.

4 Experimental Setup

We fine-tuned eight models in Spanish and five in Arabic. For Spanish, the ConfliBERT models were base-multilingual-cased, base-multilingual-uncased, BETO-cased, and BETO-uncased. Each of these has a corresponding BERT model. For Arabic, the fine-tuned ConfliBERT models were Arabic-v2-araBERT, Arabic-v2-multilingual-uncased, and Arabic-v2-Scratch. The comparable base models were BERT-base-araBERT and BERT-base-multilingual-uncased.

We assessed our model's performance using two standard metrics: Exact Match and F1 Score. For fine-tuning, we maintained consistency by using the same hyperparameters across all models. These hyperparameters follow best practices for fine-tuning BERT-based models, as outlined in the original BERT paper [6]. We ran experiments over 5 epochs with 5 different seeds, using a batch size of 8 and a learning rate of 5e-5, a maximum answer length of 100 and a maximum context length of 384.

5 Results and Analysis

The domain-specific ConfliBERT models consistently outperform more generic BERT models for these extractive QA tasks. The QA datasets had questions about political violence, wars, elections, protests, etc., and thus the ConfliBERT models were able to make use of their domain-focused training.

5.1 ConfliBERT-Spanish QA

The results of the Spanish models fine-tuned for extractive QA are shown in Table 1, with Table 1a, showing the average performance across the two datasets. The ConfliBERT-Spanish model outperformed the comparable base BERT model, and often by a substantial margin. The F1 Score for ConfliBERT BETO-Uncased is nearly 7 points higher than its BERT counterpart. Overall, the best model was the ConfliBERT BETO-Cased.

Table 1. Results for Spanish

Model Name		(a) Extractive AQ		(b) News QA		(c) SQAC	
		F1 Score	Exact Match	F1 Score	Exact Match	F1 Score	Exact Match
ConfliBERT Spanish	Cased	70.14	48.00	62.76	33.04	77.51	62.88
	Uncased	69.92	47.90	63.01	33.38	76.83	62.39
	BETO-Cased	**72.30**	**50.21**	64.88	35.08	**79.72**	**65.34**
	BETO-Uncased	72.15	50.16	**65.53**	**35.19**	78.77	65.12
BERT	Cased	69.85	44.16	59.74	30.70	72.96	57.62
	Uncased	66.61	43.98	60.19	30.06	73.02	57.89
	BETO-Cased	71.20	48.85	63.39	33.64	79.00	64.06
	BETO-Uncased	65.71	43.78	59.60	30.47	71.82	57.08

5.2 ConfliBERT-Arabic QA

Table 2 shows the results for the Arabic models, with Table 2a showing the average across datasets. Once again, the ConfliBERT models performed better than their BERT counterparts in every case. For each of the three datasets and for each metric, the best performing model was ConfliBERT AraBERT.

Table 2. Results for Arabic.

Model Name		(a) Extractive QA		(b) MLQA		(c) XQUAD		(d) ARCD	
		F1 Score	Exact Match	F1 Score	Exact Match	F1 Score	Exact Match	F1 Score	Exact Match
ConfliBERT Arabic-v2	AraBERT	**61.90**	**40.11**	**64.86**	**44.24**	**63.33**	**47.19**	**57.43**	**28.92**
	Uncased	60.76	37.79	64.11	43.47	62.21	46.10	55.95	23.79
BERT	AraBERT	60.18	38.64	63.41	42.95	62.29	46.20	54.84	26.78
	Uncased	58.35	35.50	62.16	41.00	60.55	44.54	52.33	20.94

5.3 Evaluation of Answers

Here, we present examples of QA pairs used to assess our models. We conducted a comparative analysis, comparing the results of the best-performing ConfliBERT-Arabic models against the best-performing base BERT models. We also provide a brief comparison with the responses generated by ChatGPT [16].

The initial set of questions focused on the context related to the former president of Egypt, Hosni Mubarak. The first question posed in Arabic was Q1: "When did Hosni Mubarak take over the reins of power in Egypt?" We supplied the models with the context for Q1, and they generated responses based on this information. The context and answer for Q1 in Arabic, along with its equivalent translation in English, are provided in Fig. 1. The correct answer to Q1 is October 1981, with its position highlighted in the context.

محمد حسني السيد مبارك وشهرته حسني مبارك (ولد في 4 مايو 1928، كفر المصيلحة، المنوفية) هو الرئيس الرابع لجمهورية مصر العربية من 14 أكتوبر 1981 خلفا لمحمد أنور السادات، وحتى في 11 فبراير 2011 بتنحيه تحت ضغوط شعبية وتسليمه السلطة للمجلس الأعلى للقوات المسلحة. حصل على تعليم عسكري في مصر متخرجا من الكلية الجوية عام 1950، ترقى في المناصب العسكرية حتى وصل إلى منصب رئيس أركان حرب القوات الجوية، ثم قائداً للقوات الجوية في أبريل 1972م، وقاد القوات الجوية المصرية أثناء حرب أكتوبر 1973. وفي عام 1975 اختاره محمد أنور السادات نائباً لرئيس الجمهورية، وعقب إغتيال السادات عام 1981 على يد جماعة سلفية إسلامية مصرية تقلد رئاسة الجمهورية بعد استفتاء شعبي، وجدد فترة ولايته عبر استفتاءات في الأعوام 1987، 1993، و1999 وبرغم الانتقادات لشروط وآليات الترشح لانتخابات 2005، إلا أنها تعد أول انتخابات تعددية مباشرة وجدد مبارك فترته لمرة رابعة عبر فوزه فيها. تعتبر فترة حكمه (حتى إجباره على التنحي في 11 فبراير عام2011) رابع أطول فترة حكم في المنطقة العربية - من الذين هم على قيد الحياة آنذاك، بعد السلطان قابوس بن سعيد سلطان عمان والرئيس اليمني علي عبد الله صالح والأطول بين ملوك ورؤساء مصر منذ محمد علي باشا.

Muhammad Hosni Al-Sayyid Mubarak, known as Hosni Mubarak (born on May 4, 1928, Kafr Al-Masaylaha, Menoufia) is the fourth president of the Arab Republic of Egypt from 14th October, 1981, succeeding Muhammad Anwar Sadat, until February 11, 2011, when he stepped down under popular pressure and handed over power to the Supreme Council of the Armed Forces. He received a military education in Egypt, graduating from the Air Force College in 1950. He rose through the military ranks until he reached the position of Chief of Staff of the Air Force, then Commander of the Air Force in April 1972, and led the Egyptian Air Force during the October 1973 War. In 1975, Muhammad Anwar Sadat chose him as Vice President of the Republic. Following Sadat's assassination in 1981 at the hands of an Egyptian Islamic Salafist group, he assumed the presidency of the republic after a popular referendum. He renewed his term through referendums in the years 1987, 1993, and 1999. Despite criticism of the conditions and mechanisms for running for the 2005 elections, they are considered the first direct pluralistic elections. Mubarak renewed his term for a fourth time by winning it. His reign (until he was forced to step down on February 11, 2011) was considered the fourth longest in the Arab region - among those alive at the time, after Sultan Qaboos bin Said, Sultan of Oman, and Yemeni President Ali Abdullah Saleh, and the longest among the kings and presidents of Egypt since Muhammad Ali. Pasha.

Fig. 1. Q1 Context and Highlighted Answer

The ConfliBERT-Arabic model predicted the answer as "1981," correctly identifying the year but omitting the month. In contrast, base BERT incorrectly predicted "Year 1950" as the answer, corresponding to the year when Hosni Mubarak graduated from the Air Force College, as indicated in the context.

After introducing additional context, we prompted the models with a more intricate question concerning Hosni Mubarak: "To whom did Hosni Mubarak hand power after the 2011 protests?" The correct answer for Q2 is "to the Supreme Council of the Armed Forces." Fig. 2 provides the context, along with the highlighted answer for Q2. The ConfliBERT-Arabic models accurately predicted the answers, providing the exact response. Conversely, the base BERT model generated an incorrect answer, offering the date when Mubarak handed over power as "11 2022 February."

محمد حسني السيد مبارك وشهرته حسني مبارك (ولد في 4 مايو 1928، كفر المصيلحة، المنوفية)
هو الرئيس الرابع لجمهورية مصر العربية من 14 أكتوبر 1981 خلفا لمحمد أنور السادات، وحتى
في 11 فبراير 2011 بتنحيه تحت ضغوط شعبية وتسليمه السلطة للمجلس الأعلى للقوات المسلحة.

```
Muhammad Hosni Al-Sayyid Mubarak, known as Hosni Mubarak (born
on May 4, 1928, Kafr Al-Masaylaha, Menoufia) is the fourth
president of the Arab Republic of Egypt from October 14, 1981,
succeeding Muhammad Anwar Sadat, until February 11, 2011, when
he stepped down under popular pressure and handed over power to
the Supreme Council of the Armed Forces.
```

Fig. 2. Q2 Context and Highlighted Answer.

Overall, ConfliBERT-Arabic extractive QA models consistently outperformed base BERT, providing accurate responses to most of the questions. Notably, base BERT faced challenges, particularly with political questions. Also, our ConfliBERT-Arabic demonstrated proficient handling and accurate representation of date formats and numbers in answers. In contrast, base BERT struggled, presenting dates and numbers in incorrect orders and formats.

We also conducted a comparative analysis with ChatGPT. For instance, we prompted ChatGPT with "Answer questions based on the following text:" and then provided the same Q1 context in Arabic as shown in Fig. 1. We then asked ChatGPT the same Q1 question, "When did Hosni Mubarak take over the reins of power in Egypt?" It provided the answer in Arabic, and we then requested it to be translated into English. ChatGPT provided a correct answer (October 14, 1981) but introduced inaccuracies by stating that Hosni Mubarak assumed power after the *resignation* of President Mohamed Anwar Sadat. In reality, Sadat did not resign, he was assassinated.

In our experiments with ChatGPT, particularly in the analysis of political text, we observed a tendency to infer extra details not present in the text, potentially leading to incorrect answers and biased responses based on question wording. These issues highlight the importance of caution when relying on models like ChatGPT for detailed and accurate analysis of political texts. In contrast, our domain-specific ConfliBERT models provide answers directly from the provided text and context without introducing extraneous details or generating nonsensical responses. This is an extremely important attribute for researchers who want to use these models for information extraction tasks.

6 Conclusion and Future Work

We introduced extractive QA for ConfliBERT-Arabic and ConfliBERT-Spanish models. This involved crafting an extractive QA methodology from the ground up for the unique aspects of Arabic and Spanish. We curated extractive QA datasets and undertook the translation of an English dataset to Spanish, contributing to the needs of low-resource languages in NLP. Our evaluation compared model performance against the base BERT models showing how our Arabic and Spanish models excelled especially in the domains of politics and violence.

Our future research develops more QA datasets for the political and conflict domains. We also plan to enhance our dataset translation methodologies and introduce closed-generative QA. A significant challenge in extractive QA for low-resource languages is the dearth of non-machine-translated datasets. Crafting specialized datasets for Arabic and Spanish will prove instrumental to advance the use of low-resource languages in NLP. Furthermore, refining translation techniques will extend linguistic resources for low-resource languages by enabling the translation of existing QA datasets. Our future ConfliBERT tasks aspire to create closed generative QA and expand into areas like summarization.

7 Ethical Considerations

The tools generated here provide NLP resources tailored for languages with low resources to reduce bias in academia and the policy sector. This study relies exclusively on secondary sources of information such as news reports and does not engage with human subjects to gather information from primary sources. To address concerns of biased ML inputs, it complies to select corpora and training data [3]. Due to copyright protecting the original sources, we cannot share the raw data.

Acknowledgements. This research was supported in part by NSF awards 1737978, 2039542, 1828467, 1931541, 2311142, and 1906630; ONR awards 17-1-2995 and 20-1-2738; and Army Research Office Contract No. W911NF2110032. This work used Delta at NCSA/University of Illinois through allocation CIS220162 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants 2138259, 2138286, 2138307, 2137603, and 2138296. This material is based upon High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII).

References

1. Alsarra, S., et al.: Conflibert-arabic: a pre-trained arabic language model for politics, conflicts and violence. In: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, pp. 98–108 (2023)
2. Artetxe, M., Ruder, S., Yogatama, D.: On the cross-lingual transferability of monolingual representations. arXiv preprint arXiv:1910.11856 (2019)
3. Barberá, P., Boydstun, A.E., Linn, S., McMahon, R., Nagler, J.: Automated text classification of news articles: a practical guide. Polit. Anal. **29**(1), 19–42 (2021)
4. Carrino, C.P., Costa-juss, M.R., Fonollosa, J.A.R.: Automatic Spanish translation of the squad dataset for multilingual question answering (2019)
5. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: Legal-bert: the muppets straight out of law school. arXiv preprint arXiv:2010.02559 (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

7. Gutiérrez-Fandiño, A., et al.: Maria: Spanish language models. arXiv preprint arXiv:2107.07253 (2021)
8. Häffner, S., Hofer, M., Nagl, M., Walterskirchen, J.: Introducing an interpretable deep learning approach to domain-specific dictionary creation: a use case for conflict prediction. Polit. Anal. **31**(4), 481–499 (2023)
9. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In: International Conference on Machine Learning, pp. 4411–4421. PMLR (2020)
10. Hu, Y., et al.: ConfliBERT: a pre-trained language model for political conflict and violence. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5469–5482 (2022)
11. Lee, K., Salant, S., Kwiatkowski, T., Parikh, A., Das, D., Berant, J.: Learning recurrent span representations for extractive question answering. arXiv preprint arXiv:1611.01436 (2016)
12. Lewis, P., Oğuz, B., Rinott, R., Riedel, S., Schwenk, H.: MLQA: evaluating cross-lingual extractive question answering. arXiv preprint arXiv:1910.07475 (2019)
13. Mozannar, H., Maamary, E., El Hajal, K., Hajj, H.: Neural Arabic question answering. In: Proceedings of the Fourth Arabic Natural Language Processing Workshop, pp. 108–118. Association for Computational Linguistics, Florence (2019). www.aclweb.org/anthology/W19-4612
14. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. arXiv e-prints arXiv:1606.05250 (2016)
15. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Dig. Med. **4**(1), 86 (2021)
16. Ray, P.P.: Chatgpt: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys. Syst. **3**, 121–154 (2023)
17. Trischler, A., et al.: Newsqa: a machine comprehension dataset. In: Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 191–200 (2017)
18. Yang, W., et al.: ConfliBERT-Spanish: a pre-trained Spanish language model for political conflict and violence. In: Proceedings of The 5th IEEE Conference on "Machine Learning and Natural Language Processing: Models, Systems, Data and Applications" (2023)

Accountability in Search Engine Manipulation: A Case Study of the Iranian News Ecosystem

Peter Carragher(✉) and Kathleen M. Carley

Carnegie Mellon University, Pittsburgh, USA
{pcarragh,kathleen.carley}@andrew.cmu.edu

Abstract. This article investigated the influence of Search Engine Optimization (SEO) on the Iranian news ecosystem. The political alignments of 348 news sites were labeled using GPT-4 and clustering patterns corresponding to political alignments in the webgraph were found. Topic modeling uncovered topics relevant to each alignment's framing and bias. More importantly, a criterion for distinguishing between paid-for SEO services and information operations was developed. The identification of a unique clique of link schemes suggested a potential information operation, with links to both Iranian and Mandarin-language sites. This distinction raises several challenges, including the lack of transparency and accountability in SEO practices, and highlights the need for future research into the role of SEO in influence operations.

Keywords: search engine optimization · political bias · topic modeling · webgraph

1 Introduction

The apparent influence of Search Engine Optimization (SEO) on news dissemination has raised questions about transparency and accountability within the media ecosystem [4,15]. Despite the importance of determining the nature and intent of SEO services in creating accountability within the information space, the distinction between information operations and regular paid-SEO services has not been explored [5]. This paper addresses that gap taking the Iranian news ecosystem as a case study.

The primary research question is how to untangle paid SEO from unpaid, potentially strategic, SEO operations. If SEO is paid for, the *target* of the link is accountable for leveraging these services. Conversely, if SEO is unpaid and appears to be part of a deliberate strategy, the *source* of the link may be engaging in an information operation, warranting scrutiny. As such, establishing accountability is important and this distinction can inform policies for maintaining the integrity of news networks and information ecosystems.

Following studies of political alignment in Iranian news networks [18], the clustering patterns of these news domains in the webgraph are analyzed and the political alignments which benefit from SEO practices are identified. Based

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 154–163, 2024.
https://doi.org/10.1007/978-3-031-72241-7_15

on previous work detecting framing and political bias in news articles [11], our topic modeling approach reveals how specific topics are covered by news outlets based on political alignment. The prominent link schemes influencing these news networks are identified[1], and paid and unpaid SEO services are contrasted according to their link profiles, the topics they link to, and their intrinsic characteristics. The role of unpaid SEO in information operations raises concerns about international influence and coordination.

This research builds on the emerging science of social cybersecurity [3], which explores the intersection of cybersecurity and information dissemination. By applying a social cybersecurity lens to the Iranian news ecosystem, this study aims to determine the underlying incentives of link scheme sites and shed light on the opaque practices of SEO. In doing so, the mechanisms of information manipulation can be better understood and strategies to ensure accountability and integrity in the news ecosystem can be developed. This research highlights the need for greater transparency in SEO practices and calls for further investigation into the role of SEO in influence operations.

This paper is organized into four parts; the construction of a labeled Iranian news network dataset, link and content analysis of the news sites, analysis of SEO and link scheme influences on these news sites, and a discussion on accountability in SEO.

2 Iranian News Networks

Previous work has described the main factions within the Iranian News ecosystem and the prominent outlets for the different political alignments [18]. The political alignment of these news sources is generally cast along an axis from conservative, fundamentalist, pro-government sources to left-leaning reformist sources that are affiliated with the long-standing Reformist political movement in Iran [18]. Neutral or centrist sites are those that adhere to apolitical reporting. Finally, a fourth group of news sources, labeled as anti-government, typically consist of English-language news sources seeking regime change.

Zanconato and Sabahi [18] give an in-depth discussion of these factions alongside the most prominent news networks from each (n = 15). In addition, two other sources to construct a list of Iranian News Networks are utilized—namely the Armenian School of Languages and Cultures (ASPIRANTUM) (n = 62)[2] and the w3newspapers global database of news sites (n = 271)[3].

GPT Labeling. Only 31 of the 348 news sites were labeled with their political alignment by their respective sources. This was problematic because, as researchers unfamiliar with Iranian politics and Farsi, the authors were not confident in coding these sources themselves. Fortunately, the ability of LLMs to classify news sites has been demonstrated [17].

[1] A link scheme is an organized strategy to manipulate search engine rankings through unnatural link building practices [4,14].

[2] https://aspirantum.com/blog/iranian-news-media-outlets-and-newspapers.

[3] https://www.w3newspapers.com/iran/.

GPT4 labeled all 348 news sites with the following prompt: ‘You label Iranian news domains as supportive of the government, reformist, anti-government, or neutral.’ GPT4 labels were evaluated against the 31 labeled sites—with agreement on 27 out of 31 news sites, 87% accuracy was sufficient for the following analysis. The sample was heavily skewed towards sources with a fundamentalist political alignment, with 150 fundamentalist, 137 neutral, 29 reformist, and 17 anti-government labels. Neutral was the second largest category, which can be explained by the prevalence of self-censorship in the Iranian information space [18].

3 News Network Analysis

Webgraph. Using Ahrefs API [1], the top 50 backlinking domains for each of the news sites in our list were pulled and any backlinking domain that was not already in the list was discarded[4]. Within the backlink network, assortativity based on political alignment is observed (Fig. 1). Conservative sources tended not to link to reformist sources. Government-run sources and anti-government sources were at opposite extremes, with paths between them spanning the diameter of the network.

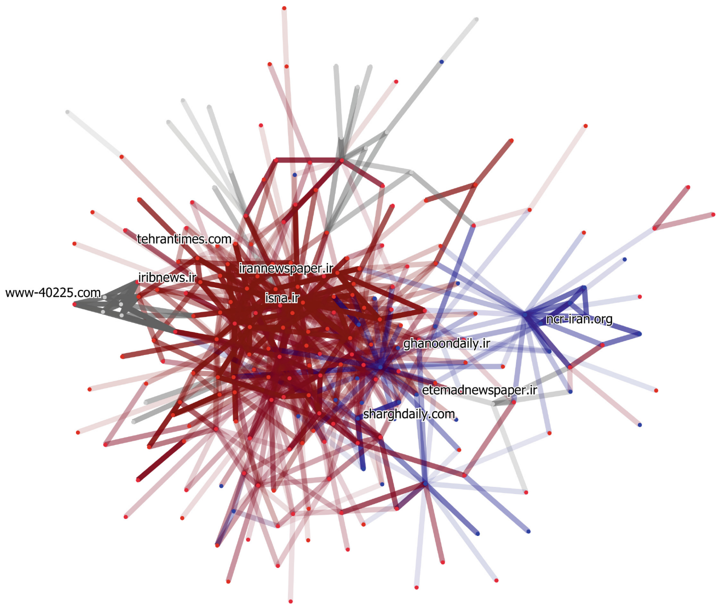


Fig. 1. The backlink network with edges colored by source node. Convservative and neutral sources are red, reformist and anti-government are blue. Link schemes are grey. (Color figure online)

[4] With the exception of link scheme sites, which were kept in the link network for further analysis (see Sect. 4).

Table 1. The top 10 topics on economics, politics, and religion per political alignment. Topics are ordered by the % of articles in that topic that fall under the given alignment. Topic words are translated from Farsi using Google Translate.

Conservative Topics (n = 150)	Reformist Topics (n = 29)
Student Palestine University Universities	Spinosa Existence Hegel Philosopher
Prisoners Judiciary Educational Power	Price Ethereum Queen Thousands
Narrations Muslims Ramezan Prayer	Nazif Constituency Tahan Candidate
Commander Forces Brigadier General	Execution Zanjani Imprisonment Condemned
Celebration Dignity Fatima Shrine	Ghalibaf Twelfth Parliament Bills
Wastewater Drink Village Governor	Charge Prosecutor Arrest Fakhor
Field Academic Science University	Uranium Conference Resolution Atomic
Aegean Justice Power Judicial	Currency Exchanges Digital Cryptocurrency
Commission Representatives Budget General	Javad Mohammad Controversial Delicate
Student Protesters Columbia Suppression	Canada Educational College Migration
Farsi Missile Military Israel	Agents Capture Hostage Extract
Yemen Missile Ocean Ballistic	Aegean Justice Power Judicial
Khamenei Muslims Ebrahimi Farry	Blinkin Riyadh Allies Washington
Security Policy Excellency Fatah	Federation Discipline Stadium Soccer
Reading Transformed Quran Signs	Faith Clashes Shari'a Collision Chastity Hijab

Scraping. Newspaper3k was used to scrape articles from 348 news sources. The scraper fetched links to every article appearing on the front page of the news site, and then scraped the content for each article individually. This resulted in 13,000 articles. Those that contained less than 200 characters were filtered out, leaving 10,000 articles. Most articles were between 1,000 and 40,000 characters. In addition, more than 9,000 of the 10,000 were in Farsi, with a small number of English-language articles. Articles were collected over the first two weeks of May 2024. Most were published in this time period, with a long tail published as far back as February 2024.

Topic Modeling. Topic modeling is a well-established method for large-scale news content analysis [6,8,20], particularly for revealing framing effects [7,9,11]. The results of training a language-agnostic Doc2Vec topic model [12,13] on the scraped articles is reported (Table 1), ranking politically relevant topics by the proportion of their articles under the given affiliation. The most predominant topics among each affiliation are in line with expectations; Conservative sources preferred topics on the justice system, national security, and religious ceremonies. Reformist sources centered around philosophy, elections, and governance.

While topics do touch on religious matters and national security, there is a difference in framing and they tend to focus on conflict (clashes of faith, uranium resolutions) rather than celebration. On the other hand, anti-government sources are more outwardly focused, with multiple foreign policy-related topics on US and Russian relations with the Arab world. That topics are aligned with their sources' aims validates the topic modeling approach. Of particular interest is the slanted coverage of US Student Protests against the war in Gaza. Reformist and Anti-Government sources do not cover this topic, and it is most heavily covered by fundamentalist-aligned sources. This warrants further investigation into how Pro-Palestine news articles published by Middle Eastern sources interact with ongoing social unrest on US college campuses.

The differences in topics between neutral and anti-government sources is also explored, with tables given in the appendix. Neutral sources tend to cover apolitical and uncontroversial topics such as technology, lifestyle, and commerce. In contrast, anti-government topics tend to focus on governance and the justice system. Our methods are validated by agreement between political alignment labels and the topics covered in the respective news sources.

4 Link Scheme Analysis

A set of link schemes are identified among the backlink domains according to the link scheme identification method proposed by Carragher et al. [4]; these are seen in Fig. 1 as a clique of grey nodes on the left side. In addition, a second set of link schemes self-identify as SEO service providers based on their domain name containing any of the substrings 'SEO', 'directory', 'rank', 'link', 'article', or 'site'. We term these multi-category link schemes, as these link schemes have previously been found to link to many disparate categories of domains [5]. They comprise a separate cluster of grey nodes at the top of Fig. 1.

Multi-category link schemes are not considered to be information operations and sell blackhat backlinking services to any customer, resulting in a unique link distribution that bridges sites of unrelated categories and topics. In contrast, link schemes engaging in information operations do not explicitly sell services and instead tend to target a narrower range of sites to promote the specific messaging of a small range of categories or topics, typically news-related [15].

Camouflage. The link scheme clique seems to engage in camouflage by masquerading as gambling sites instead of SEO sites. When investigated, the gambling aspect of the site was falsified and these sites as a whole appear to be 'shell' sites, without any clear purpose. In contrast, the multi-category link schemes have a clear purpose in that they explicitly sell a blackhat SEO service and do not attempt to hide this. Example screenshots are provided in Fig. 2.

Topics and Political Alignment. To determine whether the link scheme clique is multi-category, or potentially part of an information operation, the content of the news sites that are promoted by the clique is analyzed. The list of top political, economic, and religious topics for SEO-supported sites is given in the appendix (Table 2), which compares topics supported by the clique and the multi-category

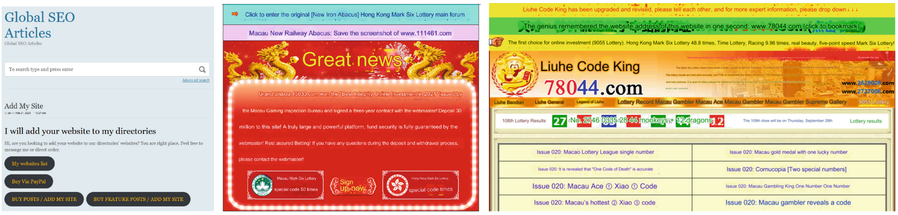


Fig. 2. Left: a paid link scheme. Center and right: unpaid link scheme sites disguised as gambling sites. These have been translated from Mandarin.

link schemes. In particular, the four Iranian news sources linked to by the clique (vaghtesobh, moniban, shomanews, fardanews) are all labeled as having conservative alignment, with religious and national security-related topics featuring prominently in the 499 articles scraped from these sites.

Table 2. The top political, economic, and religious topics for Iranian news sites linked to by a) the link scheme clique, and b) a set of multi-category link schemes.

Link Scheme Clique Topics (n = 4)	Multi-Category Link Scheme Topics (n = 5)
Major-general Military Corps Ballistic	Presidency Selected Eleventh Fundamentalists
Murder Interrogation Criminal Murderer	Colonel Suspects Judicial Capture
Students Protesters Demonstrations Movement	Citizens Asphalt Development Municipality
Klein Shias Narrations Innocent	States Allies Washington Saudi
Encouraging Cooperative Revenue Households	Income Guilds Declaration Non-commercial
Presidency Selected Eleventh Fundamentalists	Second Chair Constituency Candidates
Rights Livelihood Pension Sunni	Development Economic -operative Participation
Emirates Toman Exchange Banknotes	Biden Netanyahu Civilians Weapons

In contrast, multi-category link schemes spread links to sites of varying political alignment (Fig. 3). Topics covered by these sites are not as conservative as for the clique; instead of religious topics, they are concerned with the justice system. Instead of national security, they cover international relations (Table 2).

External Links. Multi-category link schemes rarely link to any site more than 10 times while the clique links to the Iranian news networks at least 20M times each. No Iranian news sites appear in the top 1000 outlinked domains for multi-category link schemes while the top 14 linked domains from the clique feature 5 Iranian news sites, 6 popular Chinese domains, and 3 link scheme sites. According to Similarweb data [2], there are no news sites in top 100 outlinked domains of multi-category link schemes, while there are 12 news sites within the top 100 clique outlinks. Interestingly, several US news sources rated as unreliable sources of misinformation by mediabiasfactcheck.com are found among the top 1000 clique outlinks (Fig. 4) [19].

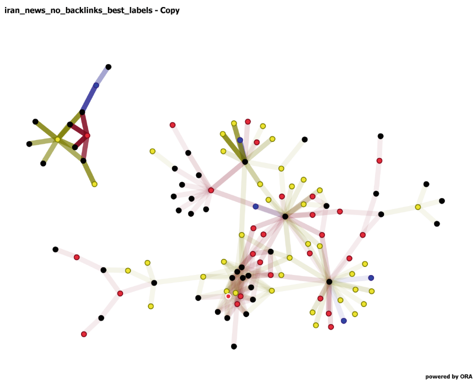


Fig. 3. Links between multi-category link schemes (black) and Iranian news sites of varying political alignment; red = conservative, blue = reformist, yellow = Neutral. (Color figure online)

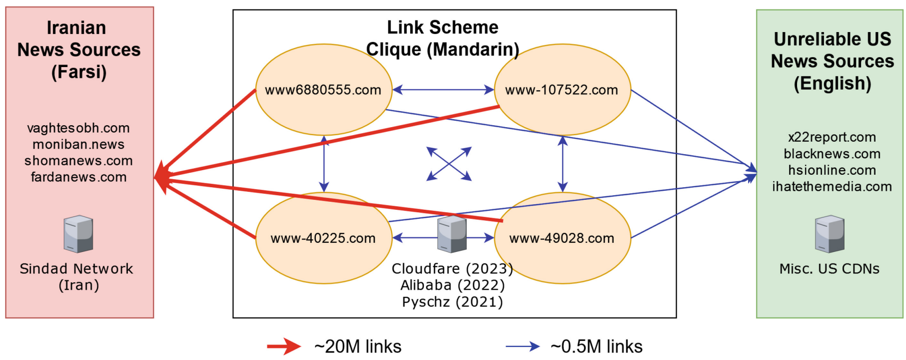


Fig. 4. A clique of link scheme sites link to conservative Iranian news sites and unreliable US news sources. Host server location and site languages are given.

Hosting. In addition, while the current IP address is hosted by Cloudfare, all four site DNS histories show that they were previously hosted on Alibaba and Psychz Networks before switching to Cloudfare[5]. The link clique sites appear to be jointly operated given their overlapping DNS histories, identical link distributions, and identical outlink targets. Likewise, the four Iranian news sites they link to are all hosted on Iranian servers owned by Sindad Network Technology Ltd[6] and also appear to be jointly operated.

[5] https://securitytrails.com/domain/www6880555.com/history/a.
[6] https://whatismyipaddress.com/ip/31.193.186.243.

Table 3. Summary of differences between multi-category and clique link schemes.

Link Scheme	Hosting	Camo	Link Counts	Political Alignment		Outlink Category	
				Topics	Webgraph	Top 10	Top 100
Multi-Category	Joint	No	10	Mixed	Mixed	SEO	SEO
Clique	Joint	Yes	20M	Conservative	Conservative	News	News

5 Discussion

This case study aims to elucidate an unknown news ecosystem that is difficult to penetrate due to contrasting language and cultural contexts. The study begins by compiling a list of 348 Iranian news sites, of which 31 are labeled according to their political alignment (test set). A few-shot approach with GPT4 was employed to label the majority. While the GPT labels proved accurate when compared with the test set, webgraph and topic modeling analysis provided additional validation.

Previous research has found homophily in the linkage patterns of US news sources, resulting in news sources clustering together based on their political alignment [4]. Additionally, it has been found that document topics cluster based on political alignment, and that topic models can capture differences in political framing [7,9,11]. As such, GPT labels were further validated by the clustering of fundamentalist, neutral, reformist, and anti-government news sources in the webgraph (Fig. 1), and the emergence of topics and framings relevant to each alignment by topic modeling (Table 1).

Having validated the labeling method, an analysis of the various SEO services that link to news sources in the webgraph was conducted. In particular, a set of multi-category link schemes, as defined in previous work [5], and a unique clique of link scheme sites were identified. These two groups of link scheme sites were contrasted along several dimensions, summarized in Table 3. This comparison suggested that the link scheme clique did not fit the multi-category link scheme definition and instead represented an information operation.

However, the use of camouflage distinguishes these clique sites from link schemes that promote pro-Kremlin misinformation and masquerade as legitimate blogs addressing current affairs [15]. Additionally, the shared ownership of the targeted news domains is atypical of an information operation where links are spread to sites based on content and political framing instead of ownership.

The Iranian news network serves as an example of an ecosystem where it is unclear whether the offending sites are 'paid for' SEO services, if so who might have contracted them, or if not, who is behind the influence campaign. Unfortunately, as a result of the distinct lack of accountability in SEO practices, there is no way to know for certain which is the case, and this is exacerbated when SEO services operate in spaces with a lack of transparency.

If the services of these link schemes were not purchased, there is reason to suspect that these sites are part of an information operation. Previous reports of coordination between Iran & China on misinformation are relevant to this finding [10,16], given that content on these sites is in Mandarin (Fig. 2) and popular Chinese domains are the second biggest benefactors of links from this clique after conservative Iranian news site. The possibility of coordination makes the lack of transparency and accountability in SEO a pertinent problem and this joint issue is established as an avenue for future work.

6 Limitations

To scale this investigation and uncover similar activities across different ecosystems, several steps are necessary. First, robust methodologies for automated labeling of news sites based on political alignment need to be further developed and validated. As in prior work [17], the evaluation of LLMs for few-shot labeling on international and multilingual news networks is limited by training materials for the LLMs, as well as their multi-lingual capabilities. Likewise, more sophisticated topic modeling approaches than Top2Vec should be explored.

Second, comprehensive web scraping and data collection tools must be deployed to capture link patterns across a wide range of sites. To mitigate this, our list of news domains was compiled from several sources and Ahrefs data was utilized for webgraph construction [1]. However, topic analysis is subject to artifacts of the scraping process that may bias results.

Thirdly, the framework developed to identify and characterize differences between paid and unpaid SEO is mainly qualitative (Table 2). While indicative, these criteria are phenomenological, and certain differences such as the use of camouflage (Fig. 2) or the political alignments of linked domains (Table 3) may be coincidental. Future work should focus on more quantitative approaches.

Finally, this study motivates future work on SEO accountability, the paid vs. unpaid SEO distinction, and the role of SEO in information operations. Collaborations with international research organizations and cybersecurity experts would serve to enhance the reach of such investigations.

7 Conclusion

There is a need for transparency and accountability in SEO practices to safeguard the integrity of information ecosystems. Understanding whether SEO activities are paid or unpaid services has implications for media trust, public perception, and geopolitical influence. By highlighting the potential for link schemes to be used as part of information operations in the Iranian news ecosystem, this study underscores how external pressures can exert influence on a regional information ecosystem and risk undermining its integrity. As such, this study motivates the development of regulatory frameworks and technological solutions to mitigate these risks and promote SEO accountability.

Acknowledgements. This work was supported in part by Office of Naval Research grants (N000142112749, N000142112229) and the Knight Foundation. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research, or the U.S. Government.

References

1. Ahrefs: Seo tools to grow your search traffic (2024). https://ahrefs.com/
2. Similarweb (2024). https://www.similarweb.com/corp/ourdata/
3. Carley, K.M.: Social cybersecurity: an emerging science. Comput. Math. Organ. Theory **26**(4), 365–381 (2020)
4. Carragher, P., Williams, E.M., Carley, K.M.: Detection and discovery of misinformation sources using attributed webgraphs. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 18, pp. 214–226 (2024). https://doi.org/10.1609/icwsm.v18i1.31309
5. Carragher, P., Williams, E.M., Carley, K.M.: Misinformation resilient search rankings with webgraph-based interventions. ACM Trans. Intell. Syst. Technol. (2024). https://doi.org/10.1145/3670410
6. Ghasiya, P., Okamura, K.: Investigating covid-19 news across four nations: a topic modeling and sentiment analysis approach. IEEE Access **9**, 36645–36656 (2021)
7. Heidenreich, T., Lind, F., Eberl, J.M., Boomgaarden, H.G.: Media framing dynamics of the european 'refugee crisis': a comparative topic modelling approach. J. Refugee Studies **32**(Special_Issue_1), i172–i182 (2019)
8. Jacobi, C., Van Atteveldt, W., Welbers, K.: Quantitative analysis of large amounts of journalistic texts using topic modelling. Digit. J. **4**(1), 89–106 (2016)
9. Joseph, K., Carley, K.M.: Frames and their affective dimensions: a case study of climate change news. In: Social, Cultural, and Behavioral Modeling: 15th International Conference, SBP-BRiMS 2022, Pittsburgh, PA, USA, 20–23 September 2022, Proceedings, vol. 13558, p. 57. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-17114-7_6
10. Jozwiak, R.: State report: Russian, Chinese and Iranian disinformation narratives echo one another. Politico (2020)
11. Kim, M.Y., Johnson, K.M.: CLoSE: contrastive learning of subframe embeddings for political bias classification of news media (2022)
12. Lau, J.H., Baldwin, T.: An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint arXiv:1607.05368 (2016)
13. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196. PMLR (2014)
14. Malaga, R.A.: Search engine optimization-black and white hat approaches. In: Advances in Computers, vol. 78, pp. 1–39. Elsevier (2010)
15. Williams, E.M., Carley, K.M.: Search engine manipulation to spread pro-kremlin propaganda. Harvard Kennedy School Misinf. Rev. (2023)
16. Woodruff Swan, B.: Eu monitors see coordinated covid-19 disinformation effort by Iran, Russia, China. rferl.org (2020)
17. Yang, K.C., Menczer, F.: Large language models can rate news outlet credibility. arXiv preprint arXiv:2304.00228 (2023)
18. Zanconato, A., Sabahi, F.: Iran-media landscape. Eur. Journalism Centre (2019)
19. Zandt, D.V.: Media bias fact check: a comprehensive media bias resource (2024). https://mediabiasfactcheck.com/methodology
20. Zhao, W.X., et al.: Comparing twitter and traditional media using topic models. In: Clough, P., et al. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 338–349. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20161-5_34

Analyzing and Predicting Meetup Mobs Outcome Via Statistical Analysis and Deep Learning

Samer Al-khateeb[1](✉), Jack Burright[1], Steven L. Fernandes[1], and Nitin Agarwal[2]

[1] Creighton University, Omaha, NE 68178, USA
{SamerAl-khateeb1,JackBurright,StevenFernandes}@creighton.edu
[2] COSMOS Research Center at UA-Little Rock, Little Rock, AR 72204, USA
nxagarwal@ualr.edu

Abstract. We refer to a "mob" as an event that is organized via social media, email, SMS, or other forms of digital communication technologies in which a group of people (who might have an agenda) get together online or offline to collectively conduct an act and then disperse (quickly or over a long period). To an outsider, such an event may seem arbitrary. However, a sophisticated amount of coordination is involved. Meetup.com is an "Event-Based Social Network" (EBSN) focused on bringing like-minded people together. Meetup hosts a wide range of events, making it crucial and well-suited for studying various events in general and mobs in particular. In this research, we collected data from Meetup and employed statistical analysis to help us better understand the data. Additionally, we utilized a deep neural network-based method to create two classifiers capable of predicting the Meetup mob outcome (success or failure) with great accuracy.

Keywords: Mobs Modeling · Deep Learning · Meetup · Statistical Analysis · Event-Based Social Network

1 Introduction

Meetup.com is one of the most used Event-Based Social Network (EBSN) sites, offering numerous events and extensive data through its API. It allows organizers to plan events related to any topic, ranging from very formal business meetings to casual events, e.g., movie nights [10]. Users can create and join groups based on their interests, and organize events such as mobs [7]. These mobs can be online (in cyberspace) or in person (in physical space). A mob does not have to be deviant (i.e., illegal and involving violence); some mobs are benign and aim to bring fun to a community, e.g., dance flash mobs. This makes Meetup.com a suitable platform to study mobs and mobbers' behaviors. Hence, in this research, we collect data from Meetup.com to help us answer the following research questions: **RQ1**: How can we utilize Descriptive Statistical Analysis (DSA) to comprehend

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 164–173, 2024.
https://doi.org/10.1007/978-3-031-72241-7_16

the collected Meetup data, and what insights can we derive from it? **RQ2**: How can we utilize Predictive Statistical Analysis (PSA) via deep learning to create a mob classifier capable of predicting the mob outcome (success or failure)? **RQ3**: How can we leverage and estimate a set of critical success factors (CSFs) and key performance indicators (KPIs) using event data collected from Meetup.com? Furthermore, how does utilizing these factors impact the performance of the deep learning classifier? To address the aforementioned research questions, we utilize DSA and PSA via deep learning.

2 Literature Review

There is a huge body of research related to the topics addressed in this paper; however, we briefly mention a few due to the paper's length requirements. For example, Schneider et al. [11] aimed to assess the potential of Meetup.com as a platform for promoting physical activity and fostering a sense of community. Huang et al. [7] created a deep learning model for predicting the growth and success of Meetup *groups*. Grundke et al. [3] developed a web crawler to collect data from Meetup.com and then utilized this data to conduct experiments aimed at enhancing the events recommendation page known as the "cold-start" page. Weinberg and Williams [13] pursued two research questions during Howard Dean's 2004 Democratic Party presidential nomination campaign when he used Meetup.com to arrange electronic events that would transition into in-person gatherings (referred to as "e2f" - electronic to face-to-face). Ricken et al. [10] conducted a study focusing on Meetup groups related to software development and aimed at addressing several key questions. Finally, Li et al. [9] focused on predicting the popularity of new groups on Meetup.com. This paper focuses on the method we used to answer our research questions. To the best of our knowledge, this work is the first to provide an estimation method for the CSFs and KPIs using Meetup data and to train a deep-learning model capable of predicting the outcome of Meetup mobs.

3 Methodology

In this section, we explain the methodology we followed to collect data from Meetup.com, the data prepossessing and enrichment steps needed to train our deep learning model, the DSA we conducted to have an overall understanding of the collected Meetup data, and the PSA we undertaken to predict the success or failure of mobs based on various inputs.

Data Collection: using Meetup.com's GraphQL API, we gathered data from 27 distinct Meetup groups, all featuring the topic "Flash Mobs". For each group, we collected data from every mob (event) they had organized. This data encompassed details such as the number of attendees, RSVP times, event description, and all comments and replies associated with the mob. This resulted in $3,536$ mobs with over $18,000$ RSVPs. We stored this data in a MySQL database.

Data Preprocessing and Enrichment: to be able to train our deep neural network, we needed to enrich the data, first, by running two dictionary-based tools, namely: Linguistic Inquiry and Word Count (LIWC) and the extended Moral Foundations Dictionary (eMFD) [6] which is based on the Moral Foundation Theory and Biblical Ethics (the six returned scores are split by vice and virtue) on the mob comments and the mob description. LIWC gives a percentage (from 0–100%) of all words in the text that fit a specific linguistic category, e.g., a *pos_emo* score of 4.5 means 4.5% of the words in the document have positive emotion. Also, the eMFD score is a percentage of all words in a text, but it returns scores that range from 0 to 1. For example, a *care.virtue* score of 0.24 would mean 24% of the words in the document are part of the care virtue words in the dictionary. All of these scores were stored in the same MySQL database.

The second data enrichment step involved calculating 13 more data attributes. These data attributes are called "Critical Success Factors" (CSFs) and "Key Performance Indicators" (KPIs) published by [4]. Table 1 shows the 13 CSFs and KPIs we estimated and used to run the second deep learning model which is explained later.

As a preprocessing step, we removed all events from private groups since we couldn't collect their comments. Moreover, events for which we couldn't calculate the number of individuals invited (using Eq. 1) were also excluded. As a result of these filtration steps, 459 mobs (24 cyber mobs and 435 physical mobs) remained for our analytical examination.

$$\#ofInvitedPeople = \#ofEventOrganizers + [\#ofMaxAllowedTickets * (1 + \#ofAllowedGuests)] \quad (1)$$

After the aforementioned preprocessing steps, two individuals in our lab randomly selected 378 mobs and manually labeled each mob as either "successful" or "failed" based on the following three criteria:

1. Presence of a photo depicting people at the event (participants in the mob).
2. Existence of a comment indicating someone's attendance or intention to attend the event.
3. Achievement of the event's target number of attendees.

If none of these three criteria were met, the mob was labeled as a failed mob. By applying these criteria, we manually labeled 211 mobs as successful and 167 as failed. This manually labeled data will serve as the ground truth for validating the two deep learning models explained below.

Statistical Analysis: *Descriptive Statistical Analysis* (DSA) describes the characteristics of the data (i.e., summarize the data) by measuring its central tendency (such as mean, median, and mode), variability (using variance and standard deviation), and frequency distribution (count). DSA aids analysts in gaining insights into the collected data without needing to examine individual data points. For example, a student's grade point average (GPA) offers valuable insight into their academic performance without the need to inspect each class grade [5]. We used DSA to have an understanding of the different mob types (topics), public vs. private mobs, the most popular locations of the mobbers (countries and cities), cyber vs. physical mobs, participation rate of these mobs, average mob *advertisement* time, average mobber *decision* time, and the correlation between participation rate and average mobber *decision* time.

DSA is good at narrating past events and understanding the data attributes through tables, graphs, and textual explanations. However, it cannot be used to make inferences or predictions about future values. To accomplish this, *Predictive Statistical Analysis* (PSA) is required. PSA is used to predict future values of an attribute, trends, or events using historical data. PSA helps in making strategic decisions and has a wide array of applications in finance, entertainment, marketing, and manufacturing, among others. PSA can be performed manually (a slower and more limited approach per application) or through the use of machine learning and deep learning algorithms (a faster and more versatile method) [2]. Therefore, in this paper, we employed an artificial deep neural network to forecast the success or failure of mobs based on various inputs.

Deep Learning Model-1: we used the *Sequential()* model provided by the Keras library from TensorFlow to train an artificial deep neural network with one input layer, two hidden layers, and one output layer (with 128, 64, 32, and 1 neurons, respectively). We added Batch Normalization to enhance the training speed, stability, and performance of the model. We also used LeakyReLU() activation function to take care of the negative values and a dropout of 0.5 after each layer to avoid overfitting. As an input for this model we used our manually labeled mob data that contained *85* attributes (i.e., the input_size) of information about the mobs. We have split the 378 labeled mobs into 70%:10%:20% for training, validation, and testing and used 5-fold cross-validation. This model was trained for 200 epochs where the data was grouped into batches of size 32.

Deep Learning Model-2: for this model, we used the same settings as model-1 above, however, the input_size for this model was 13 (which are the attributes we estimated using the methods highlighted in Table 1). The same training, validation, and testing splits, k-folds, epochs, and batches as model-1 were used here.

All data collection, preprocessing, enhancement, storage, and deep learning models training were done using a *Mac Pro - Tower* with the following specifications: 3.2 GHz 16-core Intel Xeon W processor, Turbo Boost up to 4.4 GHz Processor, 96 GB (6×16 GB) of DDR4 ECC memory. It also has 4TB SSD storage and the Radeon Pro W5500X with 8 GB of GDDR6 memory.

4 Results and Findings

In this section, we present and discuss our findings, grouping the results from the descriptive statistical analysis and the deep learning models based on the research questions addressed in this paper.

Descriptive Statistical Analysis Findings: to address the first research question, we used DSA to better understand the characteristics of the collected

Table 1. shows the CSF and KPIs we estimated and used to train our deep learning model 2. Columns 1 and 2 are borrowed from [4]. Column 3 explain how we estimated these factors using the collected Meetup data.

CSF	KPIs	Estimation Method using Our Meetup Data
Be Unique	Vividness	According to [4] for each user to be unique, the user has to have Vividness and Entertaining Content. We measure the event Vividness by counting the number of users that have attended events of only 1 group
Be Unique	Entertaining Content	We measure the event Entertaining Content by counting the number of URLs in the comments and replies of the event and the number of photos posted on the event
Interactivity	Interaction Rate	The number of replies to the comments of the event
Interactivity	Num Of Postings	The number of comments on the event
Interactivity	Recurring Rate	The number of mobbers that have attended at least 2 mobs from that group (if they just went to one event then they wouldn't have came back)
Increase Customer Happiness	Num of Positive Mentions	The average of *emo_pos* (positive emotion) from all the comments and replies posted on the event
Creative Ways To Address Users	Num Of Attended Events	This value is the number of unique topics hosted by the users that hosted the event divided by the number of organizers of that event
Address Target Group Consistent	Reach Within Target Group	The number of mobbers that have attended at least 2 mobs from that mob *topic*. For example, if the event topic is Flash Mob, how many mobbers attended events with the topic Flash Mob
Be Active	Net-reach	The size of the group (number of members) that hosted the event
Be Active	Num of Postings	The number of events that the group hosted the event have hosted
Unprofessionalism	Num of Slang Words	This score is calculated using this formula: the *(big word score + clout score - swear score)*. The values used in the formula were calculated using the LIWC of the event description. If this score is positive, it means the event is professional, and if it is negative, it means the event is unprofessional
Building a Reputation	Num of Positive Mentions	The number of comments and replies that have a higher *emo_pos* score than *emo_neg* score
Privacy Protection	–	If the mob is online it gets a 1, if it is in person it gets a 0

data. Initially, we examined the "topics" of the mobs under analysis, which can also be regarded as types of mobs. This analysis helps us understand the nature of the data (the mobs). The 459 mobs we analyzed were organized by 16 different groups, with sizes ranging from 33 members to 9,203 members. These mobs were tagged with 21 different topics such as "Social" (45 mobs), "Outdoors", "Theater", "Fun Times" (each with 29 mobs), etc. Note that more than one topic can be assigned to a single mob. Upon examining these topics, we found that all analyzed mobs were benign (no deviant mobs were included), which is expected considering that the groups that organized these mobs are public.

We also wanted to examine the diversity of the mobs data, so we analyzed the location of the mobbers. We found that mobbers are from different parts of the world, and most of them are located in big cities such as New York, Sydney, and London.

Given that out of the 459 mobs we collected, 345 are physical mobs (in-person) and 24 are cyber mobs (online), we wanted to examine the difference in participation rates to determine whether mobbers participate more in online or in-person mobs. We estimated the participation rate of these mobs using Eq. 2. We found that cyber mobs exhibit a higher average participation rate compared to physical mobs. This trend could be attributed to various factors, including the assumption that cyber mobs entail less risk or that cyber mobs are easier to participate in compared to physical (in-person) mobs.

$$ParticipationRate = \frac{\#ofInvitedPeopleRespondedWithYes}{\#ofInvitedPeople} \tag{2}$$

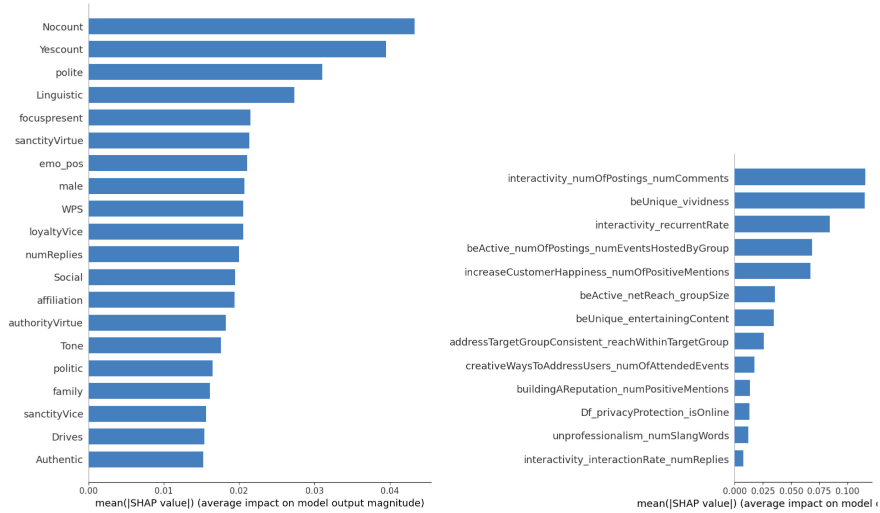
It's also important to understand the lead time mobbers take to advertise their mobs (i.e., the "*recruitment phase*" [1]). Hence, we measured the time difference between the creation of the event on Meetup.com and its scheduled occurrence. By analyzing the time difference we found that mob organizers tend to advertise their mobs well in advance, averaging around 23 days for all the collected mobs. Moreover, we observed that mobs requiring substantial training and effort, such as "Singing Lessons" or "Choir" (ranked top 1 longest time out of 51, averaging $90,230$ minutes which is equivalent to 62.7 days) and "Dance Fitness" (ranked top 6 out of 51), are advertised for significantly longer periods compared to mobs with lower training requirements like "Partying" (46 out of 51), "Music" (50 out of 51), or "Brazilian Culture" which was ranked last with an average advertisement time of $1,634$ minutes, i.e., 1.13 days. It's worth noting that mobs with the "Flash Mobs" topic ranked 15 out of 51, averaging an advertisement time of $45,370$ minutes, i.e., 31.51 days.

Besides considering the advertisement time for mobs, it's also important to examine the duration mobbers take to decide whether they will participate in a mob or not. Therefore, we measured the time difference between the event creation time on Meetup and the moment each mobber responded with either a "Yes" (to attend) or "No". By analyzing the time differences, we found that, on average, individuals invited to participate in a mob take longer to decline (respond with NO) than to accept (respond with YES). The average time taken for a "Yes" response across all mobs is 22,857.16 minutes (equivalent to 15.87 days), whereas the average time taken for a "No" response across all mobs is 33,394.06 minutes (which amounts to 23.2 days).

Finally, we calculated Spearman's correlation coefficient (SCC) to determine the relationship between the 459 Meetup.com mobs participation rate and the average mobbers' time to respond with "Yes". We found a strong, positive monotonic correlation (i.e., a high SCC value) between the mob participation rate (calculated using Eq. 2) and the average mobbers' time to say yes ($SCC = 0.68$, $n = 459$, $p < 0.001$), the average mobbers' time to say no ($SCC = 0.64$, $n = 459$, $p < 0.001$), and the average mobbers' time to respond with either yes or no ($SCC = 0.67$, $n = 459$, $p < 0.001$). Positive monotonic correlation means as the values of one variable increase, the values of the other variable also tend to increase. It doesn't mean that the increase is constant; it only means that higher values of one variable are associated with higher values of the other, even if the relationship is curved or uneven [14].

Predictive Statistical Analysis (via Deep-Learning) Findings: as stated earlier, PSA can be performed manually or through the use of machine learning and deep learning algorithms [2]. Therefore, in this paper, we employed a deep learning algorithm to forecast the success or failure of mobs based on various inputs. Our goal here is to answer *RQ2* and *RQ3*.

Result of Deep Learning Model-1: to address the second research question, we used $Model-1$ described earlier. Using the training, validation, testing data, and 5-fold cross-validation, we found that the mean accuracy for the validation set is 88.64%, and the test set accuracy is 92.11% with a loss of 0.2715. *Accuracy* measures the number of times the model can correctly detect the positive and negative classes. We also calculated the *precision*, which measures "the success probability of making a correct positive class classification" [8], and *recall*, which measures the models ability to minimize false negative scores and were found to be 0.913043 and 0.954545, respectively. This gave a harmonic mean of precision and recall (i.e., F1-score, which "takes into account the type of errors - false positive and false negative - and not just the number of predictions that were incorrect" [12]) of 0.933333. To determine which attribute (out of the 85 attributes used in this model) is most important in determining the success and failure of a mob and to understand the predictions of the machine learning models, we used the SHAP Python library to calculate the SHAP values. SHAP is an additive feature attribution method, meaning the prediction is explained as the sum of the effects of each feature. It is derived from Shapley values, a concept in cooperative game theory developed by Lloyd Shapley. This method provides local and global insights into feature contributions (importance) and model behavior. We found that the number of people responding with "No" or "Yes" and the "Polite" score of the *event description* are the top three most critical attributes in determining success and failure (see Fig. 1-a). Conversely, other linguistic measures calculated from the event description text, such as "Drives", "Authentic", etc., hold less importance but are still in the top 20 most important attributes (out of 85 attributes).



(a) (b)

Fig. 1. Rank of the most important attributes in classifying mobs for both models. (a) shows the average impact of the top 20 attributes of Model-1 while (b) shows the average impact of each of Model-2's attributes.

Result of Deep Learning Model-2: To address the third research question, we estimated the CSFs and KPIs using the estimation methods shown in Table 1. We then used these measures as an input to $Model - 2$ described earlier.

Using the training, validation, testing data, and 5-fold cross-validation, we found that the mean accuracy for the validation set is 93.17%, and the test set accuracy is 98.68% with a loss of 0.0417. We also calculated the precision and recall scores, which were 0.973684 and 1.000000. This gave an F1-score of 0.977778. The higher test accuracies for both models compared to the mean validation accuracies in k-fold cross-validation suggests that the models are not over-fitting. Since Model-2 performed very well and better than Model-1, we used it to predict the label of another set of unseen mobs data (containing 80 physical mobs and 1 online mob). Using Receiver Operating Characteristic Curve (ROC curve) best threshold of 0.336709, the model labeled 54 mobs of the unseen data as successful and 27 as failed mobs. Finally, we examined which CSFs and KPIs, out of the 13 listed in Table 1, are most important in determining the success and failure of a mob. We found the CSF "Interactivity" and the KPIs "Num of Postings", estimated via the number of comments the mob received, to be the most important in determining the success or failure of the mob. Many successful mobs had high scores in this measure while many failed mobs had low scores in this measure, indicating mobs with a lot of interaction from the group have a higher chance of succeeding. Also, the CSF "Be Unique" and KPI "Vividness", is ranked as the second most important factor. Also, many mobs had a high score in this measure were successful. Finally, the CSF "Interactivity" and the KPI "Recurrent Rate", estimated via the number of mobbers that have attended at least 2 mobs from that group, to be the third most important factor in determining the success or failure of the mob. Many successful mobs had high scores in this measure, indicating that the participation of committed mobbers will most likely make a mob succeed. Conversely, the CSF "Interactivity" and the KPI "Interaction Rate" measured via the number of replies seems to be the least important factor in determining the success and failure of a mob. See Fig. 1-b for information about the other measures.

5 Conclusion and Future Research Directions

Meetup differs from other social media sites such as Facebook in how members develop their connections. Users take their offline connections on Facebook and then connect with them online. Meetup is the opposite; users can join groups, connect with people online, and then meet them face-to-face. Another example of a platform that has this same interaction would be the dating site Match.com, where users match online and then can go on dates in person [13]. This makes Meetup a crucial platform to study mob creation and mobbers' behaviors because in a mob, a group of like-minded people, who may or may not know each other, get together online or offline to collectively conduct an act and then disperse. In this paper, we collected data from Meetup.com and conducted two types of statistical analysis: descriptive (DSA) and predictive (PSA). For the PSA, we trained two deep-learning models: one using 85 attributes, while the other used 13 attributes and achieved better performance. Additionally, we ranked the importance of all the attributes used in both models. To the best of our knowledge, this work is the first to provide an estimation method for the CSFs and KPIs using Meetup data and to train a deep-learning model capable of predicting the success or failure of Meetup mobs with high accuracy.

Even though we ran multiple experiments that resulted in the aforementioned models' accuracy, our work is limited by the public Meetup data we collected. The models should be able to predict the success or failures of the mobs with very high accuracy, but only for mobs organized by public Meetup groups. We could not test our model on mobs organized by private groups due to a lack of data and privacy issues. Also, our model is trained on Meetup data, so it might not be able to predict the success or failure of mobs organized on other social media sites such as Facebook or X (formerly known as Twitter).

So for future research direction, we plan to leverage the findings of this research to build an agent-based model to simulate mobs. The simulation model will provide a more generic method to study mobs beyond Meetup mobs.

Acknowledgements. This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-23-1-0011, W911NF-24-1-0078), U.S. Office of Naval Research (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research

Alliance, and the Jerry L. Maulden/Entergy Endowment at the UA-Little Rock. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

1. Al-khateeb, S., Agarwal, N.: Flash mob: a multidisciplinary review. Soc. Netw. Anal. Min. **11**(1), 97 (2021)
2. Cote, C.: What is predictive analytics? 5 examples — HBS online (2021). https://online.hbs.edu/blog/post/predictive-analytics
3. Grundke, M., Jung, J., Sachse, J.P., Wiese, G.: Meetup miner:measuring event interestingness on meetup (2023). https://www.parnswir.de/blobs/smm_meetup.pdf
4. Hammerl, T., Schwaiger, J.M., Leist, S.: Measuring the success of social media: Matching identified success factors to social media kpis. In: Proceedings of the 52nd Annual Hawaii International Conference on System Sciences, Maui, Hawaii, 8–11 January 2019, pp. 2427–2436. ScholarSpace, Honolulu (2019). https://epub.uni-regensburg.de/41362/
5. Hayes, A.: Descriptive statistics: definition, overview, types, example (2023). https://www.investopedia.com/terms/d/descriptive_statistics.asp
6. Hopp, F.R., Fisher, J.T., Cornell, D., Huskey, R., Weber, R.: The extended moral foundations dictionary (emfd): development and applications of a crowd-sourced approach to extracting moral intuitions from text. Behav. Res. Methods **53**, 232–246 (2021)
7. Huang, T., Liang, Y., Sun, Z.: Predicting group success in meetup with graphs (2020). https://snap.stanford.edu/class/cs224w-2019/project/26422233.pdf
8. Jankay, Y.: The 3 pillars of binary classification: accuracy, precision & recall (2018). https://medium.com/@yashwant140393/the-3-pillars-of-binary-classification-accuracy-precision-recall-d2da3d09f664
9. Li, G., Liu, Y., Ribeiro, B., Ding, H.: On group popularity prediction in event-based social networks. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12, no. 1 (2018). https://doi.org/10.1609/icwsm.v12i1.15067. https://ojs.aaai.org/index.php/ICWSM/article/view/15067
10. Ricken, S., Barkhuus, L., Jones, Q.: Going online to meet offline: organizational practices of social activities through meetup. In: Proceedings of the 8th International Conference on Communities and Technologies, pp. 139–148 (2017)
11. Schneider, K.L., et al.: An online social network to increase walking in dog owners: a randomized trial. Med. Sci. Sports Exerc. **47**(3), 631 (2015)
12. Sharma, N.: Understanding and applying f1 score: a deep dive with hands-on coding (2023). https://arize.com/blog-course/f1-score/
13. Weinberg, B.D., Williams, C.B.: The 2004 US presidential campaign: impact of hybrid offline and online 'meetup' communities. J. Direct Data Digit. Mark. Pract. **8**, 46–57 (2006)
14. Weir, I.: Spearman's rank correlation- introduction (2023). https://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf

Drivers of True and False Information Spread: A Causal Study of User Sharing Behaviors

Ling Sun[1,2(✉)], Kathleen M. Carley[2], and Yuan Rao[1]

[1] Xi'an Jiaotong University, Xi'an 710049, Shaanxi, China
sunling@stu.xjtu.edu.cn, raoyuan@mail.xjtu.edu.cn
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA
{lings,carley}@andrew.cmu.edu

Abstract. Analyzing and predicting user information-sharing behavior on online social platforms is a crucial task in social sciences. While current prediction tasks primarily emphasize accuracy, they often neglect the underlying motivations that drive user behavior, hindering a fundamental understanding and control of the information spreading environment. To address this, we analyze and quantify potential factors that may drive user sharing behavior based on social theories. Our limited derived feature set achieves over 85% accuracy in predicting user behavior on two real-world datasets, demonstrating its effectiveness. Notably, through employing causal inference techniques, our analysis on true and false information spread reveals that users with lower authority are more susceptible to being misled by false information. In contrast, the propagation of truthful news is often driven by personal preference or influenced by users' social circles. By uncovering these underlying motivations, our approach facilitates a deeper comprehension of the online information ecosystem, contributing to more effective management strategies for false information mitigation.

Keywords: Behavior Prediction · False Information · Causal Inference

1 Introduction

Online social media enables individuals to obtain information in a cheap and handy way, yet it also promotes the spread of misinformation. Predicting and analyzing users' information sharing behavior, in environments where true and false information coexist, is a crucial task in the field of information governance.

Current user behavior prediction work primarily focuses on utilizing machine learning or deep learning methods to improve prediction accuracy, but largely overlooks the underlying driving factors behind the behavior, resulting in low interpretability and credibility of the results, making it difficult to apply in real-world scenarios. Furthermore, current research on misinformation dissemination mainly focuses on comparing the spread patterns of true and false information

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 174–183, 2024.
https://doi.org/10.1007/978-3-031-72241-7_17

from a macro perspective. They have found that false information tends to spread faster, deeper, and more broadly than real information [16,18], and pointed out there are differences in novelty, topic distribution, and sentiment distribution between true and false information. However, their analyses are relatively independent, overlooking the interplay between features, and the research primarily concentrates on positive samples, i.e., users who participated in sharing, without combining and contrasting with negative samples, i.e., users who were exposed to the news but did not share it, making it difficult to uncover the true driving factors behind users' behavior. Due to the numerous and interrelated factors determining user behavior, accurately predicting user behavior and analyzing the underlying reasons is a challenging task.

Given the limitations of current research, we integrate the discovery of motivations behind information spreading with the task of predicting user behavior. From both theoretical and practical perspectives, we provide a more detailed explanation on how various factors drive user behavior. Specifically, our approach draws upon social theories to identify and extract the most crucial driving factors from complex social data, and achieves high accuracy in behavior prediction through a very limited feature set. Moreover, instead of relying on feature importance rankings from the prediction task, we introduce causal graphs to describe the interactions between features, through cross-analysis of bot accounts, human accounts, and the propagation of true and false information, we unveil the key underlying reasons genuinely influencing user behavior. In summary, this paper:

- Identifies and quantifies the potential factors influencing whether users share a news post based on reliable social theories, and validates the effectiveness of the derived features through user behavior prediction task.
- Constructs a valid causal graph to intuitively illustrate the relationships between various factors, and uncovers the motivations of users in the propagation of fake and real news through cross causal analysis and comparison.

2 Related Works

2.1 User Sharing Behavior Prediction

The user sharing behavior prediction task aims to predict whether a user will share a specific news posts based on relevant features. Researchers have proposed various machine learning-based prediction methods that integrate multimodal data and different types of features. Zhang et al. [17] developed an attention-based convolutional neural network based on the content being shared. Firdaus et al. [7]comprehensively modeled users' past tweets and sharing behavior, analyzing interest, sentiment, and personality traits of users to predict the likelihood of sharing. Sun et al. [14] employed sequential hypergraph neural networks and attention mechanisms to model time-varying user preference and predict the next infected user in information propagation.

However, the aforementioned methods primarily aim to improve prediction accuracy, making it challenging to interpret the results from a social science perspective, thereby limiting their practical application. Recently, Sun et al. [15] designed a causal-enriched deep attention (CEDA) framework to evaluate the causal effects of input variables on retweet behaviors during prediction, improving the interpretability of model.

2.2 Information Propagation Analysis

Many researchers have attempted to measure and analyze the prevalence of false information on social media. Vosoughi et al. [16] found that falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all fields. Zhou et al. [18] concluded four patterns for false information propagation, namely, More-Spreader, Further-Distance, Stronger-Engagement, and Denser-Network.

Several studies further delved into investigating the underlying reasons behind the viral propagation of false information. Based on social theories, the factors that motivate users to spread information can be reflected in four aspects: 1. News attributes, such as news topic, sentiment, source, etc. News topics influence the writing style of posts and the sharing tendency of users, and often invoke emotional responses [10]. Therefore, news with specific attributes may attract more attention from users; 2. User attributes, such as gender, age, number of friends, activity level, authority level, etc. For instance, Altay et al. [1] found that users with more friends want to keep a positive self-image of themselves, so they share less fake news; 3. User interest, users typically follow what they like [4], and prefer information that confirms their preexisting attitudes [6]. Therefore, user interest is a key determinant of information sharing, yet it is influenced by numerous factors, the interests expressed in user posts may not reflect genuine preferences, but rather stem from the echo chamber effect, which limits exposure to diverse information [4]; 4. Social influence, social identity theory reveals that users tend to conform to the viewpoints prevalent within their community to gain acceptance and achieve a sense of belonging [3]. Gimpel et al. [9] found that fake news shared by users is often shared by their trusted family and friends.

Cheng et al. [4] classified user sharing behavior into intentional and unintentional to identify suspicious users. Bui et al. [3]further analyzed the influence of factors such as social identity and news polarity on sharing intentions. In this study, we provide a more detailed explanation on how various factors drive propagation by integrating user behavior prediction with causal inference technology.

3 Analysis and Calculation of Sharing Driving Factors

Our experiments and analysis based on the publicly available fake news detection dataset, FakeNewsNet [13], which comprises news data linked to two fact-checking platforms: GossipCop[1] and PolitiFact[2]. The PolitiFact dataset predominantly addresses political topics, whereas GossipCop focuses primarily on entertainment news. These datasets encompass comprehensive information, including

[1] https://www.gossipcop.com/.
[2] https://www.politifact.com/.

news text, sources, user profiles and their historical behaviors. Based on social theories, we categorize potential factors influencing user sharing behavior into four categories, and utilize additional computations and tools to enhance more valuable features, details are provided in Table 1.

News Attributes. Textual features of news posts, such as topics and sentiments, have been shown to be closely related to their diffusion effects [3,16]. Meanwhile, the source website and the number of engagements indicate the credibility and popularity of news, which are also valuable. Thus, we identify news sentiments from the content using the Google Cloud Natural Language API[3], then obtain the source ratings from the Web of Trust API[4]. The total number of tweets, retweets, and comments for each news are used to measure its popularity.

Table 1. Features included, calculation methods/tools used, and range of values.

	Metrics	Description	Formulation/Tool	Scale
News Attributes	Source	The source website for publishing the news	–	–
	Source score	Security score of the website	WOT API	[0,1]
	Popularity	# tweets, retweets and replies	#tweets+#retweets +#replies	–
	Sentiment	Sentiment of news content	Google Cloud Natural Language API	[–1,1]
	Topic	Topic of news content	Google Cloud Natural Language API	–
User Attributes	Basic attributes	'user_id', 'created_at', '#favorites', '#friends', '#listed', '#followers', '#statuses', 'verified'	–	–
	Activity	User's activity level	Norm(0.5 *(#statuses/(time - created_at))+ 0.5*((#favourites + #friends)/2))	[0,1]
	Authority	User's authority level	Norm((#followers +#listed)/(1+#friends)* (1.2 if verified else 0.8))	[0,1]
	Bot score	Probability of being a bot	BotHunter	[0,1]
	Negativity	Proportion of posts with negative sentiments	Vader API	[0,1]
	Emotional	Proportion of posts with strong sentiments	Vader API	[0,1]
User Interest	Interest score	User's interest in news	SimCSE	[0,1]
Social Influence	Neighbor influence	Number and influence of users' neighbors who shared the news before	–	

User Attributes. Users' behavior is significantly influenced by their selection bias, which are closely related to their attributes. Research [4] found that users' verification status, status and friend count are associated with the probability of being suspicious. Therefore, we utilize the user profiles as basic attributes, and calculate the activity and authority scores based on user behavioral data. Since

[3] https://cloud.google.com/natural-language.
[4] https://www.mywot.com/.

numerous bot accounts exist on social platforms, we employ the BotHunter [2] to estimate the probability of an account being a bot. Furthermore, as emotional arousal is a crucial factor driving information sharing [10], we leverage the Vader API [11] to identify the sentiment of each post published by users, and calculate the proportion of negative sentiment (¡ 0) posts as the user's negativity score and the proportion of extreme sentiment (¿ 0.5 or ¡ –0.5) posts as their emotional score.

User Interest. User interests are influenced by various factors, such as the interests of their friends and the biases of recommendation algorithms [4]. Therefore, we treat interest scores as independent from user attributes and calculate them separately. Specifically, since typical language models like BERT [5] are unsuitable for computing similarity between short texts, we employ the SimCSE, an improved method based on contrastive learning [8]. To reduce computational costs, we concatenate all tweets of a user into a single document and design a sliding window, computing the similarity between each window and the target news text, and obtain the final interest score by averaging these similarity scores.

Social Influence. User behavior is easily influenced by their friends or family [3]. However, networks on social platforms are very large and sparse. To retain the most valuable influences, we construct a directed user network based on their historical behaviors, where an edge exists only between users who have directly or indirectly shared each other's posts, and calculate the social influence exerted on a user by quantifying the influence of their neighbors.

4 User Sharing Behavior Prediction

4.1 Data Sampling and Experimental Settings

Based on the constructed user interaction network, we sample negative instances from the neighbors of known positive instances at a 1:1 ratio. To ensure that the negative instances had the potential to encounter the news, we only sample from nodes that have previously received information from the positive instances. To compare the behavioral differences between human and bot accounts, we classified accounts with a bot score greater than 0.6 as bot accounts. Conversely, accounts with a bot score less than 0.4 were defined as human users. The final statistics is shown in the Table 2. We split the data into training and test sets with a 7:3 ratio and conducted baseline experiments using multiple machine learning classifiers, including Support Vector Machines (SVM), Logistic Regression (LR), Random Forests (RF), and Decision Trees (DT). The results demonstrated that RF achieved the best predictive performance. Consequently, all experiments and analyses presented in this work are based on the RF classifier.

Table 2. Data statistics

Dataset	News	Positive samples	Negtive samples	Bots	Human
Politifact-fake	341	276,748	231,011	162,846	121,432
Politifact-real	240	311,923	250,733	163,276	172,500
Gossipcop-fake	3430	965,641	826,866	466,798	613,531
Gossipcop-real	6903	812,788	698,528	641,471	452,280

4.2 Feature Importance Analysis

As shown in Table 3, the experimental results indicate that fully utilizing all features always leads to the highest accuracy, highlighting the importance of integrating diverse features. Besides, user attributes performs better than other features in predicting users' tendencies to share both real and fake news, particularly in Gossipcop dataset, where the model attains an accuracy of 85.90%. This approves that user behavior is significantly influenced by their own preferences. Social influence also plays a pivotal role, especially for fake news in Gossipcop dataset. This could be because entertainment news is less contentious than political news, so people rely more on social engagements and are easily deceived by fake news with specific characteristics, rather than out of interest.

Table 3. Prediction results with different features.

Features	Politifact						Gossipcop					
	All		Fake		Real		All		Fake		Real	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
All	**86.44**	**86.74**	**86.65**	**86.83**	**86.26**	**86.64**	**89.97**	**90.31**	**88.89**	**89.41**	**91.42**	**91.56**
News atts.	54.90	70.77	54.30	70.28	55.31	71.18	53.83	69.00	59.74	73.12	51.39	65.58
User atts.	77.82	78.08	77.16	77.18	76.30	77.34	85.90	86.52	83.03	84.02	88.06	88.26
Interest	58.11	59.28	59.04	58.26	57.77	62.00	57.48	59.13	58.43	60.76	57.46	56.84
Social inf.	62.86	67.49	62.49	65.98	63.30	68.81	60.78	72.07	62.33	72.47	58.87	70.86

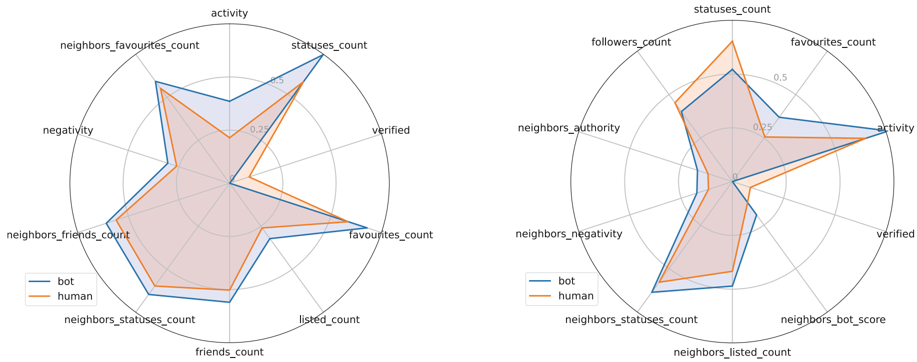
4.3 Bots and Human Behavior Analysis

Given the differing behavior patterns and driving factors between bot accounts and human accounts, we extracted potential bot and human accounts from the dataset and analyze the distinctions between them. We first extract the ten most distinguishing features and visualize them in Fig. 1. Notably, across both datasets, human accounts exhibit a higher proportion of being verified, while bot accounts display more active.

Additionally, we conducted separate predictions, results are presented in Table 4. It can be observed that the prediction accuracy for bot accounts is higher than for human accounts across both datasets. Notably, the accuracy for bot accounts is about 1.6% and 5% higher than human in the Politact and Gossipcap datasets, respectively, and the F1 score for real news in Gossipcop is even 27.7% higher than that for human accounts. This indicates that while bot accounts may mimic human-like features, they exhibit different behavioral pattern, which is simpler and more predictable.

Table 4. Prediction results with bots and human accounts.

Identity	Politifact						Gossipcop					
	All		Fake		Real		All		Fake		Real	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Bots	**82.01**	**74.45**	**81.71**	**74.14**	**82.25**	**75.60**	**89.56**	**89.96**	**84.59**	**80.16**	**93.18**	**94.53**
Human	80.46	70.44	80.37	65.28	80.62	73.25	84.57	70.06	83.34	71.68	86.36	67.33



(a) Politifact (b) Gossipcop

Fig. 1. Visualization of the key features of bots and human accounts

5 Motivations Discovery

Given that classifiers can only identify the correlation between features and behavior, rather than causal relationships, and often neglect the interactions between features, we constructed a causal graph based on social science theories, and utilized causal intervention strategy to calculate the causal effect of each feature on user behavior, and ultimately proved that the driving factors exhibit significant differences across various dissemination scenarios.

5.1 Causal Graph Construction

Based on social science theories, we first construct a causal graph for four factors and the results: user behavior, as illustrated in the Fig. 2. As we illustrated in Sect. 3, all four factors can directly influence user behavior. Additionally, based on the echo chamber effect, user attributes and social influence may affect user interest. For instance, users are more likely to encounter information shared by those around them and content recommended by algorithms based on their attributes [1,4]. Furthermore, the social influence a user experiences is simultaneously affected by both user attributes and news attributes, that is, users with more friends generally experience greater social influence, and users may share highly popular news based on social conformity theory [3].

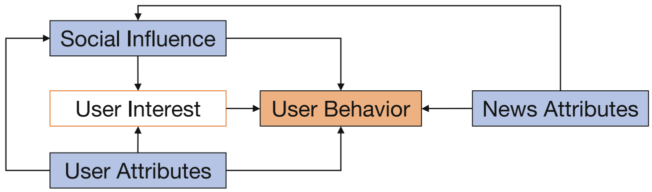
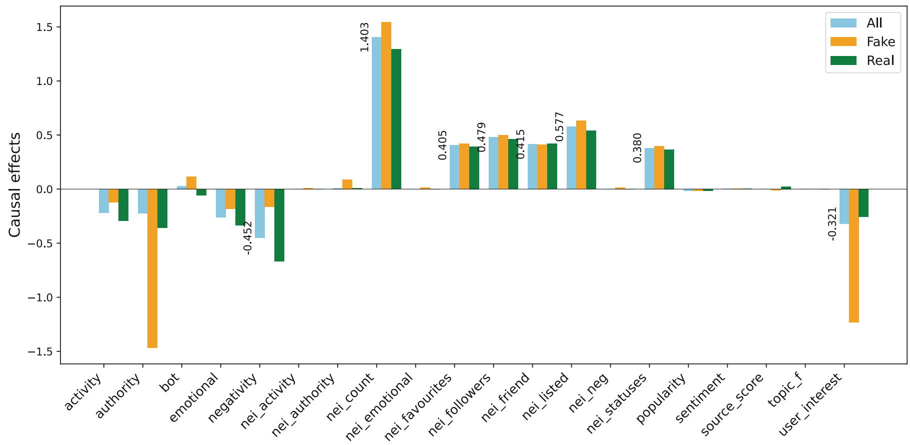
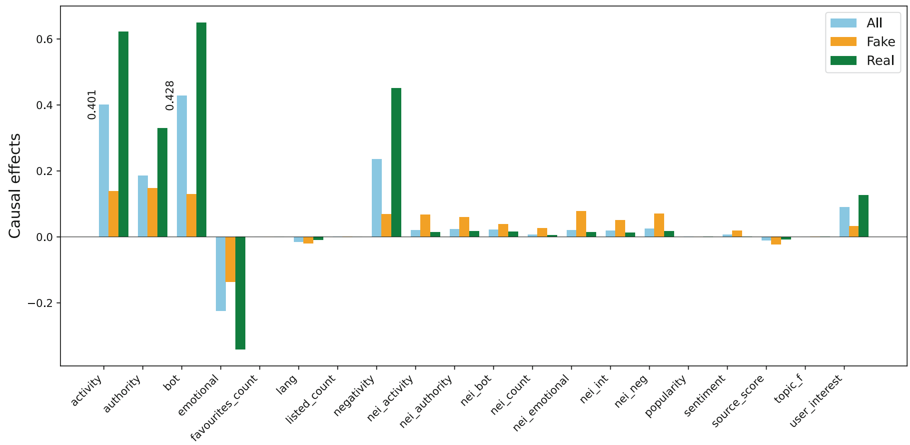


Fig. 2. The causal relationship between four types of features and user behavior



(a) Politifact



(b) Gossipcop

Fig. 3. The causal effects of features on user behavior (top 20 significant)

5.2 Causal Calculation and Motivation Discovery

Based on the causal graph, we aim to estimate the causal effect of each treatment (feature) on outcome(user behavior) using the DoWhy [12] tool. Given a Treatment (X), and an Outcome (Y), the estimated effect can be calculated by intervening the value of X:

$$\mathbb{E}[Y|do(X = x^{'})] - \mathbb{E}[Y|do(X = x)] \tag{1}$$

Specifically, for a given feature X (e.g., "User Interest"), we first utilize the backdoor criterion to identify potential confounders, which are variables that simultaneously affect both the treatment variable and the outcome variable "User Behavior". According to the backdoor criterion, we need to control for the variables "News Attributes" and "Social Influence" to block backdoor paths and eliminate confounding effects. Subsequently, we estimate the causal effect using Linear Regression (LR) due to its high computational efficiency and ease of interpretation. The coefficient of the LR directly represents the causal effects of the treatment on the outcome.

We computed the causal effects of all features on user behavior for both datasets, as illustrated in Fig. 3. For the PolitiFact dataset, it can be observed that the authority of users has the most significant impact on the dissemination of fake news, showing a pronounced negative correlation –1.48. This indicates that users with lower authority are more likely to be deceived by fake news. In contrast, for the dissemination of real news, the number of neighbors who have shared the news plays the most crucial positive role (1.32), suggesting that user behavior is largely driven by social conformity.

The causal distribution in Gossipcop differs significantly from PolitiFact. User attributes such as activity, authority, bot score, and negativity all exhibit relatively strong positive correlations with sharing behavior, while the influence from neighbors is comparatively lower. This indicates that for entertainment news, user behavior is more influenced by personal characteristics and interests. Moreover, we observe instances where bot accounts actively engage in retweeting truthful news, likely as a strategy to enhance their authority. These findings demonstrate that news propagation drivers vary across different topics, suggesting that diverse measures may be needed in information control.

6 Discussion and Conclusion

To better understand the driving mechanisms behind information propagation and facilitate control of the online information environment, we analyze and quantify the underlying reasons for user information sharing behavior from multiple aspects, grounded in social theories. Through user behavior prediction experiments and causal analysis, we demonstrate the effectiveness of our extracted features. We find that although bot accounts mimic human features, their behavioral patterns are still distinguishable and more predictable. Moreover, the veracity and topic of information lead to different distributions of driving factors, suggesting that distinct strategies should be employed in various scenarios.

References

1. Altay, S., Hacquin, A., Mercier, H.: Why do so few people share fake news? it hurts their reputation. New Media Soc. **24**(6), 1303–1324 (2022)
2. Beskow, D.M., Carley, K.M.: Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In: SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, vol. 3 (2018)
3. Bui, Q.N., Moriuchi, E.: Sharing intention of politicized news on social media: mediators and moderators. In: 57th Hawaii International Conference on System Sciences, HICSS 2024, Hilton Hawaiian Village Waikiki Beach Resort, Hawaii, USA, 3–6 January 2024, pp. 6086–6095 (2024)
4. Cheng, L., Guo, R., Shu, K., Liu, H.: Causal understanding of fake news dissemination on social media. In: KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, 14–18 August 2021, pp. 148–157 (2021)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019, vol. 1, pp. 4171–4186 (2019)
6. Dmj, L., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018)
7. Firdaus, S.N., Ding, C., Sadeghian, A.: Retweet prediction based on topic, emotion and personality. Online Soc. Netw. Media **25**, 100165 (2021)
8. Gao, T., Yao, X., Chen, D.: Simcse: simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual, 7–11 November 2021, pp. 6894–6910 (2021)
9. Gimpel, H., Heger, S., Olenberger, C., Utz, L.: The effectiveness of social norms in fighting fake news on social media. J. Manag. Inf. Syst. **38**(1), 196–221 (2021)
10. Horner, C.G., Galletta, D.F., Crawford, J., Shirsat, A.: Emotions: the unexplored fuel of fake news on social media. J. Manag. Inf. Syst. **38**(4), 1039–1066 (2021)
11. Hutto, C.J., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: The Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, 1–4 June 2014 (2014)
12. Sharma, A., Kiciman, E.: Dowhy: an end-to-end library for causal inference. CoRR arxiv:2011.04216 (2020)
13. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data **8**(3), 171–188 (2020)
14. Sun, L., Rao, Y., Zhang, X., Lan, Y.: MS-HGAT: memory-enhanced sequential hypergraph attention network for information diffusion prediction. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, pp. 4156–4164 (2022)
15. Sun, W., Liu, X.F.: Deep attention framework for retweet prediction enriched with causal inferences. Appl. Intell. **53**(20), 24293–24313 (2023)
16. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
17. Zhang, Q., Gong, Y., Wu, J., Huang, H., Huang, X.: Retweet prediction with attention-based deep neural network. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, 24–28 October 2016, pp. 75–84 (2016)
18. Zhou, X., Zafarani, R.: Network-based fake news detection: a pattern-driven approach. SIGKDD Explor. **21**(2), 48–60 (2019)

Examining Socio-Economic Isolation in Urban Spaces

Hend Alrasheed[1,2](✉), Alina Fonseca Flores[3], Vinicius Andrade Brei[2,3], and Alex Paul Pentland[2]

[1] Department of Information Technology, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
halrasheed@ksu.edu.sa

[2] Institute for Data, Systems, and Society, MIT Connection Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
{hrasheed,pentland}@mit.edu

[3] School of Management, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil
{alina.flores,brei}@ufrgs.br

Abstract. Urbanization offers the potential to reduce socio-economic disparities by fostering interaction among diverse population groups. However, its success is often hindered by factors leading to socio-economic segregation. Utilizing anonymized mobility data, this study investigates socio-economic isolation in Salvador, Brazil, with a focus on gender and age isolation. We analyze isolation at both the census and individual levels, measuring isolation among gender and age groups in public spaces and examining interactions among individuals from different social groups. Our results highlight significant gender and age disparities at the census-level, with social class having a major influence at the individual-level.

Keywords: Socio-economic isolation · Mobility patterns · Social network analysis · Group exposure · CDR data

1 Introduction

Urbanization presents an opportunity to reduce socio-economic disparities by enabling access to public amenities, while also encouraging interaction among diverse population groups. However, its effectiveness in addressing these disparities is significantly shaped by factors such as income and race. Within urban environments, a notable challenge arises in the form of isolation, as groups become spatially segregated based on their socio-economic status [1]. This isolation significantly affects their access to essential resources and opportunities [2].

Traditional census data often fall short in capturing the nuances of socio-economic isolation, partly due to their inability to monitor individual movements to 'third places'-spaces neither homes nor workplaces, yet essential for

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 184–194, 2024.
https://doi.org/10.1007/978-3-031-72241-7_18

understanding daily social interactions [3]. Furthermore, the influence of urban mobility on socio-economic dynamics has led researchers to increasingly rely on mobility data to delve deeper into patterns of isolation [4–7]. Prior research has primarily focused on economic isolation, with other forms of isolation receiving comparatively less attention. This discrepancy is largely attributed to the limited availability of detailed data on other demographic attributes, primarily due to concerns surrounding privacy. However, both gender [5,8–10] and age [2,11–13] have been identified as influential factors shaping mobility dynamics.

In this work, we investigate socio-economic isolation in Salvador, Brazil, emphasizing the interplay of gender, age, and social class. Using mobility traces from Call Detail Records (CDRs) of a large cohort of anonymized mobile phone users during October and December 2023, we conduct both census-level and individual-level analyses through statistical measurements and social network techniques. At the census-level, we quantify gender and age-group isolation in public spaces. At the individual-level, we assess the probability of exposure between individuals from different socio-economic groups.

Our preliminary findings reveal considerable levels of isolation, with gender and age-groups exhibiting pronounced effects at the census-level, while social class plays a substantial role at the individual-level. Specifically, women experience higher levels of isolation compared to men, with more tracts (56%) male-dominated and 30% female-dominated. Among age-groups, both youth and seniors are more prone to isolation. Additionally, individuals from higher social classes also experience higher levels of isolation. These findings have several implications for urban planning and policy making. The observed socio-economic isolation underscores the need for targeted interventions to promote integration and reduce disparities.

2 Related Work

Prior research has explored the connections between gender, age, income, and mobility patterns, revealing distinct trends across demographics [2,4,6,10,11]. In Santiago, Chile, studies using mobility datasets found that women generally visit fewer unique locations than men and distribute their time less evenly among these locations, a pattern attributed to urban affordances [5]. However, contrasting findings from travel surveys across 13 countries indicated that women are more inclined to walk and use public transport, with these tendencies varying by age and increasing gender disparities as age increases [10]. Research on age-related mobility shows that both younger and older age-groups often face greater socio-economic isolation, influenced by affordability issues [2,11–13].

While numerous studies have examined socio-economic influences on mobility, the impact of these choices on isolation within specific areas or Points of Interest (POIs) remains underexplored. Income-based segregation within the US has been analyzed using statistical indices [4] and social network analysis [6], highlighting substantial variations in income isolation even among closely situated POIs and the significant role of local business sectors in this segregation.

There remains a lack of comprehensive analysis integrating multiple socio-economic dimensions to understand isolation within specific areas. In this work, we aim to bridge this gap by examining socio-economic isolation based on multiple factors, including gender, age, and social class.

3 Method

We aim to evaluate the isolation level between various socio-economic groups by analyzing their visits within census tracts. We start by examining group isolation at the census-level. Subsequently, we utilize a network-based model to assess isolation at the individual-level.

3.1 Datasets

Mobility Data. We utilized two months of anonymized, privacy-enhanced human mobility data from October and December 2023 for the city of Salvador, Brazil, provided by the mobility data analytics company Logan. This dataset comprises census-level aggregated records detailing visits and dwell times across socio-economic groups. Each record includes demographic information across three categories: gender (Female, Male), age group (18–24, 25–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80+), and social class (A, B, C, D, E), generating 80 unique demographic profiles per census tract. Brazil uses social class ranks to categorize households based on income and living standards, with Class A being the wealthiest and Class E the poorest. The dataset spans 4,548 census tracts, with an average of 3,274.8 visits per tract and 837.4 unique visitors. To maintain data integrity, we omitted records with fewer than two visitors or with dwell times shorter than five minutes or longer than 24 h. Additionally, the dataset is structured to provide the count of visitors for each time block throughout the week, with each day segmented into four time blocks: 6:00–9:59, 10:00–13:59, 14:00–17:59, and 18:00–21:59.

Point of Interest (POI) Data. POI data for each census tract in Salvador were sourced from the Brazil Federal Government's Open Data (dadosabertos.rfb.gov.br/CNPJ/). This dataset comprises 275,685 POIs, with each entry containing details such as the name, coordinates, and sector. The coordinates provided for each POI correspond to the centroid of its respective census tract, rather than precise locations. Using the POI dataset, we can analyze the type of business in each tract. On average, each tract contains about 65.2 POIs. More details about the datasets are available at [14].

3.2 Census-Level Analysis of Socio-Economic Isolation

This section measures socio-economic isolation across various groups by examining their mobility patterns within each census tract. While a more granular analysis would typically be conducted at the Point of Interest (POI) level, we utilize census-level data from our dataset for this examination.

We use adapted versions of the Income Segregation of Places measure proposed by [4]. Each measure is designed to quantify deviations from an ideal state of integration, where the total time spent at a census tract's POIs is uniformly distributed across all groups. We first introduce isolation based on a single demographic attribute, such as gender, then expand to encompass two attributes.

The gender-based isolation measure (I_c^g) is defined as $I_c^g = |T_{fc} - \frac{1}{2}| + |T_{mc} - \frac{1}{2}|$, where T_{fc} and T_{mc} represent the proportion of total time spent at census tract c by female (f) and male (m) visitors, respectively. The value of I_c^g ranges from 0 to 1. A tract with $I_c^g = 0$ indicates a tract visited equally by both genders representing ideal integration. Conversely, a tract with $I_c^g = 1$ indicates that the tract is visited exclusively by a single gender. Generally, higher I_c^g values indicate a tract predominantly visited by one gender, reflecting higher levels of gender isolation. Note that a value of $I_c^g = 0.5$ suggests a significantly skewed gender composition, with up to 75% from one gender. Whereas $I_c^g = 0.3$ indicates a moderate level of gender isolation, with one gender representing 65% of visitors.

Age-group isolation within each census tract is quantified as $I_c^a = \frac{4}{7}\sum_j |T_{jc} - \frac{1}{8}|$, where T_{jc} is the proportion of total time spent at c by each of the eight age-groups. A census tract visited by five age-groups will have $I_c^a = 0.2$, indicating moderate isolation, while one visited by three age-groups will have $I_c^a = 0.7$, indicating high isolation.

We then measure isolation based on gender and age-group within each census tract c (denoted I_c^{ga}). The isolation index I_c^{ga} quantifies the deviation from ideal integration, which is achieved when the total time spent within a census tract is distributed evenly across all groups. Here, each group is represented as a pair (g, a), where g corresponds to a specific gender and a represents an age-group.

$$I_c^{ga} = \frac{8}{15}\sum_i \sum_j |T_{ijc} - \frac{1}{16}|, \tag{1}$$

where T_{ijc} is the proportion of total time spent at census tract c by each group of gender i and age-group j. Isolation is considered moderate to high when $I_c^{ga} \geq 0.3$, indicating that a quarter of the groups might not visit the tract. We excluded consideration of social class to maintain result interpretability.

3.3 Individual-Level Analysis of Socio-Economic Isolation

We utilize a social network model to assess the likelihood of exposure among individuals from different socio-economic groups. In our model, the probability of forming connections between groups is determined by the number of visitors from each group during each time block and the density of POIs in the census tract. A higher density of POIs spreads potential interactions over a wider area, reducing the chance of individuals from different groups encountering each other at the same location. For this analysis, we utilize the smallest available time

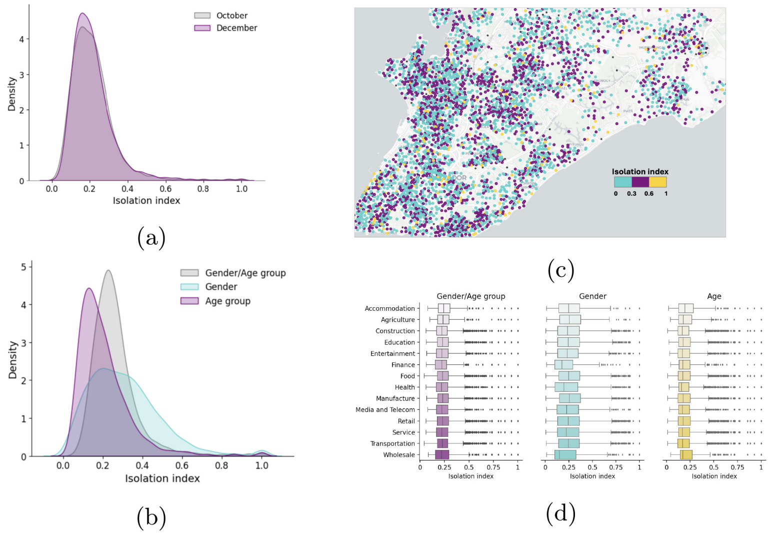


Fig. 1. Census-level isolation in Salvador. (a) Distribution of gender/age-group isolation for October and December. (b) Distribution of the isolation index across all census tracts. (c) Gender-based isolation for each census tract, depicted by circles at their centroid coordinates. (d) Distribution of isolation by POI sectors

interval in the dataset, which consists of time blocks (Sect. 2.1). We then use network centrality metrics to serve as indicators of group isolation.

We construct a weighted, undirected exposure network $G_c = (V_c, E_c, W_c)$ for each census tract c, where each node represents a distinct group defined by unique socio-economic attributes (gender, age-group, and social class). Let v_i and v_j denote the nodes for groups i and j, respectively. An edge forms between v_i and v_j with probability $P_{ij}(t)$, defined as: $P_{ij}(t) = \frac{m_i^t \cdot m_j^t}{n_c}$, where m_i^t and m_j^t are the numbers of visitors from groups i and j during time t, and n_c represents the number of POIs in the census tract c. The edge weights within the network capture the cumulative probabilities of interactions. I.e., $w_{ij} = \sum_k P_{ij}(t_k)$, where k is the number of time blocks in the dataset.

To gain insights into the overall potential for exposure between groups across census tracts, we aggregate the individual tract-specific networks into a Global Exposure network $G = (V, E, W)$. This network retains the same set of nodes as those in the individual tract networks and it calculates the average weights of edges between corresponding nodes across these networks. This aggregation allows us to apply network-based metrics to evaluate the roles of these nodes. Here we use two centrality metrics: degree centrality and closeness centrality.

Degree centrality measures the direct ties of a node with others in the network. In weighted networks, the degree of a node is equal to the sum of the weights of its incident edges. In our Global Expose network, this metric quantifies the level of exposure a demographic group has with other groups. Nodes with higher weighted degree centrality are indicative of groups that experience more frequent exposure to others, reducing their isolation.

Closeness centrality indicates a node's proximity to all other nodes in the network, representing the speed with which it can interact with them. In our Global Exposure network, nodes with higher closeness centrality hold more central positions, facilitating direct exposure or easy establishment of connections with other groups. Conversely, nodes with lower closeness centrality are more peripheral, indicating fewer or less direct connections within the network. We utilize weighted closeness centrality, considering edge weights in our analysis.

4 Results

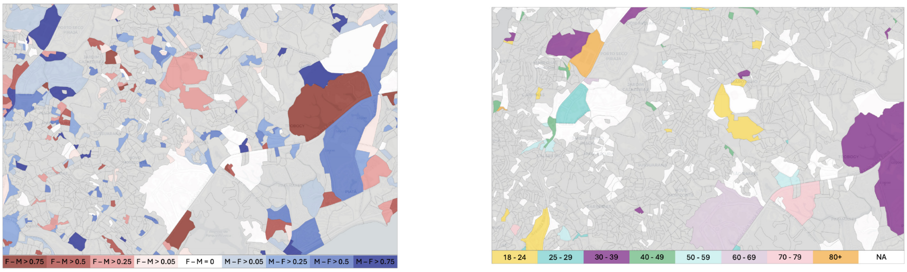
4.1 Census-Level Analysis of Socio-Economic Isolation

Census-level group isolation by gender and age-group was quantified using equation (1) for each observed month, October and December. Trends in gender and age-group isolation were found to be consistent across both months (Fig. 1(a)). Figure 1(b) illustrates the distribution of gender and age-group isolation, as well as the isolation based solely on gender and age-group across all census tracts, selecting the maximum observed values from the two months. Gender isolation is shown in the map in Fig. 1(c). Overall, approximately 24% of tracts exhibit moderate to high isolation based on gender and age-group, while 46% show similar levels of isolation based solely on gender, and 16% based solely on age-group. Notably, the patterns of isolation manifest differently between gender and age-groups, with gender displaying a more pronounced influence on the levels of isolation.

To investigate the relationship between POI sectors and group isolation, we associate the isolation index of each census tract with its respective POIs. This involves linking the isolation index of each tract to the POIs it contains, although it is important to note that the actual distribution of visits across these POIs is not available in our dataset. Figure 1(d) illustrates the distribution of isolation indices by POI sectors in each census tract, differentiated by gender, age, and a combination of both gender and age. Notably, within the Education, Food, and Health sectors, over 20% of POIs display moderate to high segregation based on gender and age-groups, with more than 30% exhibiting such segregation based solely on gender.

The left panel of Fig. 1(d) reveals slight variability in isolation indices within each sector. Sectors such as Education, Entertainment, and Retail exhibit notable outliers where the isolation index is significantly higher than the sector's median, suggesting that certain gender/age-groups are particularly isolated. In contrast, sectors like Finance, Health, and Manufacturing demonstrate lower median isolation indices, indicating a more balanced representation among visitors based on gender and age-groups within these sectors.

The middle and right panels of Fig. 1(d) suggest that gender differences contribute more significantly to isolation than age. In contrast, the distribution of



(a) (b)

Fig. 2. Maps illustrating group dominance in Salvador's census tracts. (a) Gender dominance. (b) Age-group dominance, where dominance is defined by an age-group comprising at least 5% of tract visitors

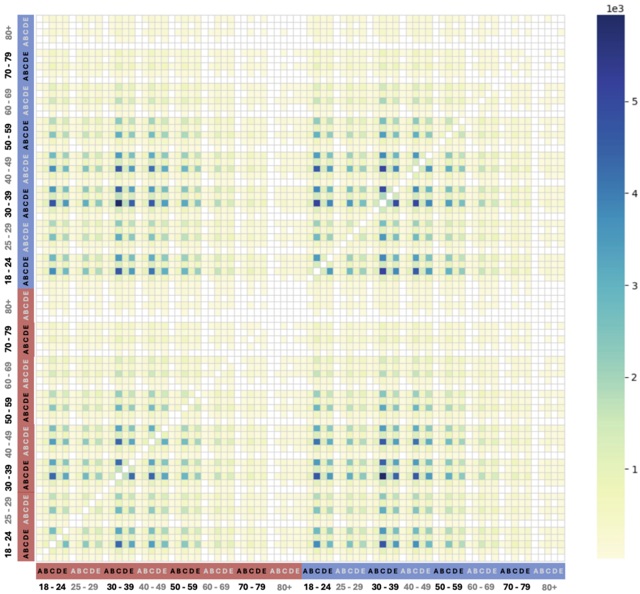


Fig. 3. The cumulative probability of two individuals from different groups visiting the same POI within a census tract

isolation indices by age-group appears more uniform across sectors, suggesting that age alone may not be a significant factor in isolation within these sectors.

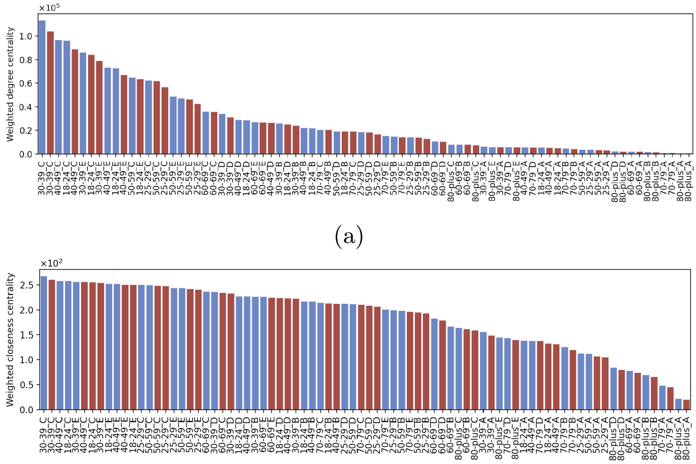
We further explore census tracts that exhibit moderate to high gender and age-group isolation (isolation index $I_c^{ga} \geq 0.3$. Figure 2 illustrates the predominant gender and age-groups, respectively. In Fig. 2(a), there is a notable trend of male dominance over female dominance across the majority of tracts, with 56% male-dominated, 14% neutral, and 30% female-dominated. Conversely, Fig. 2(b) indicates a relatively uniform distribution of dominance among different age-groups, with no single age-group consistently prevailing.

4.2 Individual-Level Analysis of Socio-Economic Isolation

To explore individual-level socio-economic isolation, we construct an exposure network for each census tract using the mobility dataset (Sect. 2.3). Each census tract network comprises 80 nodes representing all possible socio-economic combinations. Edge weights indicate cumulative exposure probabilities between individuals from different groups. Tract-specific networks were then aggregated to create the Global Exposure network. Figure 3 illustrates accumulative exposure probabilities between groups for the month of October in the Global Exposure network.

The bottom-left and top-right corners of Fig. 3 (same-gender exposures) exhibit similar patterns, suggesting that gender does not significantly influence the probability of individual exposure within POIs. However, male-to-male exposure is slightly higher across most age-groups and social classes. Figure 3 also shows some degree of social class isolation, indicated by the darker squares. For instance, social classes C and E show higher exposure probabilities both within their groups and with each other. Conversely, social classes A and B exhibit lower exposure probabilities with all other groups. Overall, Fig. 3 indicates that younger (18–24) and older (70–79, 80+) age-groups have distinct patterns, showing less exposure to other age-groups compared to middle-aged groups.

Figure 4 illustrates each group's weighted degree and closeness centralities in the Global Exposure network. In Fig. 4(a), it is evident that the highest weighted degree centrality values are predominantly associated with male groups, particularly those in the young to mid-age ranges, and with groups from social classes C and E. This suggests that these groups experience higher exposure to other demographic groups. It appears that social class has a more significant influence on exposure than gender or age-group. Groups from higher social classes (A and B) and older age-groups tend to show more isolation. Similar patterns are observed in Fig. 4(b). The weighted closeness centrality highlights that social class and age-group play a more substantial role in determining how central groups are within the exposure network. The figure indicates a more uniform distribution of how quickly different gender groups can reach the rest of the network. This suggests that gender has a less pronounced impact on centrality in this context.



(a)

(b)

Fig. 4. Weighted degree (a) and closeness (b) centralities for each demographic group in the Global Exposure network. Red bars represent female groups, while blue bars represent male groups. (Color figure online)

5 Discussion

Urban planning should actively promote interaction among diverse socio-economic groups, addressing the imperative need to identify factors contributing to community isolation. Recent empirical studies shed light on gender disparities in mobility dynamics, delineating distinct behavioral patterns between men and women [5,9]. In Brazil, mobility trends are influenced by multifaceted factors including gender, age, and income levels [15–17].

This study evaluates the extent of isolation among various socio-economic cohorts within Salvador, a significant Brazilian urban center. Our analysis reveals divergent levels of isolation, with gender and age-groups exhibiting pronounced effects at the census-level, while individual-level social class plays a significant role. Our findings indicate that women experience higher levels of isolation compared to men, and among age-groups, both the youth and seniors, along with individuals from higher social classes, are more prone to isolation. Notably, there's evident social segregation, particularly among higher-class individuals.

An enduring inquiry surrounds the origins of these isolation trends, whether stemming from socio-economic constraints such as affordability, access to transportation, and public safety, or from personal preferences leading individuals with similar socio-economic backgrounds to gravitate towards homogeneous environments. Similar to findings by [18], we observe that the middle class experiences relatively lower levels of isolation, while individuals from lower classes tend to interact more broadly across socio-economic strata. These observations persist irrespective of the characteristics of the neighborhood.

Isolation within certain sectors, such as Education, among specific age-groups may be attributed to the sector's focus on particular demographics. Conversely, isolation in sectors like Construction and Health might signify a bias in job opportunities favoring specific gender and age-groups.

Our study has several limitations. First, the aggregation of the dataset at the census-level limits our capacity to distinguish subtle isolation patterns, including those pertaining to Points of Interest. Second, the dataset lacks granularity to distinguish between residents and visitors within each census tract, crucial for understanding factors such as spatial dynamics. Additionally, our network model assumes a linear scaling of interaction probability with the number of visitors from each group, potentially oversimplifying real-world dynamics.

6 Conclusion

Through our analysis employing mobility data and social network techniques, we have unveiled significant disparities and complexities within urban socio-economic dynamics in Salvador. This study illustrates how various demographic groups experience differential levels of isolation, with notable influences of gender and age at the census tract level, while social class plays a pivotal role at the individual-level. The insights gained are critical for informing urban planning and policy making efforts aimed at mitigating social isolation. Although our analysis was based on aggregated data-necessitated by the dataset's structure-we believe it reflects the underlying trends that are likely to persist in more granular analyses. Future research should aim to expand our analysis to incorporate additional factors such as employment status and educational attainment to provide a deeper understanding of the dynamics of isolation. Longitudinal studies are needed to explore how these patterns evolve over time, offering insights into the temporal stability of our findings. Additionally, a detailed examination of the role of Points of Interest in relation to isolation will enhance our comprehension of spatial factors in urban social dynamics. Finally, further research is needed to assess whether our findings can be generalized beyond the city of Salvador.

Acknowledgement. We thank Logan for providing the mobility data for Salvador.

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. Tammaru, T., Marcińczak, S., Van Ham, M., Musterd, S.: Socio-Economic Segregation in European Capital Cities. Routledge, Abingdon (2014)
2. Abramson, A.J., Tobin, M.S., VanderGoot, M.R.: The changing geography of metropolitan opportunity: the segregation of the poor in US metropolitan areas, 1970 to 1990. Hous. Policy Debate **6**(1), 45–72 (1995)
3. Oldenburg, R., Brissett, D.: The third place. Qual. Sociol. **5**(4), 265–284 (1982)

4. Moro, E., Calacci, D., Dong, X., Pentland, A.: Mobility patterns are associated with experienced income segregation in large US cities. Nat. Commun. **12**(1), 4633 (2021)
5. Gauvin, L., et al.: Gender gaps in urban mobility. Human. Social Sci. Commun. **7**(1), 1–13 (2020)
6. Nilforoshan, H., et al.: Human mobility networks reveal increased segregation in large cities. Nature **624**(7992), 586–592 (2023)
7. Liao, Y., Gil, J., Yeh, S., Pereira, R.H., Alessandretti, L.: Socio-spatial segregation and human mobility: a review of empirical evidence. arXiv preprint arXiv:2403.06641 (2024)
8. Goel, R., Sharma, R., Aasa, A.: Understanding gender segregation through call data records: an Estonian case study. PLoS ONE **16**(3), e0248212 (2021)
9. Reisch, T., Heiler, G., Hurt, J., Klimek, P., Hanbury, A., Thurner, S.: Behavioral gender differences are reinforced during the COVID-19 crisis. Sci. Rep. **11**(1), 19241 (2021)
10. Goel, R., et al.: Gender differences in active travel in major cities across the world. Transportation, 1–17 (2022)
11. Browning, C.R., Calder, C.A., Boettner, B.: Socioeconomic segregation of activity spaces in urban neighborhoods: does shared residence mean shared routines? RSF J. Social Sci. **3**(2), 210–231 (2017)
12. Cook, C., Currier, L., Glaeser, E.L.: Urban Mobility and the Experienced Isolation of Students (No. w29645). National Bureau of Economic Research (2022)
13. Couture, V.: Urban youth most isolated in largest cities (2024)
14. https://github.com/halrashe/socio-economic_isolation
15. Duryea, S., de Freitas, L.B., Trevisan, G.: Universities and intergenerational social mobility in Brazil: examining patterns by race and gender. J. Econ. Race Policy **2**, 240–256 (2019)
16. da Cruz, G.F., Pero, V.: Gender differences in intergenerational income mobility in Brazil. In: Advances in Women's Empowerment: Critical Insight from Asia, Africa and Latin America, vol. 29, pp. 65–93. Emerald Publishing Limited (2020)
17. Moreira, G.C., Ceccato, V.A.: Gendered mobility and violence in the São Paulo metro. Brazil. Urban Stud. **58**(1), 203–222 (2021)
18. Macedo, M., Lotero, L., Cardillo, A., Menezes, R., Barbosa, H.: Differences in the spatial landscape of urban mobility: gender and socioeconomic perspectives. PLoS ONE **17**(3), e0260874 (2022)

Ablation Studies in Protest Networks: The Role of Influential Agents in Shaping Protests

Sayantan Bhattacharya(✉), Nitin Agarwal, and Diwash Poudel

COSMOS Research Center, University of Arkansas - Little Rock, Little Rock, USA
{sbhattachary,nxagarwal,dpoudel}@ualr.edu

Abstract. This study investigates how different content modalities, like images and text, contribute to mobilizing support and driving engagement in collective action campaigns that thrive on multimedia representation propelled by social media. Focusing on the 2022 Brazilian insurrection, it examines how escalating outrage and shifting sentiments manifested online during societal flashpoints. Grounding the analysis in Resource Mobilization Theory and Moral Foundations Theory, it explores how texts and images contributed to collective identity formation, understanding emotion and morality's role in mobilizing audiences. A socio-technical approach analyzed emotion and moral dimensions in Instagram posts by analyzing protest networks to identify influential agents. Findings demonstrated text and image captions playing crucial roles in collective identity formation and mobilization, with morally charged responses of rage and victimization driven by influential agents circulating reinforcing multimedia.

Keywords: Influential Agent · Ablation · Collective Action

1 Introduction

Images and videos enhance the persuasive power of collective action campaigns. This study examines how visual media and text contribute to mobilizing support and driving campaign engagement. Collective action communication thrives on multimedia representation—a blend of images, videos, captions, and other symbolic content—propelled by the power of social media, given the potential influence of multimedia content in shaping collective narratives and mobilizing [19]. Quantitative analysis is needed to trace the fluctuations and appeals to emotion and morality within images and texts.

Recognizing this need, the present study examines how the escalation of outrage and shifting emotion manifest online during significant societal flashpoints, such as the recent insurrection in Brazil [4,10]. Emerging research indicates a significant shift towards multimedia-centric online communication, establishing it at the forefront of communication research. Computational methods can pave the way for new approaches to understanding the psychological and social processes

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 195–204, 2024.
https://doi.org/10.1007/978-3-031-72241-7_19

evoked by multimedia data. Emotionally resonant content, like images showing protests and acts of violence, can influence behavior and encourage collective actions. To investigate the influence of text and images on mobilization, we explore how they contribute to collective identity formation during the protests that followed the 2022 Brazilian presidential elections, drawing from the principles of Resource Mobilization Theory (RMT) and Moral Foundations Theory (MFT) to understand the role of emotion and morality in mobilizing audiences.

The Brazil anti-government protest campaign unfolded following the announcement of the Brazilian presidential election runoff results on October 30, 2022, when supporters of the then-president Jair Messias Bolsonaro cited electoral fraud. This incident led to a surge of polarizing content on social media, with nationalists calling for a military coup. The subsequent nationwide demonstrations culminated in riots on January 8, 2023, involving around 4,000 pro-Bolsonaro supporters who stormed multiple congress buildings [15]. Social media platforms began sharing images of alleged riot participants, utilizing different hashtags to crowdsource information and aiding authorities in identifying and punishing those who evaded arrest.

We employed a socio-technical approach, applying computational techniques to analyze Instagram posts' emotional and moral dimensions by comparing text and imagery. Throughout the observation period, we identified shifts in emotion and morality, unveiling the effects driven by influential agents who circulate multimedia that reinforce their respective ideas. The findings demonstrate that text and image captions played an essential role in collective identity formation and mobilization, though they contributed differently. Notably, the response was morally charged, with prominent emotions of rage and victimization.

Using this process, our research examines the following research questions:

RQ1: What is the role of content modality in the collective action process?
RQ2: What is the role of influential agents on the collective action process in terms of evoking emotions through different content modalities?
RQ3: What is the role of influential agents on the collective action process during mobilization through different content modalities?

2 Literature Review

Existing literature often relies on interviews and surveys to study collective action and identity formation. However, there is a need to explore how computational methods examine the role of visual and textual content in shaping online collective identities and mobilizing users.

2.1 Collective Action

Lance Bennett introduced the concept of connective action, which describes a type of collective action in which individuals use online social networks to form informal, locally organized communities around a protest or social movement without the need for explicit collective identity framing or formal coordination [3]. Previous studies have explored collective identity from various perspectives. Koller proposes an approach to analyzing collective identity in discourse by examining linguistic and semiotic features and their socio-cognitive interpretation, suggesting that discourse practices, social context, and dominant ideologies influence the construction of collective identities [13]. Others investigate how members of opinion-based groups contribute to collective identity formation and collective action on social media, proposing a model of collective group/social identity for collective action [5]. In contrast, Kann et al. (2023) investigate the role of collective identity in protest participation during the 2020 BLM protests, finding that while individuals demonstrate increased interest in the movement, there is an insignificant effect on subsequent collective identity [12].

2.2 Resource Mobilization Theory

Further examining the theoretical underpinning, Resource Mobilization Theory (RMT) suggests that the success of social movements depends on available resources, including moral resources like legitimacy, solidarity, and sympathetic support [8]. This study focuses on examining the moral resources associated with the protest movements that followed the 2022 Brazilian presidential elections, adopting the Moral Foundations Theory (MFT) framework to analyze virtue-based and vice-based moral foundations present in the online discourse [9]. A fundamental principle of RMT is that mobilizing moral resources, such as shared commitment and consensus, is crucial for the emergence and sustainability of protests [20]. Previous qualitative approaches have examined moral discourse to identify shared grievances and collective identities [2]. Computational methods like moral foundations analysis enable quantitative evaluation of moral appeals in text, as Dehghani et al. (2016) demonstrated in the study of homophily in social networks with tweets [7]. However, this study is the first to examine collective identity and mobilization on Instagram and multimedia networks, utilizing a computational approach to analyze influential agents and their role in mobilization efforts.

2.3 Multimedia in Social Movements

Digital performances on social media platforms are recognized as modular forms of protest with high potential for virality, facilitating rapid diffusion of inspiration and ideas across social networks. Sharing personal photographs is a deliberate strategy to capture attention and draw new participants into a movement, with curating and presenting content in digital spaces bolstering commitment and engagement [18].

Recent literature explores social media's transformative role in shaping social movements' agendas and aiding collective action online and offline [1]. It highlights the shift to digital media and increased public engagement, with social network multimedia reporting contributing to organizational routines [3]. Studies examine multimedia adoption in different phases of social movements, influenced

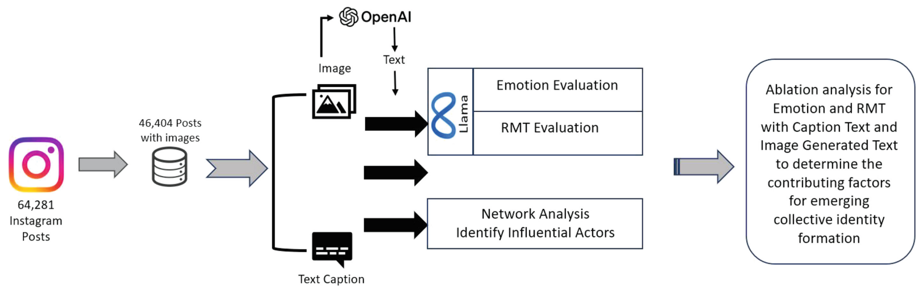


Fig. 1. Core Steps for identifying Influential Agents and factors for Collective Action

by innovation diffusion and social network dynamics [17]. Research also investigates online media's role in mobilizing large-scale collective action, analyzing hierarchy and closure's impact on mobilization [6] (Fig. 1).

3 Data Collection

The APIFY data extraction tool collected data based on seed hashtags. The data collection process encompassed the time frame from November 1, 2022, to January 31, 2023. The commencement of the 'Brazil anti-government protests' arose subsequent to allegations of electoral fraud during the Brazilian presidential elections. The supporters of Jair Messias Bolsonaro demanded the nullification of the election outcomes and disseminated divisive content through various social media platforms, leading to outbreaks of riots on January 8, 2023. The demonstrators who stood in support of Bolsonaro called for military intervention and employed hashtags such as #Brazilwasstolen, #brazilelectionfraud, #brazilianspring(s), #sosffaa, and #SOSbrasil [15]. These hashtags were very focused and specific for this research. The total number of Instagram posts collected was 64,281, and only publically accessible data was considered. Following the preprocessing stage for the communication network, our dataset comprised 46,404 image posts with textual captions. To handle any non-English posts, we utilized the Google Translate API for translation purposes.

4 Methodology

The methodology employed in this research comprises a multi-faceted approach to investigate the role of influential agents in shaping collective identity formation and mobilization during the protests following the 2022 Brazilian presidential elections.

4.1 Diffusion of Innovation Theory

We divided our weekly dataset into three distinct phases to study the impact and adoption rate of the Brazil insurrection campaign. By categorizing the data consistent with the Diffusion of Innovation theory, we observed the time at which different stages occurred, the rate of adoption by participants within the movement. The first four weeks, known as the Initiation Phase, identify the onset of the protests and the early adopters. The following four weeks, the Amplification phase, mark a period of rapid growth and increased adoption rate. The final phase is the Saturation Phase, during which growth stabilizes and eventually cease [18].

During the Initiation phase, a group of highly driven and passionate individuals, known as early supporters or pioneers, instigated the protest movement. For the Brazil protests, this phase began in late October and November 2022, after Lula's narrow presidential election victory over Bolsonaro. Bolsonaro's supporters reacted with initial acts of civil disobedience, including appeals for military intervention, construction of road blockades by truckers, rallies calling for military involvement, and Bolsonaro's legal challenge against the election results.

As the protest became more popular, it entered a stage of growth that captured the attention of a larger audience. This period, spanning from mid-December 2022 to January 2023, witnessed a surge in the scale, intensity, and radicalism of the pro-Bolsonaro movement. The discord between authorities and Bolsonaro supporters escalated due to their rejection of the election results. This conflict peaked on January 8th with the breach of government premises in Brasilia, disrupting the transition following Bolsonaro's refusal to acknowledge the election outcome. Leading up to this climax were incidents like an attempted breach at police headquarters on December 12th, the arrest of an individual trying to detonate explosives on December 24th, arrests linked to coup plots towards late December, and Bolsonaro's departure for Florida on December 30th, leaving his fervent supporters unsettled. These occurrences underscored a rise in extremism, organization, and desperation among fringe Bolsonaro factions that culminated in violent outbursts in the capital on January 8th [21].

During the Saturation Phase, the protest movement peaked in terms of involvement and impact. Most people were exposed to the protest message, and additional expansion leveled off as the movement gained social acceptance. In this case, no significant events were observed after the Amplification phase, leading to the assumption that the intensity of the protest had reduced, and the situation gradually returned to normalcy.

The phase classification of the protest into Initiation, Amplification, and Saturation provides a temporal framework for identifying influential agents and analyzing their role in shaping collective emotion and identity formation. By examining network interactions and content during each phase, we can observe the most influential individuals driving the narrative. Comparing emotional and moral content across phases reveals shifts in strategies as the movement progresses. This phase-based approach contextualizes the analysis within the protest's trajectory, offering insights into the mechanisms through which influential agents leverage emotional and moral appeals to mobilize sustained collective action over time.

4.2 Network Analysis for Detection of Influential Agents

After phase classification, next step involved conducting a network analysis to generate a user communication network based on the interactions and mentions between users on Instagram. This network was the foundation for identifying influential agents within the protest movement.

We focused on top community and employed several network metrics to determine influential agents, including degree, betweenness, and closeness centralities. These measures provided insights into an individual's influence and connectivity within the network, allowing us to identify the most prominent, well-connected users. By isolating the data specific to these influential agents, we analyzed their contributions and impact on the protest network, aligning with the principles of ablation theory. We found out two users, blog_do_almeida (Degree: 88, Betweenness Centrality: 1251.0, Closeness Centrailty: 1.0) and gayscomlucidez ((Degree: 87, Betweenness Centrality: 767.0, Closeness Centrailty: 0.85)) having the highest centrality measures that made them influential agents for the protest network. The analysis of influential agents' data focused on understanding their role in evoking collective emotion and mobilizing moral resources, as Resource Mobilization Theory (RMT) outlined. Leveraging computational techniques, we examined the emotional dimensions expressed in the textual and visual content shared by these agents.

4.3 Use of LLM

We intended to use the Llama model for all analyses, but due to the lack of image-to-text conversion in Llama 8b Instruct, we used GPT-4 for that task. As the first part of modality analysis, we performed an image-to-text analysis using an OpenAI GPT-4 model. In this case, we relied on the basis of the benchmark testing and model validations with other peer-reviewed research. In 2023, OpenAI published research evaluating the performance and reliability testing of GPT-4 models that showed that GPT-4 outperforms existing large language models for a collection of NLP tasks [14,16]. In this case, we used the prompt "Describe the content of the following image". After that, we employed Llama 3 models to evaluate emotional assessment and determine the morality foundation for Resource Mobilization Theory. We fed all the text (caption text, image-generated text) to the model and got the emotion percentages for anger, fear, disgust, joy, and sadness. In the case of emotional analysis, we used the prompt, "Analyze the following text and identify the emotions expressed in terms of anger, joy, sadness, disgust, and fear. Provide the result in percentage".

For calculating RMT analysis with morality foundation, we used the prompt "Analyze the following text and identify the percentage of elements reflecting Sympathetic Support, Solidarity, and Legitimacy according to Resource Mobilization Theory".

The need for specialized expertise at each analysis stage further justified the decision to utilize two separate models. While GPT-4 excels in the visual-to-verbal domain, Llama 3 8b Instruct offers superior performance in emotion detection and classification [11]. This bifurcated approach ensures that each aspect of

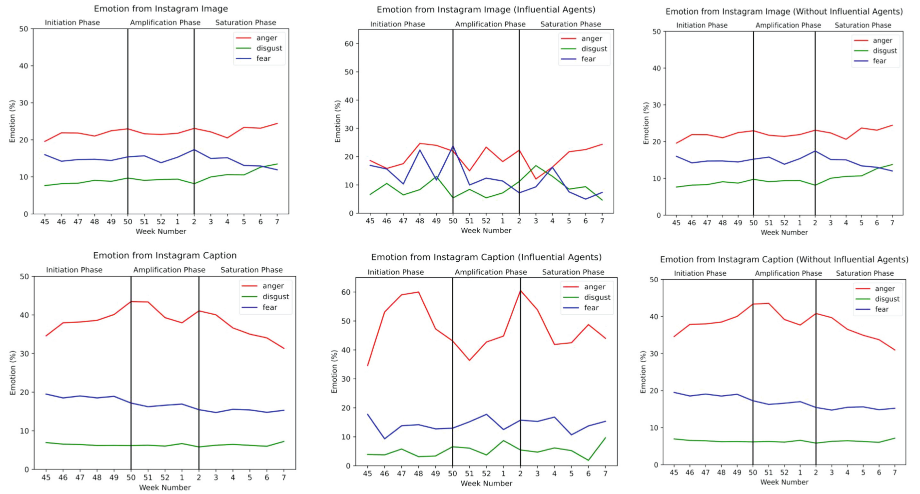


Fig. 2. Ablation for Emotion Analysis with whole data (left most), with influential agent (middle), without influential agent (right most)

the analysis benefits from the most capable model, thereby enhancing the overall accuracy and reliability of our findings. Moreover, the use of different models aligns with the principle of methodological triangulation, which advocates for the application of multiple methods to cross-validate results. By employing two distinct models, we not only capitalize on their individual strengths but also reinforce the validity of our analysis through the convergence of independent methodologies.

5 Result and Discussion

Addressing the first research question (RQ1) on the role of content modality in the collective action process, the analysis revealed that images and captions played a crucial role, contributing differently to shaping emotions and mobilization efforts. Instagram images effectively conveyed emotions like anger, fear, and disgust, with anger consistently being the highest-expressed emotion among all users (Fig. 2). As the protest progressed, there was a gradual increase in anger and fear in images, suggesting heightened emotional responses as the movement gained momentum. Meanwhile, Instagram captions effectively expressed heightened emotions, particularly anger, with noticeable peaks during the Amplification phase (Fig. 2). This indicates that captions played a vital role in mobilizing support by tapping into intense emotional expressions.

Regarding the second research question (RQ2) on the role of influential agents in evoking emotions through different content modalities, the findings demonstrated their significant impact on the emotional dynamics within the protest network, specifically in amplifying emotional responses through images and captions. For images, posts from influential agents exhibited a more volatile pattern

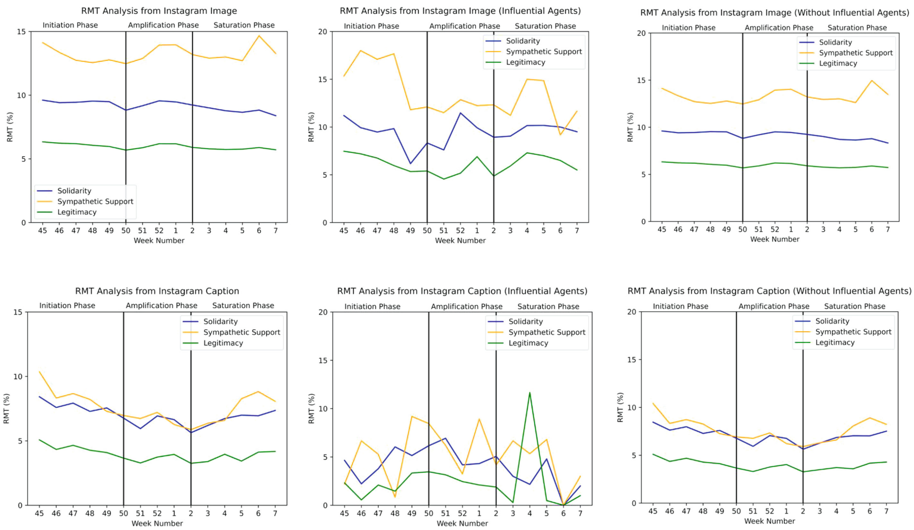


Fig. 3. Ablation for RMT Analysis with whole data (left most), with influential agent (middle), without influential agent (right most)

in anger, with noticeable peaks during the Amplification Phase (Fig. 2). These spikes in emotional reactions highlight the influential agents' ability to drive emotional intensity, potentially fueling increased participation and engagement. Similarly, when examining captions, influential agents showed a pronounced increase in anger, particularly during the Amplification phase, with higher overall levels than non-influential agents (Fig. 2). This pattern demonstrates that influential agents played a crucial role in intensifying emotional engagement and solidarity within the protest network through their textual content.

Turning to the third research question (RQ3), which explored the role of influential agents in evolving mobilization through different content modalities, the analysis revealed their pivotal role in shaping the collective narrative and mobilizing support through their emotional and moral appeals conveyed via images and captions. The prominent peaks in anger, especially among influential agents, suggest their crucial role in converging emotional responses and fostering a sense of unity and shared purpose among participants. Influential agents enhanced the collective identity by amplifying emotional responses like anger, making the cause more compelling and urgent (Fig. 3). While general participants contributed to the baseline emotional tone, influential agents created the necessary emotional spikes through their content, galvanizing collective action. This finding underscores the critical role of influential agents in driving mobilization efforts and reinforcing the movement's collective identity through their strategic use of visual and textual content modalities.

6 Conclusion

Understanding the role of influential agents in evoking collective emotion and mobilizing moral resources like solidarity, sympathetic support, and shared grievances is crucial. These elements are key to collective identity formation and sustained collective action, as per RMT. By evaluating the emotional and moral appeals conveyed through their content, we uncovered strategies employed by these agents to mobilize support and shape the collective narrative surrounding the protests. Throughout this process, we maintained a comparative lens, contrasting the presence and absence of influential agents to assess their impact on the overall discourse, emotional landscape, and mobilization efforts. This approach provided insights into how visual and textual content facilitated by influential agents drives collective identity formation and collective action during significant societal events. The methodologies presented herein can effectively contribute to future predictive models, enabling analysts and policymakers to effectively monitor the emerging strength of a collective entity. Consequently, these findings augment our understanding of social and collective identity formation, while also complementing the conventional survey-based techniques commonly employed in this domain. Furthermore, our approach offers a distinctive perspective within the field of computational social science.

Acknowledgments. This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189, W911NF-23-1-0011, W911NF-24-1-0078), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

1. Agarwal, N.: Developing socio-computational approaches to mitigate socio-cognitive security threats in a multi-platform multimedia-rich information environment. In: NATO Science & Technology Organization (STO) Symposium on Mitigating and Responding to Cognitive Warfare (STO-MP-HFM-361), Madrid, Spain, 13–14 November 2023 (2023). ISBN: 978-92-837-2521-3. https://doi.org/10.14339/STO-MP-HFM-361
2. Benford, R.D., Snow, D.A.: Framing processes and social movements: an overview and assessment. Ann. Rev. Sociol. **26**, 611–639 (2000)

3. Bennett, W.L., Segerberg, A.: Digital media and the personalization of collective action: social technology and the organization of protests against the global economic crisis. Inf. Commun. Soc. **14**(6), 770–799 (2011)
4. Bernroider, E.W.N., Harindranath, G., Kamel, S.: From connective actions in social movements to offline collective actions: an individual level perspective. Inf. Technol. People **35**(8), 205–230 (2022)
5. Brünker, F., Deitelhoff, F., Mirbabaie, M.: Collective identity formation on Instagram - investigating the social movement fridays for future. In: Australasian Conference on Information Systems (2019)
6. Chen, Z., Oh, P., Chen, A.: The role of online media in mobilizing large-scale collective action. Soc. Media+ Soc. **7**(3) (2021)
7. Dehghani, M., et al.: Purity homophily in social networks. J. Exp. Psychol. Gen. **145**(3), 366 (2016)
8. Edwards, B., McCarthy, J.: Resources and social movement mobilization. In: Snow, D.A., Soule, S.A., Kriesi, H. (eds.) The Blackwell Companion to Social Movements, pp. 116–152. Wiley-Blackwell (2004)
9. Graham, J., et al.: Moral foundations theory: the pragmatic validity of moral pluralism. In: Advances in Experimental Social Psychology, vol. 47, pp. 55–130. Academic Press (2013)
10. Grover, T., Mark, G.: Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, pp. 193–204 (2019)
11. Huang, W., et al.: How good are low-bit quantized LLaMA3 models? An empirical study (2024)
12. Kann, C., Hashash, S., Steinert-Threlkeld, Z., Alvarez, R.M.: Collective identity in collective action: evidence from the 2020 summer BLM protests. Front. Polit. Sci. **5**, 1185633 (2023)
13. Koller, V.: How to analyse collective identity in discourse - textual and contextual parameters. Crit. Approaches Discourse Anal. Across Disciplines **5**(2), 19–38 (2012)
14. Lian, Z., et al.: GPT-4V with emotion: a zero-shot benchmark for generalized emotion recognition. Inf. Fusion **108**, 102367 (2024)
15. Nicas, J., Spigariol, A.: Bolsonaro supporters lay siege to Brazil's capital. The New York Times (2023)
16. GPT-4 technical report. OpenAI (2023)
17. Shaik, M., Cakmak, M., Spann, B., Agarwal, N.: Characterizing multimedia adoption and its role on mobilization in social. In: Proceedings of the 57th Hawaii International Conference on System Sciences (2024)
18. Soule, S.A., Roggeband, C.: Diffusion processes within and across movements. In: The Wiley Blackwell Companion to Social Movements (2018)
19. Spears, R., Postmes, T.: Group identity, social influence and collective action online: extensions and applications of the SIDE model, pp. 23–46 (2013)
20. van Stekelenburg, J., Klandermans, B.: The social psychology of protest. Eur. J. Soc. Psychol. **61**(5-6) (2013)
21. Supporters of Brazil's Bolsonaro engage in post-election unrest. Reuters (2023). https://www.reuters.com/world/americas/supporters-brazils-bolsonaro-engage-post-election-unrest-2023-01-09/

Unveiling Bias in YouTube Shorts: Analyzing Thumbnail Recommendations and Topic Dynamics

Mert Can Cakmak(✉), Nitin Agarwal, Selimhan Dagtas, and Diwash Poudel

COSMOS Research Center, University of Arkansas - Little Rock, Little Rock, AR, USA
{mccakmak,nxagarwal,sedagtas,dpoudel}@ualr.edu

Abstract. In the dynamic landscape of digital media, YouTube Shorts have emerged as a popular format, captivating users with their brief and engaging content. However, the recommendation algorithms driving these videos often exhibit biases that influence which thumbnails are prominently displayed. This study delves into the biases present in YouTube's recommendation algorithms, focusing on the thumbnails of YouTube Shorts, which play a crucial role in attracting viewers. Thumbnails, as powerful visual elements, significantly impact user decisions and engagement. By utilizing advanced topic modeling and content generation techniques, we analyzed a substantial dataset of YouTube Shorts' thumbnails. Our analysis, employing generative AI and BERTopic models, reveals notable shifts in topic distribution across recommendation cycles, it highlights a preference for certain types of content. These biases not only affect content visibility but also steer user engagement towards popular, yet potentially less diverse, topics. The findings of this study enhance the understanding of algorithmic biases in digital platforms and aim to promote more equitable and transparent content recommendation practices.

Keywords: YouTube Shorts · Thumbnails · Recommender System · Algorithmic Bias · Generative AI · Topic Modelling · Topic Clustering

1 Introduction

In the contemporary digital age, humans are influenced by external digital stimuli, particularly through recommendation algorithms that impact emotions, thoughts, and actions [1,2]. On YouTube, thumbnails are powerful visual attractors that significantly affect a user's decision to view a video, making them crucial for capturing interest and guiding subsequent actions. The trend towards shorter video formats, such as YouTube Shorts, has reshaped content consumption, catering to fast-paced lifestyles seeking brief, engaging content.

This research explores biases in recommendation algorithms as they pertain to YouTube Shorts' thumbnails. By examining how these visual elements are

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 205–215, 2024.
https://doi.org/10.1007/978-3-031-72241-7_20

recommended and disseminated, the study aims to uncover patterns of bias. Specifically, it addresses the following research questions:

- **RQ1:** How does the topical content of YouTube Shorts' thumbnails change over time through recommendations?
- **RQ2:** What types of topics are more and less frequently recommended for YouTube Shorts after multiple recommendation cycles?
- **RQ3:** How does the content depicted in these thumbnails, as recommended by YouTube's algorithm, perpetuate biases on the platform?

To answer these questions, up-to-date topic modeling and content generation techniques were utilized. This research aims to enhance the understanding of the effects of recommendation algorithms on thumbnail content and their implications for content diversity and user engagement.

This paper is structured as follows: Sect. 2 reviews key concepts and relevant literature. Section 3 details our data collection and analytical methods. Section 4 presents findings on biases in YouTube Shorts' recommendations with graphical analyses. Finally, Sect. 5 summarizes key insights and discusses implications for future research and practical applications.

2 Literature Review

This literature review provides an overview of key studies relevant to various aspects of our research.

2.1 South China Sea Dispute

The South China Sea (SCS) is a critical geopolitical region with significant attention due to overlapping territorial claims and its strategic importance for maritime routes and natural resources.

The research by [3] highlights the dual factors of natural resources and freedom of navigation, emphasizing the SCS's abundant resources like oil, natural gas, and rich fishing grounds. They also stress the SCS's importance as a vital trade route. The work by [4] examines China's strategic approach, detailing efforts to consolidate claims and expand influence through diplomatic, economic, and military measures. The study by [5] analyzes China's assertiveness from 1970 to 2015, identifying key turning points and highlighting the cumulative effects of China's actions.

These studies underscore the SCS's impact on regional stability, international maritime law, and the balance of power in the Asia-Pacific region.

2.2 YouTube Shorts and Thumbnails

YouTube Shorts, introduced to meet the demand for short-form content, have quickly become a dominant format, particularly in entertainment categories [6].

These videos attract higher engagement metrics compared to regular videos (RVs) but pose monetization challenges due to fewer advertisement opportunities, requiring new revenue strategies for creators [7]. The popularity of Shorts reflects changes in viewer behavior, aligning with shorter attention spans [8].

Thumbnails on YouTube play a crucial role in attracting viewers and influencing video selection, directly impacting click-through rates (CTR) and engagement metrics. Visually appealing thumbnails with high view counts are more likely to be selected [9]. However, clickbait thumbnails can lead to viewer dissatisfaction when the content does not meet expectations [10,11]. Technological advancements like Optical Character Recognition (OCR) help detect and avoid clickbait, ensuring accurate representation of video content [12]. Thumbnails also influence algorithmic recommendations, as higher CTRs lead to more favorable placements in user feeds.

2.3 Recommendation Bias

Recommender systems significantly influence content consumption, often embedding biases and creating filter bubbles. Bias can arise from user preferences, algorithmic design, and training data. The study by [13] highlights how recommendation algorithms shape public discourse by promoting emotionally charged content [14]. Another work by [15] discusses content drift towards homogeneous and radical themes, emphasizing the importance of monitoring these shifts. Audits by [16–18] reveal the promotion of biased content, underlining the need for interventions. The work by [19] found that YouTube's algorithm promotes content with specific emotional tones, affecting user engagement. Cross-topical analysis by [20] reveals biases in diverse contexts and highlight content drift risks. Lastly, the research by [21] found that YouTube's algorithm can lead users, especially those with right-leaning ideologies, towards radical content [22]. Thus, addressing recommender bias and drift is crucial for a well-informed public.

To the best of our knowledge, no previous studies have specifically investigated biases in YouTube Shorts and their thumbnails. Therefore, our research offers a unique and novel contribution to the understanding of algorithmic biases in digital media platforms.

3 Methodology

This section details the methods used for data collection, topic modeling, and analysis of YouTube Shorts' thumbnails to investigate recommendation biases.

3.1 Data Collection

To initiate data collection, we held workshops with experts to generate relevant keywords for our search, targeting YouTube Shorts videos.

Due to the YouTube Data API's limitations with Shorts, we used APIFY's YouTube Scraper [23] to collect 1,210 unique video IDs. Finding this insufficient, we employed a snowball method to generate additional keywords, using the YouTube Data API and transcriptions from previous research [24,25].

The keywords for the South China Sea Dispute covered legal rulings, geopolitical tensions, and economic interests, enabling us to collect 2,094 unique video IDs for a detailed analysis of the conflict.

To measure bias in YouTube Shorts recommendations, we developed a custom scraping method.

Using collected video IDs as seed videos, we ensured a neutral user profile with fresh WebDriver instances. Automated with Selenium, our script scrolled through recommendations (depth) to a depth of 50, then started a new session for the next video.

We collected 104,700 videos, and after filtering out unavailable ones, our final dataset included 100,300 videos with their thumbnail images, obtained via the YouTube Data API.

3.2 Caption Generation

To investigate thumbnail images in detail and understand their context, we generated captions that describe the contents and events depicted in the images.

For this, we used GPT-4 Turbo, an enhanced version of OpenAI's GPT-4 language model. This model is optimized for speed and cost-effectiveness while maintaining similar capabilities and performance to the standard GPT-4, making it suitable for applications requiring quick responses and scalability. GPT-4 Turbo excels in natural language understanding and generation, supporting tasks from text completion and translation to content creation and conversational AI [26].

3.3 Topic Modelling

To understand the thumbnails' captions, we classified them into topics to track their evolution through recommendations.

We used GPT-4o, a refined and efficient version of GPT-4 designed for faster performance and lower costs while maintaining advanced natural language capabilities [26]. This model generated two types of topics general topics and categorized topics with specific constraints.

We also used BERTopic [27], a technique leveraging BERT embeddings to capture semantic similarities, resulting in coherent and interpretable topics. A fine-tuned version pre-trained on approximately 1,000,000 Wikipedia pages [28] was utilized, identifying 2,377 distinct topics. This robust framework effectively analyzed the thematic content of videos.

3.4 Clustering Topics

To analyze the topics discussed in Sect. 3.3, we clustered them due to their large number.

We filtered out non-informative topics like 'Photograph(s)', 'Thumbnail(s)', 'Image(s)', and 'Video(s)'.

BERT embeddings [29] were generated to capture semantic meaning through dense vector representations, considering both preceding and following contexts for accuracy.

These embeddings were reduced to two dimensions using t-SNE for visualization [30], which reveals intricate local patterns effectively.

The reduced features were clustered using the OPTICS algorithm [31], which handles varying data densities and is more flexible than other methods.

4 Results

This section presents our findings on biases in YouTube Shorts' recommendations, supported by graphical analyses of topic shifts and distributions.

4.1 Clustered General Topics with GPT

The GPT model generated 2,314 unique general topics. To visualize the topic clusters, we plotted them in a 2D space at various depths using t-SNE components for dimensionality reduction. We visualized the top 50 high-frequency topics. The legend distinguishes clusters by color and shows noise in gray.

As shown in Fig. 1, representing depth 0, topics are clustered together. Cluster 0 includes terms like politics, history, diplomacy, war, and military, indicating political themes. Clusters 3, 4, and 5 contain terms like ships, aircraft, and fishing, relating to aircraft carriers and economic perspectives at sea. Cluster 7 includes terms like geopolitics, map, and Philippines, highlighting the geographic perspective of the topic. Other clusters depict activities such as broadcasts, meetings, and presentations, with some, like Cluster 1, showing animated explanations of the topic. Initial depth videos were highly relevant to the investigated topic.

At depth 5, as displayed in Fig. 2, the topics are vastly different from the initial ones. Cluster 0 shows crafting topics, Cluster 1 covers machines and robotics, Cluster 3 is mostly about gaming, Cluster 4 includes dance, gym, and martial arts, Cluster 7 features child and dog-related terms, and Cluster 8 contains memes. The original topics have almost completely faded, with many new topics emerging.

We did not include all depths here because topics drifted early, and space constraints limited our illustrations to these two depths. More depths will be shown in the upcoming result subsections.

4.2 Categorized Topics with GPT

In this section, we investigated the categorized topics mentioned earlier in Sect. 3.3. Using the GPT model, we generated 20 categorized topics. These categories were determined by researching online and examining YouTube video categories, supplemented by additional categories from various news websites to ensure comprehensive coverage. Thus, we settled on 20 categories to encompass

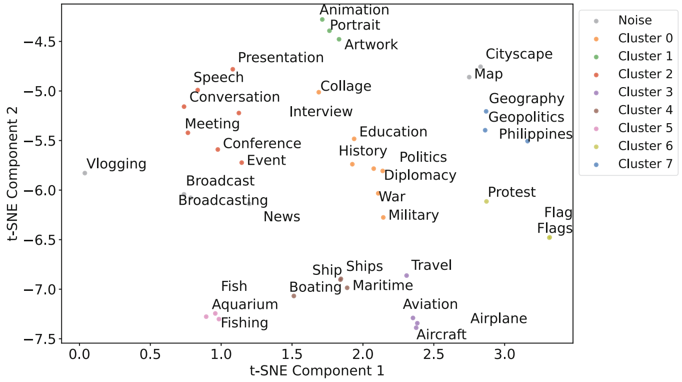


Fig. 1. General Topic Clustering for Depth 0

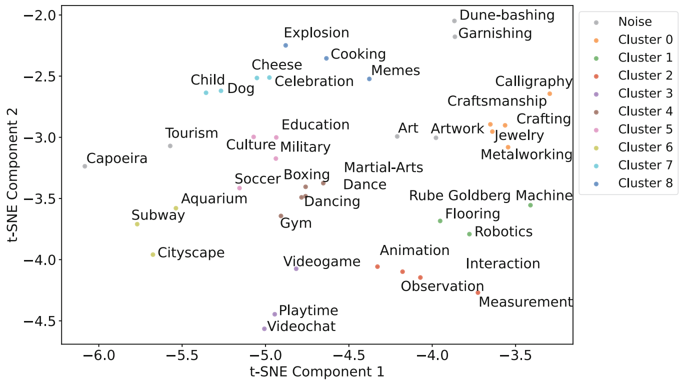


Fig. 2. General Topic Clustering for Depth 5

a broad range of topics as shown in Fig. 3. We used a lollipop chart for clarity, with the Y-axis representing the topics and the X-axis showing the topic ratios or distributions. The legend indicates depth ranges with corresponding colors. Topics for each depth range are accumulated and normalized, grouped into five classes for convenience and clarity, filtering out depths with a ratio below 0.01.

At the initial depth (depth 0), news dominates nearly 40% of the topics, followed by politics at around 15%, with other topics like history and lifestyle also present. As the recommendation algorithm suggested new videos, these new topics were neither news nor politics. News topics reduced dramatically across other depths, and political terms disappeared entirely. New topics, primarily entertainment-related, increased with each depth. Lifestyle topics remained relatively stable across depths due to their broad and encompassing nature. The graph clearly shows the topic shift happening in the recommendation algorithm.

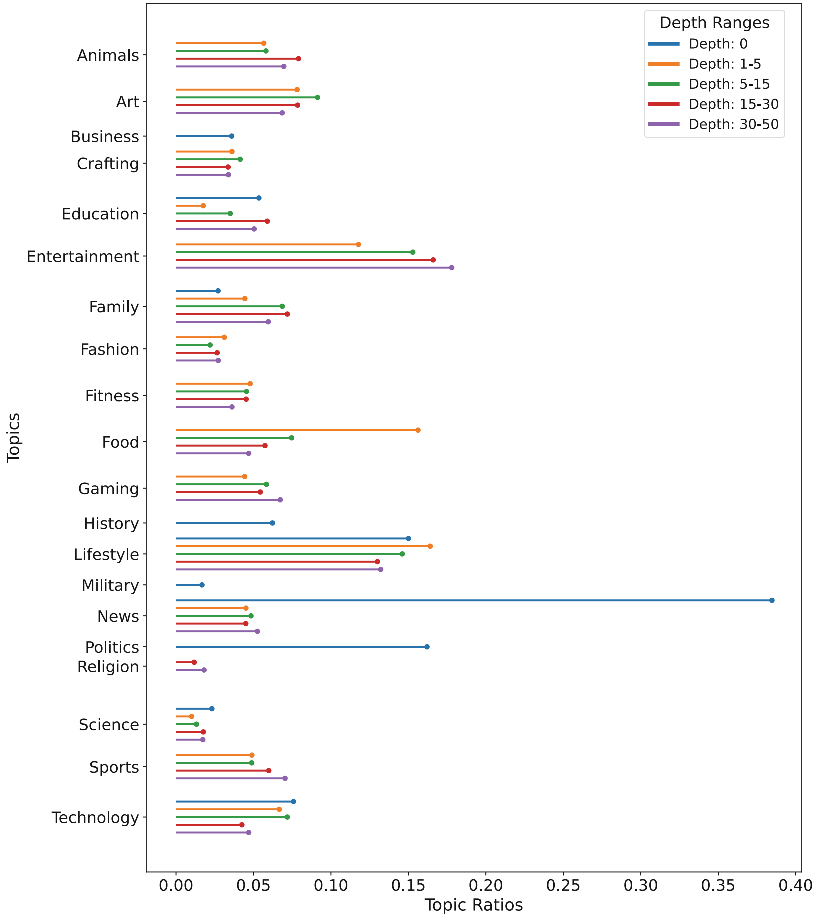


Fig. 3. The distribution of Categorized Topics Across Depths

4.3 Topic Distribution with BERTopic

In addition to generative AI, we utilized BERTopic for topic modeling, as detailed in Sect. 3.3. For illustration, we used radar charts to visualize the topics at initial, middle, and end depths as shown in Fig. 4. On the circle the topic IDs and names are represented, and the topic IDs are available on Huggingface [28] with details. We focused on the three most prevalent topics for each depth level to effectively track topic transitions.

At depth 0, the most prominent topics included flag, geography, and uniform. For example, topic 1650 (uniforms, uniformed, berets, beret) related primarily to military and soldiers, topic 935 (geography, geographic, geographical, geographer) concerned geopolitical regions around the South China Sea, and topic 111 (flags, flag, flagpole, commonwealth) referenced different nations. These topics clearly relate to the South China Sea Dispute.

As we move to deeper levels, we observed the emergence of unrelated topics, with the initial topics fading away. At later depths, we see topics like 706 (artistic, art, artwork, paintings), 5 (cuisine, cuisines, foods, culinary), and 1879 (lighting, lights, fluorescent, light). This shift indicates the recommendation algorithm's steering away from the original subject matter towards more general and unrelated topics.

Although BERTopic may not capture topics as effectively as GPT, it vividly demonstrates the topic drift across depths. The transition from focused, relevant topics to broader, unrelated ones highlights the algorithm's influence on content distribution, illustrating how quickly the focus can shift away from the initial subject matter.

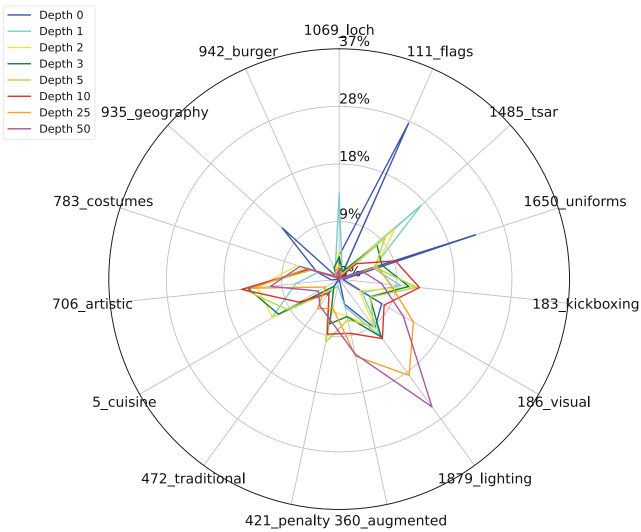


Fig. 4. Topic Distribution Using BERTopic Across Depths

5 Conclusion and Discussion

In this study, we investigated algorithmic bias in YouTube Shorts' video recommendations, focusing on thumbnail captions within the context of the South China Sea Dispute.

Our findings indicate a clear topic shift or drift in YouTube Shorts recommendations. After the initial videos, broader and less relevant topics are suggested due to YouTube's recommender system favoring high-engagement entertainment videos. This popularity bias results in the neglect of minority and serious issues, creating an algorithmic bias on YouTube Shorts. Consequently, these more popular but less serious videos are recommended more frequently.

For future work, we will address study limitations by comparing results with engagement scores to validate our assumptions. We will investigate various narratives, comparing well-known topics with niche subjects like the South China Sea Dispute, to understand recommendation levels. Additionally, we will incorporate interactive data collection (liking and commenting) to observe how user interactions affect recommendations and analyze text attributes such as titles, descriptions, and transcriptions.

This research highlights biases in YouTube's algorithmic recommendations, focusing on thumbnails. Understanding these biases is essential for fair representation of diverse topics, especially serious and minority issues. Thumbnails influence user engagement, and our study shows how algorithmic preferences can skew topic visibility. By exposing these biases, we contribute to the discourse on digital media ethics and the need for transparency in recommendation systems.

Acknowledgments. This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189, W911NF-23-1-0011, W911NF-24-1-0078), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

1. Cakmak, M.C., Shaik, M., Agarwal, N.: Emotion assessment of YouTube videos using color theory. In: Proceedings of the 9th International Conference on Multimedia and Image Processing (ICMIP). IEEE (2024)
2. Yousefi, N., Cakmak, M.C., Agarwal, N.: Examining multimodal emotion assessment and resonance with audience on YouTube. In: Proceedings of the 9th International Conference on Multimedia and Image Processing (ICMIP). IEEE (2024)
3. Macaraig, C.E., Fenton, A.J.: Analyzing the causes and effects of the South China Sea dispute: natural resources and freedom of navigation. J. Territ. Marit. Stud. **8**(2), 42–58 (2021). https://www.jstor.org/stable/48617340
4. Fravel, M.T.: China's strategy in the South China Sea. Contemp. Southeast Asia **33**(3), 292–319 (2011). http://www.jstor.org/stable/41446232
5. Chubb, A.: PRC assertiveness in the South China Sea: measuring continuity and change, 1970–2015. Int. Secur. **45**(3), 79–121 (2021). https://doi.org/10.1162/isec_a_00400
6. Violot, C., Elmas, T., Bilogrevic, I., Humbert, M.: Shorts vs. regular videos on YouTube: a comparative analysis of user engagement and content creation trends.

In: ACM Web Science Conference (WebSci 2024). ACM (2024). https://doi.org/10.1145/3614419.3644023
7. Rajendran, P.T., Creusy, K., Garnes, V.: Shorts on the rise: assessing the effects of YouTube shorts on long-form video content. arXiv preprint arXiv:2402.18208 (2024)
8. Sahu, G., Gaur, L., Singh, G.: Investigating the impact of personality tendencies and gratification aspects on OTT short video consumption: a case of YouTube shorts. In: 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), Uttar Pradesh, India, pp. 1–6 (2023). https://doi.org/10.1109/ICIPTM57143.2023.10118122
9. Park, J.: The impact of YouTube's thumbnail images and view counts on users' selection of video clip, memory recall, and sharing intentions of thumbnail images. The Florida State University (2022)
10. Qu, J., Hißbach, A.M., Gollub, T., Potthast, M.: Towards crowdsourcing clickbait labels for YouTube videos. In: HCOMP (WIP & Demo) (2018)
11. Shajari, S., Alassad, M., Agarwal, N.: Characterizing suspicious commenter behaviors. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, Kusadasi, Turkiye, pp. 631–635. ACM (2023). https://doi.org/10.1145/3625007.3627309
12. Vitadhani, A., Ramli, K., Dewi Purnamasari, P.: Detection of clickbait thumbnails on YouTube using Tesseract-OCR, face recognition, and text alteration. In: 2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST), pp. 56–61 (2021). https://doi.org/10.1109/ICAICST53116.2021.9497811
13. Cakmak, M.C., Okeke, O., Onyepunuka, U., Spann, B., Agarwal, N.: Analyzing bias in recommender systems: a comprehensive evaluation of YouTube's recommendation algorithm. In: Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2023), pp. 753–760 (2024). https://doi.org/10.1145/3625007.3627300
14. Alp, E., Gergin, B., Eraslan, Y.A., Çakmak, M.C., Alhajj, R.: Covid-19 and vaccine tweet analysis. In: Özyer, T. (ed.) Social Media Analysis for Event Detection. LNSN, pp. 213–229. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08242-9_9
15. Kirdemir, B., Kready, J., Mead, E., Hussain, M.N., Agarwal, N., Adjeroh, D.: Assessing bias in YouTube's video recommendation algorithm in a cross-lingual and cross-topical context. In: Thomson, R., Hussain, M.N., Dancy, C., Pyke, A. (eds.) SBP-BRiMS 2021. LNCS, vol. 12720, pp. 71–80. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80387-2_7
16. Srba, I., et al.: Auditing YouTube's recommendation algorithm for misinformation filter bubbles. ACM Trans. Recomm. Syst. **1**(1), 6 (2023). https://doi.org/10.1145/3568392
17. Gurung, M.I., Bhuiyan, M.M.I., Al-Taweel, A., Agarwal, N.: Decoding YouTube's recommendation system: a comparative study of metadata and GPT-4 extracted narratives. In: Companion Proceedings of the ACM on Web Conference 2024, pp. 1468–1472. Association for Computing Machinery (2024). https://doi.org/10.1145/3589335.3651913
18. Poudel, D., Cakmak, M.C., Agarwal, N.: Beyond the click: how YouTube thumbnails shape user interaction and algorithmic recommendations. In: The 16th International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (2024)

19. Okeke, O., Cakmak, M.C., Spann, B., Agarwal, N.: Examining content and emotion bias in YouTube's recommendation algorithm. In: Proceedings of the Ninth International Conference on Human and Social Analytics (HUSO 2023), Barcelona, Spain, pp. 15–20 (2023). https://www.thinkmind.org/index.php?view=article&articleid=huso_2023_1_40_80032
20. Cakmak, M.C., Okeke, O., Onyepunuka, U., Spann, B., Agarwal, N.: Investigating bias in YouTube recommendations: emotion, morality, and network dynamics in China-Uyghur content. In: Cherifi, H., Rocha, L.M., Cherifi, C., Donduran, M. (eds.) COMPLEX NETWORKS 2023. SCI, vol. 1141, pp. 351–362. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-53468-3_30
21. Haroon, M., Chhabra, A., Liu, X., Mohapatra, P., Shafiq, Z., Wojcieszak, M.: YouTube, the great radicalizer? Auditing and mitigating ideological biases in YouTube recommendations. arXiv preprint arXiv:2203.10666 (2022)
22. Shaik, M., Cakmak, M.C., Spann, B., Agarwal, N.: Characterizing multimedia adoption and its role on mobilization in social movements. In: Bui, T.X. (ed.) 57th Hawaii International Conference on System Sciences, HICSS 2024, Hilton Hawaiian Village Waikiki Beach Resort, Hawaii, USA, 3–6 January 2024, pp. 146–155. ScholarSpace (2024). https://hdl.handle.net/10125/106393
23. Streamers. Youtube Scraper. APIFY (2024). https://apify.com/streamers/youtube-scraper. Accessed 10 Jan 2024
24. Cakmak, M.C., Okeke, O., Spann, B., Agarwal, N.: Adopting parallel processing for rapid generation of transcripts in multimedia-rich online information environment. In: 2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 832–837 (2023). https://doi.org/10.1109/IPDPSW59300.2023.00139
25. Cakmak, M.C., Agarwal, N.: High-speed transcript collection on multimedia platforms: advancing social media research through parallel processing. In: 2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). IEEE (2024)
26. OpenAI, et al.: GPT-4 technical report. arXiv:2303.08774 (2024)
27. Grootendorst, M.: BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794 (2022)
28. Grootendorst, M.: BERTopic_Wikipedia. Huggingface (2024). https://huggingface.co/MaartenGr/BERTopic_Wikipedia. Accessed 1 May 2024
29. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2019)
30. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
31. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: OPTICS: ordering points to identify the clustering structure. ACM SIGMOD Rec. **28**(2), 49–60 (1999)

Moral and Emotional Influences on Attitude Stability Towards COVID-19 Vaccines on Social Media

Samantha C. Phillips[1(✉)], Lynnette Hui Xian Ng[1], Wenqi Zhou[2], and Kathleen M. Carley[1]

[1] CASOS Center, Software and Societal Systems Department, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA
{samanthp,huixiann,carley}@andrew.cmu.edu
[2] Palumbo-Donahue School of Business, Duquesne University, 600 Forbes Ave, Pittsburgh, PA 15282, USA
zhouw@duq.edu

Abstract. Effective public health messaging benefits from understanding antecedents to unstable attitudes that are more likely to be influenced. This work investigates the relationship between moral and emotional bases for attitudes towards COVID-19 vaccines and variance in stance. Evaluating nearly 1 million X users over a two month period, we find that emotional language in tweets about COVID-19 vaccines is largely associated with more variation in stance of the posting user, except anger and surprise. The strength of COVID-19 vaccine attitudes associated with moral values varies across foundations. Most notably, liberty is consistently used by users with no or less variation in stance, while fairness and sanctity are used by users with more variation. Our work has implications for designing constructive pro-vaccine messaging and identifying receptive audiences.

Keywords: Attitude strength · Emotion · Moral values · Social media

1 Introduction

Moral values and emotions are ubiquitous experiences that shape how we process information and form judgments. Efforts to curb public opinion, such as

This work was supported in part by the Knight Foundation and the Office of Naval Research grant MURI: Persuasion, Identity, & Morality in Social-Cyber Environments, N00014-21-12749. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research, or the U.S. Government.

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 216–225, 2024.
https://doi.org/10.1007/978-3-031-72241-7_21

promoting the acceptance of COVID-19 vaccines, can leverage moral and emotional reasoning to appeal to target audiences [4,12]. This practice can be more effective when the moral values or emotions are relevant to members of the target audience, that is, anti-vaccine belief holders [1,15,17,20]. Furthermore, moral values and emotions associated with an attitude can affect the stability of and openness to updating it. For example, attitudes with a moral basis have been associated with more attitude-consistent behavior and resistance to persuasion [2,9]. Similarly, emotional reactions to attitude-relevant information have been associated with greater attachment to that belief [22]. Understanding antecedents of attitude strength can reveal which individuals are most receptive to public health messaging.

Social media provides insight into how people's public-facing views change over time, and the corresponding change in justification for those views. In this work, we examine the relationship between the use of moral and emotional language in tweets about COVID-19 vaccines and the variance of stance towards COVID-19 vaccines expressed by the posting user. The variance of stance across two months is used as a proxy for attitude strength, which is characterized by stability over time, resistance to persuasion, and influence on behavior and information processing [7]. Through an empirical analysis the tweets of nearly 1 million X (formerly Twitter) users who tweeted about COVID-19 vaccines at least twice in a two month period, we aim to address the following:

RQ1: How is moral and emotional language used differently in pro- and anti-COVID-19 vaccine tweets?
RQ2: How is moral and emotional language used differently by X users who indicate variation in expressed stance versus those who have no variation in expressed stance?
RQ3: Do X users who include more moral and emotional language in their tweets about COVID-19 vaccines have more or less variation in stance towards COVID-19 vaccines?

The selected time period between March and May of 2021 coincides with some of the highest vaccination rates in the United States throughout the COVID-19 pandemic[1]. For each tweet, we applied a lexicon-based methodology to detect moral foundations and emotions. In addition, we developed and applied a classifier to label the stance of each tweet towards COVID-19 vaccines as {pro, anti, neutral}. We then calculated the average use of each moral value and emotion, as well as the standard deviation of stances, across tweets for each user.

In *RQ1*, we provide an overview of which moral foundations and emotions are used in pro- and anti-COVID-19 vaccine tweets. In *RQ2*, we compare the moral and emotional language used by the 310,258 users who expressed a single stance about COVID-19 vaccines with the 603,158 users with any variation in stance represented in their tweets. In addition, in *RQ3*, we identify associations between moral and emotional language and stance variance among users with any variation in stance, excluding users with zero variation.

[1] https://usafacts.org/visualizations/covid-vaccine-tracker-states/.

2 Related Works

Extensive previous work has examined associations between moral values and judgments. Moral foundation theory proposes six automatic intuitions (i.e., foundation or value) that affect judgments [6]. These foundations include: care (well-being of others), fairness (justice), sanctity (purity), liberty (freedoms), loyalty (in-group/out-group relations), and authority (following rules and traditions). Moral foundations most associated with vaccine hesitancy are purity and liberty, followed by authority [1]. In addition, vaccine hesitancy is associated with less need for care [1,17]. However, these studies simply ask people about their automatic moral intuitions and attitudes towards vaccines, so there is no assessment of changes in stance over time or how moral reasoning is used to justify judgments about vaccines specifically. Nonetheless, they demonstrate existing interest in how moral values affect attitudes towards vaccines.

Like moral intuitions, emotions can affect information processing and attitude formation. While there are a multitude of ways to conceptualize emotions, we use Plutchik's six-emotion model: happy, sad, anger, fear, disgust, surprise [14]. This model has been used to study emotions on social media [8,13]. Certain emotional reactions to vaccine-relevant events can predict vaccine hesitancy, like anger and anxiety [20]. We build on this work to assess which emotions are used in pro- and anti-vaccine tweets.

Scholars have evaluated moral aspects of an attitude as an antecedent to attitude strength, finding attitudes with a moral basis are more resistant to influence [2,9] and predict consistent behavior [9]. Furthermore, emotional responses to attitude-relevant information typically indicate a high degree of importance has been placed on that view, implying attitude strength [11,22]. We investigate if using emotional language to discuss a stance is also associated with stability over time. Unlike previous analyses of antecedents of attitude stability, we use stance variability in social media posts over time as a measure of attitude strength, extending these works to consider stance over time outside a laboratory environment. Taking these studies together, we expect that moral and emotional language are associated with *less* variation in stance towards COVID-19 vaccines.

3 Data and Methods

3.1 Dataset

Our dataset was initially collected using a streaming keyword search via Twitter v1 API between March 12, 2021 thru May 11, 2021. We collected tweets that contained at least one of the following terms: coronavirus, Wuhan virus, Wuhanvirus, 2019nCoV, NCoV, NCoV2019, covid-19, covid19, covid 19. We then further filtered the tweets to select those about vaccines using the following keywords: vaccine, vax, mrna, autoimmuneencephalitis, vaccination, getvaccinated, covidisjustacold, autism, covidshotcount, dose1, dose2, VAERS, GBS, believemothers, mybodymychoice, thisisourshot, killthevirus, proscience, immunization, gotmyshot, igottheshot, covidvaccinated, beatcovid19, moderna, astrazeneca, pfizer, johnson & johnson, j&j, johnson and johnson, jandj.

In total, we extracted 2,283,281 users, 1,369,865 of which only tweeted one time about COVID-19 vaccines during this time period. In this work, we analyze the remaining 913,416 users who tweeted at least twice in our time period for a total of 6,811,854 tweets.

3.2 Stance Detection

To evaluate stance consistency, we first developed a classifier to label the stance (pro, anti, neutral) of each tweet towards COVID-19 vaccines. To generate a training dataset, two annotators labelled 1034 tweets with Cohen's Kappa of 0.741 and agreement of 84.2%. A third annotator labelled the 163 tweets the initial two annotators disagreed on. We removed the 6 tweets that all three disagreed on. We then fine-tuned a BERTweet model[2] for the stance detection task to obtain an accuracy of 72.8% on the evaluation data. The stance classifier returns a stance label and confidence score for each tweet.

We applied this stance detection classifier to original tweets, retweets, replies, and quote tweets in our dataset. This gives us 3,547,645 pro-vaccine, 2,928,820 neutral, and 335,389 anti-vaccine tweets. To obtain a measure of variation in stance per user, we calculated the standard deviation of stance expressed across their tweets. Of the 913,416 users in our dataset, 310,258 express a consistent stance. Removing the 310,258 users with a standard deviation of 0, the average standard deviation is 0.62 (std. dev. 0.18) and median is 0.58.

3.3 Moral Value and Emotion Extraction

We use the Netmapper software to extract words and phrases associated with each moral value (care, fairness, liberty, sanctity, loyalty, authority) and emotion (happiness, sadness, fear, anger, disgust, surprise) [3,21]. We recorded the number of concepts representing each moral/emotion variable in each tweet.

4 Results

4.1 RQ1: Moral Values and Emotions in Pro and Anti-COVID-19 Vaccine Tweets

Figure 1 displays the average and 95% confidence interval of the number of concepts associated with each moral value and emotion in pro- and anti-COVID-19 vaccine tweets. Overall, anti-vaccine tweets contain more moral and emotional language than pro-vaccine tweets for most types of moral values and emotions. That is, anti-vaccine tweets contain more references to care, fairness, authority, sanctity, and liberty foundations than pro-vaccine tweets. On the other hand, pro-vaccine tweets contain more loyalty terms. Anti-vaccine tweets are also more

[2] https://huggingface.co/docs/transformers/model_doc/bertweet.

likely to contain sad, fear, anger, disgust and surprise emotions, while pro-vaccine tweets contain more happiness. Table 1 contains examples of tweets containing select moral values and stances.

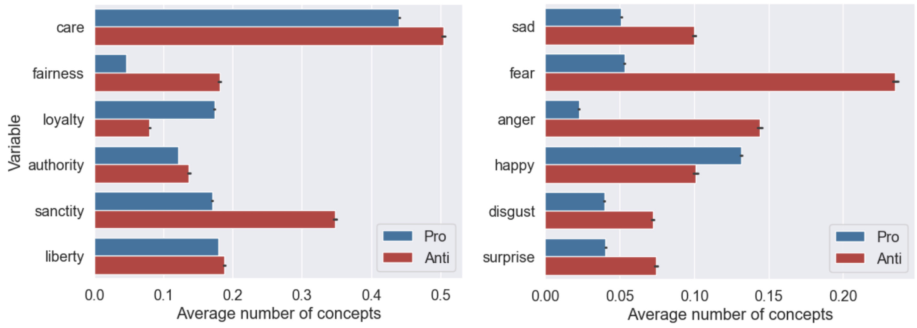


Fig. 1. Average and 95% CI number of concepts associated with each moral foundation (left) and emotion (right) in pro- and anti-COVID-19 vaccine tweets.

Table 1. Example tweets

Stance	Moral foundation	Example tweet
pro	care	If you can't see the urgency of saving humanity by granting a #TRIPSWaiver to ensure #VaccineForAll, I don't see what else you see. Time to stop putting Health in the Marketplace
anti	care	"Vaccinating" babies now with the mRNA jabs? How can this even be ethical?
pro	loyalty	His son has taken both doses of the COVID-19 vaccine, and hopefully was wearing a mask. They are both putting action behind their push towards herd immunity. I hope you'll get the vaccine too!
anti	sanctity	I am not taking the vaccine. I take 5000 units of vitamin D3. If God wants me to die from COVID-19 no vaccine will stop God's Will

4.2 RQ2: Moral Values and Emotions in Tweets by Users Expressing no Versus any Variation in Stance

Table 2 contains t-statistics from t-tests comparing the use of moral and emotional language by users with no versus any variance in expressed stance. Users that are completely consistent in their expressed stance tend to use anger, loyalty, care, and liberty in their tweets more than users that express variation in their stance. Users who express any variation in stance are more likely to use happy, disgust, surprise, fear, or sad emotional language. Furthermore, they tend to refer to sanctity, authority and fairness foundations more in their tweets.

Table 2. T-statistics from t-tests comparing the average number of concepts associated with each moral value and emotion in tweets by users who expressed no variation versus any variation in stance towards COVID-19 vaccines. Negative values indicate the variable is used more in tweets by users that express no variation in stance. All t-tests are significant ($p < 0.0001$)

Emotion		Moral value	
Variable	T-statistic	Variable	T-statistic
Anger	−44.22	Loyalty	−37.95
Happy	6.36	Care	−34.36
Disgust	17.26	Liberty	−22.57
Surprise	30.18	Sanctity	21.90
Fear	32.3	Authority	27.42
Sad	37.86	Fairness	44.43

4.3 RQ3: Moral Values and Emotions in Tweets by Users Expressing Some Degree of Variation in Stance

Figure 2 displays the coefficients for an OLS regression model predicting the standard deviation of expressed stance across tweets for each user given the use of moral and emotional language in their tweets about COVID-19 vaccines. All coefficients are significant ($p < 0.0001$) except for anger ($p > 0.1$).

More emotional language is largely associated with more variation in stance, except surprise. Liberty and loyalty are associated with less variation in stance, while the remaining moral foundations (fairness, sanctity, authority, care) are associated with more variation.

5 Discussion

The aim of this work is to investigate how moral foundations and emotions associated with an attitude affect the stability of that attitude in a social media context. To do so, we analyzed tweets expressing a stance toward COVID-19 vaccines.

First, we examined which moral foundations and emotions are used in tweets expressing pro- and anti-COVID-19 vaccine views to provide an overview of the context in which each feature is typically used. The findings in *RQ1* indicate that anti-vaccine tweets tend to contain more moral and emotional language than

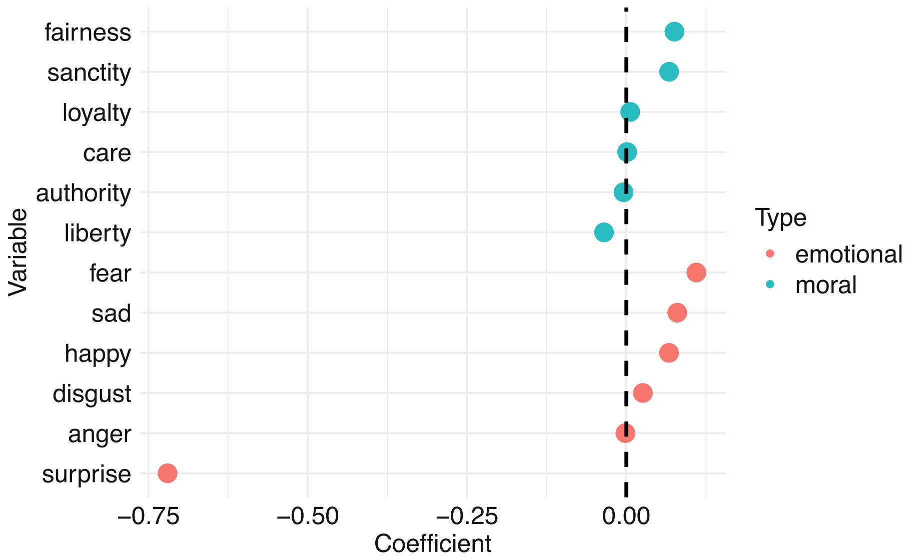


Fig. 2. Regression coefficients for the standard deviation of expressed stance across tweets. Negative values indicate the variable is associated with less variation in stance. Error bars indicate the 95% confidence intervals. All coefficients are significant ($p < 0.0001$) except for anger ($p > 0.1$) and care ($p = 0.049$)

pro-vaccine tweets except for loyalty and happiness. Many pro-vaccine tweets that contain language associated with loyalty contain information about where a community can access vaccines or celebrate that vaccines were successfully given to a community (e.g., Americans, indigenous tribes in Oklahoma). As one may expect, vaccine hesitant messaging tends to use more negative emotional language than those expressing support.

Our findings echo previous work showing vaccine hesitancy is associated with moral values like sanctity and liberty [1]. Anti-vaccine tweets that use language associated with sanctity include concerns about the efficacy and safety of COVID-19 vaccines [1], and use more negative emotional language [5]. However, while vaccine hesitancy has been associated with lack of need for care [1,17], we find care is the most used moral foundation in anti-vaccine tweets. These tweets include concern about how negative side effects from vaccines are affecting or may affect others.

In addressing *RQ2* and *RQ3*, we find emotional language is associated with more variation in stance, except for anger and surprise. That is, certain emotions (fear, sadness, happiness, disgust) indicate unstable attitudes. Users may be less tied to their view once their emotional reaction is less salient. Indeed, studies show that both positive and negative emotions can override analytical thinking processes, increasing the effect of biased information processing on truth discernment [16,18].

Conversely, anger is associated with consistent stance, but the association between anger and stance variance among those who do vary to some degree is not significant. This highlights the different ways different emotions can impact judgments depending on the individual and context. Surprise is most associated with users that express some, but small, variance in stance. Surprise drives belief updating by attracting attention to unexpected information, motivating people to reconsider and potentially rectify their understanding [10].

The role of moral language varies across stance foundations. Liberty is consistently used by users with no or less variation in stance, while fairness and sanctity are used by users with more variation in stance towards COVID-19 vaccines. Care and loyalty are associated with no or high variation, while authority is associated with some but low variation. The same moral value can play a very different role depending on the context and individual differences. For example, moral identity is how important the individual considers being a "moral" person to their self-concept. Prioritizing "binding" values (loyalty, authority, sanctity) represent a group-oriented view of morality. Adherence to binding foundations is associated with out-group derogation if moral identity is weak and willingness to help out-group members if moral identity is strong [19]. More work is needed to identify variables influencing the effect of moral values on stance consistency.

The results of this work inform how moral values and emotions associated with a belief affect the stability of that belief. Our work has implications for constructing effective campaigns to promote vaccines and debunk vaccine misinformation, as well as directing these messages to receptive audiences. Specifically, promoting and tailoring pro-vaccine messaging to people expressing moral and emotional language associated with less attitude stability (i.e., sanctity, fairness, fear, sadness, happiness, disgust) can help to optimize public health outreach.

5.1 Future Work and Limitations

Our study is limited to the platform X. Future work should assess the use of moral values and emotions towards vaccine attitudes on other social media platforms. These analyses could provide vital insight into how organizations designing public health messaging should anticipate shifts in motivations as vaccines are developed and approved for the public. Furthermore, we did not attempt to identify automated accounts ("bots") in our analysis [13]. Future work should examine the extent to which removing users displaying bot-like behavior affects the observed relationship between moral-emotional bases for an attitude and the consistency of that attitude.

Rather than separately examining which moral values and emotions predict more or less vaccine hesitancy over time, we looked at any change in stance regardless of the direction. Disentangling the effects on increases and decreases in anti-vaccine attitudes can directly inform public health messaging aiming to increase vaccinations [17].

Lastly, attitudes towards vaccines are multi-faceted, complex judgments that often include far more nuance that we include in this work. Indeed, vaccine hesitancy can be motivated by a wide range of factors, such as concerns over vaccine testing protocols, institutional distrust, and belief in alternatives [23]. Considering more fine-grained stance categories may reveal important distinctions in the role of moral values and emotions in stance (in)consistency.

6 Conclusion

We studied the influence of moral and emotional associations with an attitude on the stability of that attitude over time. Specifically, we examined the moral values, emotions, and stance expressed in tweets by nearly one million users on X over a two month period. We find that (1) moral and emotional language is used more in anti-vaccine tweets, except for happiness and loyalty. (2) Emotional language is associated with larger variation with stance, except for anger and surprise, which are associated with consistent stance. (3) The role of moral language and stance variability differs across values, e.g. liberty is associated with consistent stances while fairness and sanctity are associated with larger stance variation. Our work shows that moral and emotional reasoning in attitudes can be used to predict receptive audiences, which can be especially useful in counter-misinformation campaigns.

References

1. Amin, A.B., et al.: Association of moral values with vaccine hesitancy. Nat. Hum. Behav. **1**(12), 873–880 (2017)
2. Aramovich, N.P., Lytle, B.L., Skitka, L.J.: Opposing torture: moral conviction and resistance to majority influence. Soc. Influ. **7**(1), 21–34 (2012)
3. Carley, L.R., Reminga, J., Carley, K.M.: Ora & netmapper. In: Thomson, R., Dancy, C., Hyder, A., Bisgin, H. (eds.) Social, Cultural, and Behavioral Modeling. Demo at SBP-BRiMS 2018. Springer, Cham (2018)
4. Chou, W.Y.S., Budenz, A.: Considering emotion in COVID-19 vaccine communication: addressing vaccine hesitancy and fostering vaccine confidence. Health Commun. **35**(14), 1718–1722 (2020)
5. Faasse, K., Chatman, C.J., Martin, L.R.: A comparison of language use in pro-and anti-vaccination comments in response to a high profile Facebook post. Vaccine **34**(47), 5808–5814 (2016)
6. Graham, J., Nosek, B.A., Haidt, J., Iyer, R., Koleva, S., Ditto, P.H.: Mapping the moral domain. J. Pers. Soc. Psychol. **101**(2), 366 (2011)
7. Krosnick, J.A., Petty, R.E.: Attitude strength: an overview. In: Attitude Strength, pp. 1–24 (2014)
8. Liu, Z., Zhang, T., Yang, K., Thompson, P., Yu, Z., Ananiadou, S.: Emotion detection for misinformation: a review. Inf. Fusion **107**, 102300 (2024)
9. Luttrell, A., Petty, R.E., Briñol, P., Wagner, B.C.: Making it moral: merely labeling an attitude as moral increases its strength. J. Exp. Soc. Psychol. **65**, 82–93 (2016)
10. Maguire, R., Maguire, P., Keane, M.T.: Making sense of surprise: an investigation of the factors influencing surprise judgments. J. Exp. Psychol. Learn. Mem. Cogn. **37**(1), 176 (2011)
11. Miller, J.M., Krosnick, J.A., Holbrook, A., Tahk, A., Dionne, L.: The impact of policy change threat on financial contributions to interest groups. In: Political Psychology: New Explorations, pp. 172–202 (2016)

12. Nan, X., Iles, I.A., Yang, B., Ma, Z.: Public health messaging during the COVID-19 pandemic and beyond: lessons from communication science. Health Commun. **37**(1), 1–19 (2022)
13. Ng, L.H.X., Carley, K.M.: Bot-based emotion behavior differences in images during Kashmir black day event. In: Thomson, R., Hussain, M.N., Dancy, C., Pyke, A. (eds.) Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2021. LNCS, vol. 12720, pp. 184–194. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80387-2_18
14. Plutchik, R.: The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am. Sci. **89**(4), 344–350 (2001)
15. Rossen, I., Hurlstone, M.J., Dunlop, P.D., Lawrence, C.: Accepters, fence sitters, or rejecters: moral profiles of vaccination attitudes. Soc. Sci. Med. **224**, 23–27 (2019)
16. Scheufele, D.A., Krause, N.M.: Science audiences, misinformation, and fake news. Proc. Natl. Acad. Sci. **116**(16), 7662–7669 (2019)
17. Schmidtke, K.A., et al.: Evaluating the relationship between moral values and vaccine hesitancy in Great Britain during the COVID-19 pandemic: a cross-sectional survey. Soc. Sci. Med. **308**, 115218 (2022)
18. Sinclair, M., Ashkanasy, N.M., Chattopadhyay, P.: Affective antecedents of intuitive decision making. J. Manag. Organ. **16**(3), 382–398 (2010)
19. Smith, I.H., Aquino, K., Koleva, S., Graham, J.: The moral ties that bind... even to out-groups: the interactive effect of moral identity and the binding moral foundations. Psychol. Sci. **25**(8), 1554–1562 (2014)
20. Sun, R., Wang, X., Lin, L., Zhang, N., Li, L., Zhou, X.: The impact of negative emotional reactions on parental vaccine hesitancy after the 2018 vaccine event in China: a cross-sectional survey. Hum. Vaccines Immunotherapeutics **17**(9), 3042–3051 (2021)
21. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010)
22. Visser, P.S., Krosnick, J.A., Norris, C.J.: Attitude importance and attitude-relevant knowledge: motivator and enabler. In: Political Psychology, pp. 217–259. Psychology Press (2016)
23. Weinzierl, M.A., Harabagiu, S.M.: From hesitancy framings to vaccine hesitancy profiles: a journey of stance, ontological commitments and moral foundations. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 1087–1097 (2022)

The Life of a Tie: Social Origins of Network Diversity

Patrick S. Park(✉), Henry George Xu, and Kathleen M. Carley

Carnegie Mellon University, Pittsburgh, PA 15213, USA
{patrickp,hgx,kathleen.carley}@cs.cmu.edu

Abstract. This study examines the survival and evolution of 443K bidirectional mention ties on Twitter by merging datasets collected before 2015 and in the first few months of the COVID-19 pandemic (February to June, 2020). We hypothesize that strong pre-existing ties, marked by frequent communication and shared identities, endure and tolerate cognitive and stance differences over time. Our findings show that surviving ties are stronger than average pre-2015 ties but exhibit greater cognitive distance in COVID-19 discussions, suggesting that strong ties can tolerate different and even opposing opinions on contentious topics. This challenges traditional models of social influence and homophily, which predict increased cognitive similarity within strong ties. The findings imply the potential for old ties to function as network bridges, reducing political divides by connecting dissimilar social groups.

Keywords: strong ties · network diversity · tie survival · partisan bridging · covid-19

1 Introduction

Homophily is one of the most salient principles that give structure to social networks [11]. However, in almost any social context, social ties that diverge from this principle do exist - they tend to be weak network bridges that connect otherwise distant, dissimilar network communities. Viewed from the perspective of homophily, such non-homophilous ties are theoretically puzzling, yet the literature tends to explain away their presence with a functionalist logic, highlighting the information and control advantages that people potentially gain from bridging dissimilar groups [2,8]. This logic implies that individuals form and maintain non-homophilous ties in pursuit of such instrumental benefits [1,9]. However, recent studies show that bridging ties spanning long network distances may not necessarily rest on instrumental motivation - discussions between Twitter users who form a long-range tie tend to contain more words indicative of affective role relations (e.g., buddy, friend) and fewer words related to professional role relations (e.g., boss) [13]. The Chinese term, "guanxi", which generally denotes a particularistic relationship characterized by the warmth, trust, and obligation to another individual, can span dissimilar social communities, even in the

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 226–235, 2024.
https://doi.org/10.1007/978-3-031-72241-7_22

absence of structural support of common network neighbors [3]. How, then, do non-homophilous, structurally bridging ties come about?

Building on sociological insights highlighting the long-term cumulation of shared experience and history of a social tie for building trust [3,7], the current study explores the theory that non-homophilous bridging ties are likely the exceptionally strong ties that survive the long term, while their weaker adjacent ties decay with time [13]. A strong social tie may initially form in a tightly-knit group of people with similar attributes and in overlapping social contexts (i.e., high clustering and homophily with multiple shared foci) [6]. With the passage of time, group members tend to dissipate in social space to different schools, occupations, social classes, life styles, geographies, and political affiliations throughout the life course, leading the weaker ties that existed in the group to either naturally decay due to competing priorities or abruptly break when trust is breached. However, some of the strongest ties may endure these naturally compounding survival pressures of diverging life paths and experiences. The network neighbors who successfully sustain their strong ties over the long term may, perhaps unbeknownst to themselves, become the network bridges that structurally connect communities with different life styles, moral values, political beliefs, and group norms.

Our strategy for empirically testing this theory is to identify pairs of Twitter users that communicate over several years (@mention communication ties) and assess the changes in their relational strength, cognitive similarity, political disagreements, and the mutuality of communication. Specifically, among a sample of Twitter users who discussed COVID-19-specific topics with other users throughout the politically contentious months of the pandemic, we identify the @mention ties that had been observed previously in a large-scale Twitter data corpus collected before 2015. We analyze how the past relational and structural characteristics as well as the attribute similarities of these ties associate with those characteristics in the present and assess their capacity to maintain communication despite opposing stances on Covid vaccination, a hot-button issue in the U.S. that sparked fierce debates about individual freedom, public safety and health, and widening political divisions along pre-existing party lines. We hypothesize that communication partners who used to frequently communicate (i.e., strong ties), those who used to exhibit higher cognitive similarity, and those who shared similar identities (e.g., occupation, familial roles, political affiliation) and cultural interests (e.g., musical genre and sports) are more likely to maintain active bilateral communications over the long term. We further hypothesize that communication partners who likely initially developed the relationship based on similarity (i.e., homophily) will have developed tolerance for each other's differences in identity, cognition, and stance that tend to occur with the passage of time.

2 Data

We merge two temporally separated Twitter datasets to trace the survival and cognitive divergence of communication ties over time. The first dataset consists of 26M U.S. Twitter users and their timelines of up to 3200 tweets, collected between 2013 and 2014 (the "pre-2015 data") [13]. The second dataset consists of a sample of tweets containing COVID-19 related keywords from February to June 2020 (the "Covid data"). We identify 443K bidirected mention ties (i.e., reciprocal mentions between two users) among 380K U.S. Twitter users in the pre-2015 data that also reappear in the more recent Covid data, either as unidirected or bidirected mention ties. These are the users who may have maintained their communications for five years or longer. While the pre-2015 data contain a more comprehensive set of tweets that a user created at the time, the Covid data covers only the Covid-specific tweets, limiting our observation of communications to COVID-19 subtopics. Nevertheless, it is possible to assess with these data how past relationships may affect communications in the long term on potentially sensitive topics during challenging times.

2.1 Variables

Tie Strength. We measure the strength of a tie by the frequency of mention tweets exchanged between two users, separately from both the pre-2015 data (past tie strength) and the Covid data (recent tie strength).

Tie Range. Following [13] and [8], we compute the second shortest path length of a tie from the pre-2015 data as a measure of the network structural distance that the tie bridges between otherwise disconnected network neighborhoods.

Cognitive Distance. We use a pre-trained 100-dimension GLoVe word embedding [14] to map each user's cognitive location in high-dimensional vector space by aggregating the word vectors from their respective tweets. We subsequently measure the cognitive distance between users by computing the Euclidean distance of their respective aggregate word vectors. This cognitive distance reflects a broad range of differences in knowledge, interests, and beliefs, and can also be viewed as a measure of (the inverse of) cognitive homophily. By computing this cognitive distance of a mention tie separately from the two datasets, we gauge the relative growth or reduction in cognitive distance over time.

Stance Distance. The cognitive distance between two users based on the GLoVe word vectors may simply reflect *different* interests and not necessarily *opposing* stances on contested issues. Therefore, we measure the users' stance on COVID vaccination and compute the difference in stances between two users as an indication of opposing positions that a tie may be tolerating. Specifically, from the Covid dataset, we use a pre-trained COVID-Twitter-BERT model [12] for tweet-level vaccination stance detection. The last layer hidden state computed from this model is used as an abstract vector representation of the vaccination-related stance expressed in a tweet. We aggregate these tweet-level stance representations at the individual level and use it as a summary vector representation of the individual's overall stance on vaccination. Although this approach does not allow straight forward qualitative interpretation of a user's stance (e.g., pro- vs. anti-vaccination), it is a richer representation that renders a continuous, granular measure of stance opposition between two users. Although the hidden layer representation does not provide a readily human-readable indicator, its correspondence to the intuitive categorical labels are straightforward as shown in Fig. 1, which demonstrates the clear separation and groupings in both PCA and t-SNE plots.

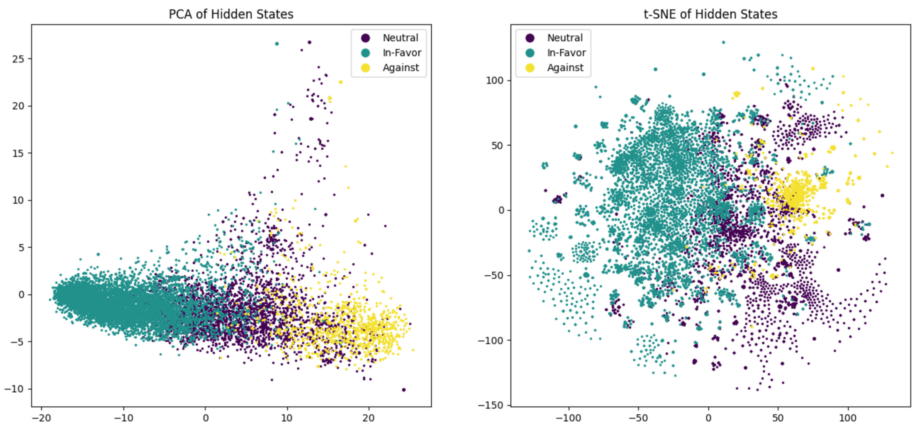


Fig. 1. PCA and t-SNE visualizations of hidden states. Colors indicate different stance labels on Covid vaccination, with each cluster representing a distinct stance group

Identity Overlap. Apart from the cognitive (dis)similarity of two users, we explore how the similarity in users' salient past identities (i.e., identity homophily) and cultural interests correlate with the survival of their communication ties. Specifically, we construct a prototype identity lexicon[1] for the categories of occupation (e.g., "reporter" at Boston Globe), familial roles (e.g., proud "father"), political affiliation (life-long "democrat"), interest in sports (e.g., football), and cultural consumption (e.g., "punkrock", "historybuff") and code the occurrence of these terms in the user profile descriptions. Based on this identity and interest codings of each user's profile description, we construct category-specific dichotomous variables that indicate whether two users who share a tie both listed the same identity terms or not in their pre-2015 profile descriptions.

[1] Lexicon and keyword occurrences available: https://doi.org/10.5281/zenodo.11430935.

2.2 Analytical Strategy

In the first set of analyses, we describe the characteristics of the pre-2015 bidirected mention ties that reappear in the Covid dataset either as unidirected or bidirected mention ties, in terms of their tie strengths, cognitive distances, and vaccination stance distances. Based on these quantitative descriptions, we assess the likelihood of the pre-2015 bidirected tie to remain in bidirected communications (i.e., tie survival), relative to one-way (unidirected) communications in the Covid data. Here, we use logistic regression to evaluate the odds of a pre-2015 tie showing up as bidirected in the Covid dataset with other tie-level covariates, including pre-2015 tie strength and shared identities on each identity category.

3 Results

Consistent with our assumption that strong ties are more likely to survive, the ties observed in the pre-2015 data that reappear in the Covid data tend to be relationally stronger than the average pre-2015 tie. As shown in Fig. 2, these "old ties" exchanged approximately 40% more mention tweets than the average tie in the pre-2015 data (red dashed line), irrespective of the network distance (i.e., tie range) they spanned in the past.

Figure 3 plots the mean cognitive distance separately measured from the two datasets. Since the Covid dataset was constructed from tweets about a narrower issue than the issue-agnostic pre-2015 tweets, the former exhibits a shorter cognitive distance (orange) than the distance in the general, topic-agnostic tweets from the pre-2015 data (blue) as one would expect. Furthermore, while the pre-2015 cognitive distances tend to increase with tie range as expected, more noteworthy is the same pattern emerging in the Covid cognitive distances. That is, the tie range from years prior continues to correlate with the cognitive distance around Covid.

Figure 4 plots the relationship between cognitive distance (y-axis) and tie strength (x-axis), crossed by the time of measurement (pre-2015 vs. Covid). Panels A and D (main diagonal) exhibit associations between cognitive distance and tie strength that are consistent with homophily - when measured in the same time periods, the users who engage in more frequent conversations also tend to exhibit greater cognitive similarity at that time. Surprisingly, however, the ties that are cognitively distant in their Covid-related discussions used to engage in more frequent communications pre-2015 (panel B), suggesting that the relationally strong ties that survive over time might develop the trust and tolerance to discuss their differing conceptions on such contentious issues as Covid. Conversely, the old ties that engage in more frequent Covid-related conversations used to exhibit greater cognitive distance pre-2015 (panel C), a result that is not readily explainable by homophily.

Granted, cognitive distance does not necessarily indicate that two users hold opposing stances on contentious issues. For example, one user might tweet mostly about the psychological isolation of the lockdown while the other might tweet about the supply shortages triggered by the lockdown. Therefore, in Fig. 5,

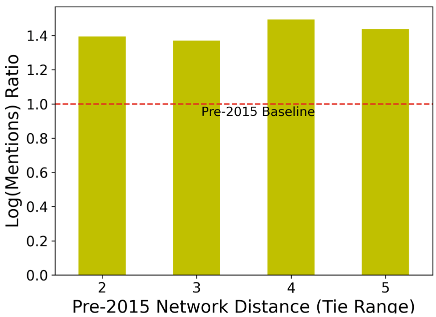


Fig. 2. Old ties observed in COVID-19 tweets (survived ties) tend to be stronger than the average old tie (red horizontal line). Tie range is the second shortest path length of a tie, which measures the network distance that the tie spans (Color figure online)

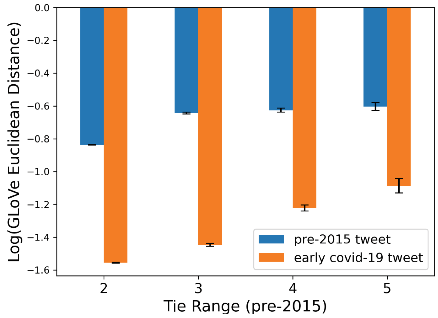


Fig. 3. Cognitive distance of old ties. Compared to the cognitive distance measured in their dyadically exchanged tweets pre-2015 (blue bars), the cognitive distance in the COVID-19 tweets they exchanged (orange bar) tends to be shorter (Color figure online)

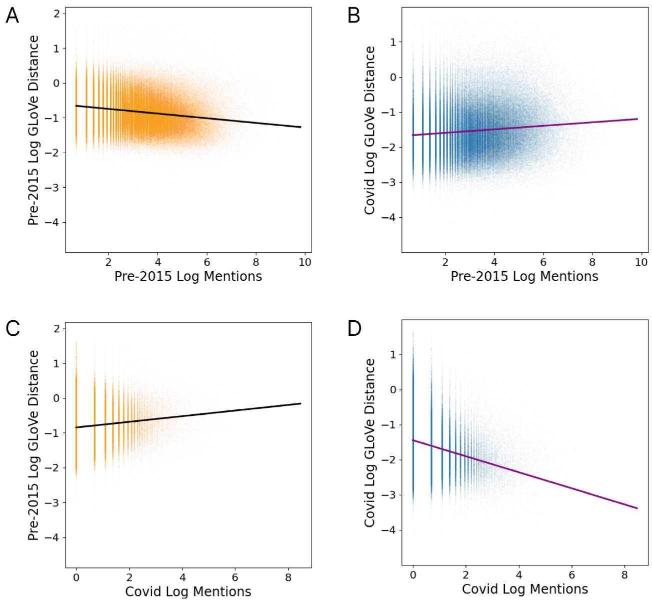


Fig. 4. Cognitive distance in conversations between old ties (pre-2015 and COVID data) and their relational strength measured by mention frequency (pre-2015 and COVID data). The strong ties in pre-2015 tend to show greater cognitive distances in their COVID conversations (panel B)

we directly measure stance opposition on the issue of Covid vaccination and explore its association with the a tie's relational strengths, measured from the two datasets. Similar to the counter-intuitive increase in Covid-related cognitive distance with the increase in pre-2015 mention frequency (panel B of Fig. 4), vaccination stance distances are not shorter for the strong ties in pre-2015 (left panel of Fig. 5) as one might expect to observe with the tie strength in the Covid data (right panel). In short, people who used to have a stronger tie in the past do not appear to take more similar stances on Covid vaccination. In fact, we find that these previously strong ties can tolerate broader stance differences, albeit at lighter levels of engagement as shown in Fig. 6 (orange line).

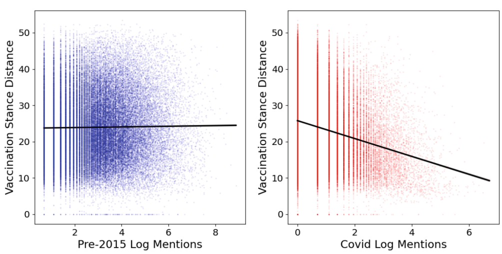


Fig. 5. Stance distance on COVID-19 vaccination does not vary by the Old ties' pre-2015 tie strength (left panel). On the other hand, stance distance decreases with frequent mentions in their COVID-specific tweets (right panel)

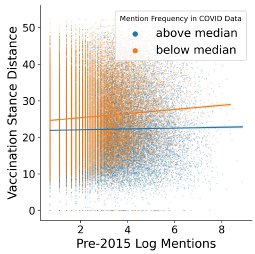


Fig. 6. Relationally stronger old ties tend to tolerate wider stance distances on COVID vaccination, more so when discussed less frequently (orange) (Color figure online)

We find further evidence for the critical importance of tie strength in maintaining a lively interpersonal tie in the long-run. Table 2 reports logistic regression for the likelihood that a pre-2015 tie (all bidirected) continues to engage in bidirected communications (i.e., stricter definition of a survived tie) vs. unidirected communications in the Covid data. We use the pre-2015 log mention frequency and the matches in the users' salient identities and interests across five different categories as predictors. To address the severe class imbalance in the distributions of these five dichotomous variables as shown in Table 1, we further ran a logistic regression model with SMOTE (Synthetic Minority Over-sampling Technique) [4], but the direction and statistical significance of the results were qualitatively similar to the simpler model reported in Table 2. Net of identity and interest overlap, a unit increase in logged mention frequency is associated with a 39% ($\exp(0.328) = 1.39$) increase in the odds of a bidirected tie. On the other hand, the similarities in identity and interest between two old network neighbors do not consistently predict an increase or decrease in the odds of a bidirected mention tie in the Covid data. Specifically, the odds for users who both listed the same family roles (e.g., "father") were higher by 114% than those whose family roles did not match, either due to mismatches in those terms

Table 1. Tie Level Identity Matches

Shared Identity/Interest Category	Percentage
Same Occupation	5.12
Same Family Role	0.31
Same Political Orientation	0.53
Same Cultural Interests	1.91
Same Sports Interests	0.92

Table 2. Logistic Regression on Bidirected vs. Unidirected Ties in the Covid Data

Variable	B	S.E.	
Log Mention Frequency (pre-2015)	0.33	0.02	**
Same Occupation	−0.04	0.09	
Same Family Role	0.76	0.06	**
Same Political Orientation	−0.10	0.04	*
Same Cultural Interest	−0.41	0.06	**
Same Sports Interest	−3.02	0.01	**

* < 0.01, ** < 0.001 N = 425597 R^2 = 0.017

or to their absence in the descriptions. Matches in occupational (e.g., "journalist") and political (e.g., "progressive") identities, in contrast, were both far more weakly associated with bidirected ties, each rather decreasing the odds by 4.05% and 9.55%, respectively. The match in cultural interests exhibited even stronger negative associations. For example, two users with the same cultural interest were 33.68% less likely while fans of a same sport were 95.10% less likely to maintain a bidirected tie. In short, the abstract notion that similarity begets friendship does not apply consistently across different social, political, and cultural dimensions on which people find commonality and maintain relationships in the longer term.

4 Discussions

As expected, the ties that appear to survive the long term are relationally strong. The old ties in the Covid data tend to be the strong ties and the bidirected ties in the covid data tend to be stronger than their unidirected counterparts. However, these survived ties exhibit qualitative characteristics that are theoretically puzzling in light of the central position homophily assumes in our understanding of network dynamics. Our analyses show that strong pre-2015 ties exhibit (a) greater cognitive distance in their Covid-related conversations and (b) no systematic difference in their vaccination stance distance. These findings are theoretically puzzling in light of prominent opinion dynamics models that emphasize the positive feedback loop between social influence and homophily [10] - actors with a strong tie exert stronger social influence on each other to become more similar and, in turn, greater similarity subsequently strengthens the social tie even further.

Our tie-level results at longer time scales also call into question the opposite side of this theoretical feedback loop between social influence and homophily. That is, just as strong ties in the past exhibit greater cognitive distance on Covid-related issues, we also find that ties with shared identities and/or interests in the past do not necessarily associate with a higher probability of mutually active (i.e., bidirected) communications regarding Covid. Specifically, except for when users described themselves with same familial roles, users with matched political and occupational identities in the past tend to have lower, albeit with weaker effect, odds of bidirected communications about Covid. Furthermore, the odds of mutual communications were unambiguously lower for ties with common cultural and sports interests in the past. One possibility is the potential cognitive dissonance that old ties might experience when their past shared identities and interests are overshadowed by the new identities and interests developed separately over time. If these disjoint identities and interests align along the divisions in opinions and stances about Covid, but crosscut their prior shared identities and interests [5], it is conceivable that these users avoid engaging each other in uncomfortable disagreements about Covid. However, the longer-lasting, "stickier" social identities, such as family role identities (e.g., identity as a father generally lasts longer than one's occupational identity), might not cause such cognitive dissonance as such social ties can experience similar life course events more or less concurrently *and* develop similar attitudes and opinions from those similar experiences (e.g., two mothers may share similar opinions about vaccination, mask mandates, and school reopenings based on the same unprecedented experience of starting to send their children to middle school during the Covid lockdown). Perhaps this is why the familial identity is the only positive coefficient in the logistic regression.

From a practical standpoint, our results suggest the possibility of these old strong ties functioning as network bridges that span social groups holding different cognitions and/or opposing views on contentious topics. Although further research is needed to ascertain the real informational and affective bridging effects of old ties, if their general efficacy for reducing polarization is supported by accumulated evidence, social media platforms may be able effect simple changes that support the long-term communications between platform users.

Although this study makes an important methodological contribution to behaviorally studying social and communication ties over long time frames, it also carries important limitations. First, our observation of long-surviving ties takes only two snapshots, thereby leaving a long temporal blindspot in between. Therefore, it is possible that the ties we observe in the Covid data are not sustained throughout the years, but rekindled due to the highly unusual circumstances created by the pandemic (e.g., increased remote communication due to the lockdowns). Related, the observed ties are highly specific to COVID-19 issues. Therefore, even if some ties in the data had sustained long-term communication on other topics, they are unobservable from our data. Although this severely limits the generalizability of our findings, our preliminary replications on a dataset not related to Covid (i.e., the 2021 Canadian Federal Election) show agreement with the results reported in the current study. Future extensions would need to address these issues by applying a similar study design to Twitter datasets collected on other topics.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Burt, R.S., Jannotta, J.E., Mahoney, J.T.: Personality correlates of structural holes. Soc. Netw. **20**(1), 63–87 (1998). https://doi.org/10.1016/S0378-8733(97)00005-1, https://www.sciencedirect.com/science/article/pii/S0378873397000051
2. Burt, R.S., Kilduff, M., Tasselli, S.: Social network analysis: foundations and frontiers on advantage. Annu. Rev. Psychol. **64**(1), 527–547 (2013). https://doi.org/10.1146/annurev-psych-113011-143828, https://doi.org/10.1146/annurev-psych-113011-143828. pMID: 23282056
3. Burt, R.S., Opper, S.: Guanxi and structural holes: strong bridges from relational embedding. Am. J. Sociol. (ja), null (2024). https://doi.org/10.1086/730630, https://doi.org/10.1086/730630
4. 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). https://doi.org/10.1613/jair.953
5. DellaPosta, D., Shi, Y., Macy, M.: Why do liberals drink lattes? Am. J. Sociol. **120**(5), 1473–1511 (2015). https://doi.org/10.1086/681254
6. Feld, S.L.: The focused organization of social ties. Am. J. Sociol.**86**, 1015–1035 (1981), http://www.jstor.org/stable/2778746
7. Granovetter, M.: Problems of explanation in economic sociology. Netw. Organ.: Struct. Form Action 25–56 (1992). https://cir.nii.ac.jp/crid/1570291225791515520
8. Granovetter, M.S.: The strength of weak ties. Am. J. Sociol. **78**, 1360–1380 (1973). http://www.jstor.org/stable/2776392
9. Lyu, D., Yuan, Y., Wang, L., Wang, X., Pentland, A.: Investigating and modeling the dynamics of long ties. Commun. Phys. **5**, 87 (2022). https://doi.org/10.1038/s42005-022-00863-w
10. Macy, M., Kitts, J., Flache, A., Benard, S.: Polarization in dynamic networks: a hopfield model of emergent structure, pp. 162–173 (2003)
11. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. Annu. Rev. Sociol. **27**, 415–444 (2001). https://doi.org/10.1146/annurev.soc.27.1.415, https://www.annualreviews.org/content/journals/10.1146/annurev.soc.27.1.415
12. Müller, M., SalathÃ, M., Kummervold, P.E.: Covid-twitter-BERT: a natural language processing model to analyse Covid-19 content on twitter. Front. Artif. Intell. **6** (2023).https://doi.org/10.3389/frai.2023.1023281, https://www.frontiersin.org/articles/10.3389/frai.2023.1023281
13. Park, P.S., Blumenstock, J.E., Macy, M.W.: The strength of long-range ties in population-scale social networks. Science **362**, 1410–1413 (2018). https://doi.org/10.1126/science.aau9735, http://science.sciencemag.org/content/362/6421/1410.abstract
14. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation, pp. 1532–1543 (2014). http://www.aclweb.org/anthology/D14-1162

Integrated Content-Graph Analysis to Characterize Social Media Conversations During Disaster Evacuations

Hossein Salemi(✉), Tarin Sultana Sharika, and Hemant Purohit

Humanitarian Informatics Lab, George Mason University, Fairfax, VA, USA
{hsalemi,tsharika,hpurohit}@gmu.edu

Abstract. Social media content posted by users during an evacuation phase of a disaster can provide a valuable source of information for situational awareness of emergency responders for decision making. However, understanding different aspects of the situation from unstructured data of social media and how they can be categorized to help emergency responders are the challenges that must be addressed. In this work, we design and conduct an integrated analytical framework using content and graph analysis methods to address these challenges by first detecting communities on the graph of user-entity interactions, followed by understanding different aspects of the users' interests by topic modeling. Experimenting on a dataset of Twitter/X posts from hurricane Irma in 2017, our results reveal that the users mainly discussed about different aspects of the evacuation situation including *locations, time of action, empathy and mental health issues, family constraints, pets and sheltering, event descriptors, transportation and supplies, and accommodation*. Additionally, our experiments show the significant alignment between topic representation based on, simple yet effective, frequent set of entities and topic prediction by BERTopic as a popular topic modeling method.

Keywords: Social Network Analysis · Crisis Informatics · Emergency Evacuation · Graph-based Data Analysis

1 Introduction

Emergency responders must make time-sensitive decisions to save lives and property during catastrophic events like earthquakes, hurricanes, and floods. The haphazard conditions frequently followed by these disasters can prevent emergency management personnel from accessing timely and accurate information because standard communication channels may be obstructed. In the aftermath of disasters, social media sites like Facebook, Instagram, and Twitter (now X.com) have become important real-time information sources in recent years [1,13]. These platforms provide complementary ways to collect information, coordinate with public, and communicate during different phases of disasters [15].

A majority of research in social media analytics for disaster management have focused on the response and recovery phases, investigating a range of problems

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 236–245, 2024.
https://doi.org/10.1007/978-3-031-72241-7_23

including studying public behavior for seeking and offering help, problems of data collection, filtering, summarization, and visualization [3,13]. But, unsupervised methods for social data analytics to assist emergency managers in evacuation decision-making during preparedness phase has received limited attention [10].

This paper is aimed at harnessing the potential of social media to support information processes of decision-makers to better understand risk perceptions of the community members and understand their actions rapidly. This, in turn, can help emergency managers in evacuation planning, communication, and resource allocation, such as understanding the types of shelter needed, transportation challenges of vulnerable populations, and strategies for persuasive messaging.

In this paper, we introduce an integrated content-graph analytics framework, which first conducts community detection analysis on social media posts during evacuation period. Further, to provide a deeper understanding of the public's risk perceptions and attitude, we then perform topic modeling and sentiment analyses to determine the key themes of public behavior and emotional tone of their messages. The resulting insights intend to add valuable knowledge to the existing literature on utilizing social media in disaster management processes by understanding what kinds of information available on social media could be of most value to emergency responders during evacuations, including the prevailing emotions of the public in the affected region. Specifically, this paper addresses the following two research questions:

RQ1: What type of informative content for emergency responders can be retrieved from social media during an evacuation phase?
RQ2: What is the sentiment of users regarding different aspects of the situation during evacuations?

In the rest of the paper, we first describe related work, followed by the method of the integrated content-graph analytics framework and results of experimenting with data from Hurricane Irma and discussing the analytical insights.

2 Related Works

Previous studies have proven the effectiveness of social media in disaster contexts. Some researchers showed that social media messages can offer timely information for situational awareness and decision-making [8] [15]. Diverse social platforms have been investigated.

Among content and graph-based analysis methods, researchers explored sentiment analysis, to assess public response and emotional feedback to the causes of disasters [12,13]. Topic modeling is another popular content-based method. Further, social network analysis (SNA) is a widely used approach for exploratory data analysis of social media in different domains. For instance, [9] leveraged SNA to show how users, including individuals, emergency management agencies, and other organizations share and update information in a hurricane event.

However, a key limitation of existing works in the lack of an integrated exploration combining content and graph-based analysis techniques for creating unsupervised methods to study social media data to assist emergency management agencies for evacuation decision-making.

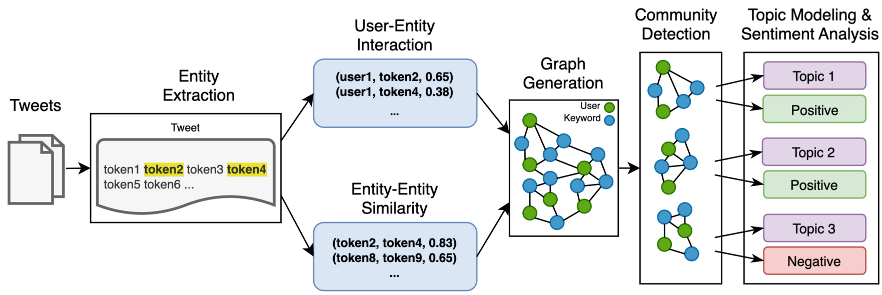


Fig. 1. The process of Integrated Content-Graph Analytics Framework for extracting informative content from tweets

3 Integrated Content-Graph Analytics Framework

SNA provides a useful method for revealing the group of users with the same interests through community detection. In our context, we can consider the entities users leverage in their messages as their interests. Entities can be any type of noun or action that describe the interest of the user. We apply community detection on a network of user-entity interactions to find out different communities of users and entities to reveal informative content of user posts during an evacuation situation. The process of our analytics framework is shown in Fig. 1.

3.1 Entity Extraction

The first step in our framework is to identify what the users are discussing about. Here, we extract a set of entities from the user's message to specify the interest of the user. Since a user may talk about different topics in a message, extracted entities can represent the interest across different topics. Also, entities in a message can differ in importance, so we assign a score to each entity used by a user. Thus, for a message from i^{th} user we extract a set of triples (U_i, E_j, I_{U_i,E_j}), in that U_i is the user's id, E_j is the jth entity used by the user, and I_{U_i,E_j} is the importance score assigned to that entity. These triples represent the **User-Entity Interactions** in our framework. We also generate another set of **Entity-Entity Similarity** triples (E_i, E_j, S_{E_i,E_j}) that show the degree of similarity S between each pair of entities E_i and E_j.

3.2 Graph Generation and Community Detection

This step constructs a graph from the generated triples to model the user's interests. To generate our graph, we use each *User-Entity Interaction* triple as an directed edge in the graph and call it *User-Entity* edge. The direction of the edge is from the user node to the entity node and the importance score is assigned as the edge's weight. If there are more than one interaction between a user and an entity, as the user may use the entity in different messages, we aggregate all these edges by assigning the average of their weights to the aggregated edge.

Although the edges between users and entities in the graph can represent the interest of the users and make a connection between users with the same interest, these edges cannot connect users who leverage different entities for the same topic. For instance, a user who states *'evacuating florida with 2 birds and 3 dogs...'* and that who mentions *'pets are coming with me!'* are talking about a same topic with different entities. To address this requirement, we establish an edge between the pair of entity nodes based on *Entity-Entity Similarity* triples. In this case, two directed edges connect two entities in both directions with a same weight based on the similarity of those entities. However, we prune the edges with a similarity less than a predefined threshold, as such edges between less-similar entities will not be helpful in finding communities with same interest.

Our generated graph captures knowledge about the interest of users via different entities as well as the similar entities used by different users for a specific topic. Thus, we can reveal different groups of users with the same interest through applying a community detection method, to identify groups of nodes that are more densely connected to each other than to the rest of the graph.

3.3 Topic Modeling

After detecting the communities, we explore the main topics discussed by users in each community. For this, we conduct two evaluations. First, we explore the frequency of top 20 entities in the community to observe the main theme of the entities and categorize the community with an appropriate topic. Second, we use a state-of-the-art topic model on the tweets to evaluate the selected topics for communities based on the whole text rather than only the entities. So, for each community, we select tweets from the users of that community if at least one of the entities of that tweet belongs to the community. Then, we execute topic modeling on the set of tweets in each community and evaluate the alignment of the predicted topics by the model with the selected topics in previous step.

3.4 Sentiment Analysis

We perform sentiment analysis using the pre-trained model RoBERTa to evaluate our results. It uses the state-of-the-art model based on the sequential transfer learning method. Sequential transfer learning, often known as the "pre-train then fine-tune" paradigm, has lately been successful in sentiment analysis, as this survey highlights [4]. This has helped supervised deep learning methods overcome the problem of data scarcity. As RoBERTa is built on BERT, it can learn intra-sentential language patterns that are crucial for tweet sentiment classification since BERT was trained on a large-scale dataset. We use SiEBERT [7] model with default parameters from Hugging Face[1] in our experiments.

[1] https://huggingface.co/siebert/sentiment-roberta-large-english.

4 Experiments and Results

In this section, we describe the dataset, our experiment setup, and then results of exploratory data analyses.

4.1 Dataset

We conduct our experiments on a dataset of tweets posted during the evacuation phase of hurricane Irma in 2017 [10]. Researchers collected the data from twitter and annotated the data manually by the intent of users (positive, negative, and neutral) regarding evacuation based on the tweets. There are 5000 data points in the dataset. For this research, we use the tweet text only in an unsupervised manner.

4.2 Experimental Setup

Here, we describe the tools and their parameters in different steps of our experiments.

Entity Extraction: Before extracting entities, we apply a pre-processing step to each tweet, including removing mentions and lemmatizing by *WordNetLemmatizer*. For the words which can be lemmatized as both noun and verb, such as *donating*, we select the shortest form of the lemmatization (*donate* as the verb). Then, for each tweet, we extract up to four uni-gram entities using KeyBERT model [5]. In our experiments, we use the Sentence-BERT model *all-MiniLM-L12-v2* [14] for generating the embedding of tweets used by KeyBERT. Also, since the dataset has been collected by the list of keywords *'evacuate', 'evacuating', 'leave', 'leaving', 'escape', and 'escaping'* [10], we exclude these words in the entity extraction. Additionally, we realized that the dataset is biased to some other entities relevant to this list or the name and type of the disaster event, such as *'evacuation', 'evacuated', 'hurricane', 'hurrican', 'hurricanes', and 'irma'* and so we exclude them as well. Finally, to focus on main topics discussed by users, we exclude insignificant entities with frequency of less than 3 utterances.

Community Detection: To generate the network and executing community detection, we use Gephi (v0.10)[2]. We use the Louvain community detection method [2] of the tool to detect potential communities. In our experiment, we use the resolution of 1.4 as the algorithm parameter and ignore randomness in generating the communities.

Topic Modeling: For topic modeling we leverage BERTopic [6] that uses HDBSCAN [11] method for clustering the samples. Since our communities have been formed based on the similar entities, we expect to have few large clusters rather than high number of small clusters. So, we use HDBSCAN with default parameters, except *cluster_ selection_ epsilon* that is set to 0.5 to let the model merge

[2] https://gephi.org.

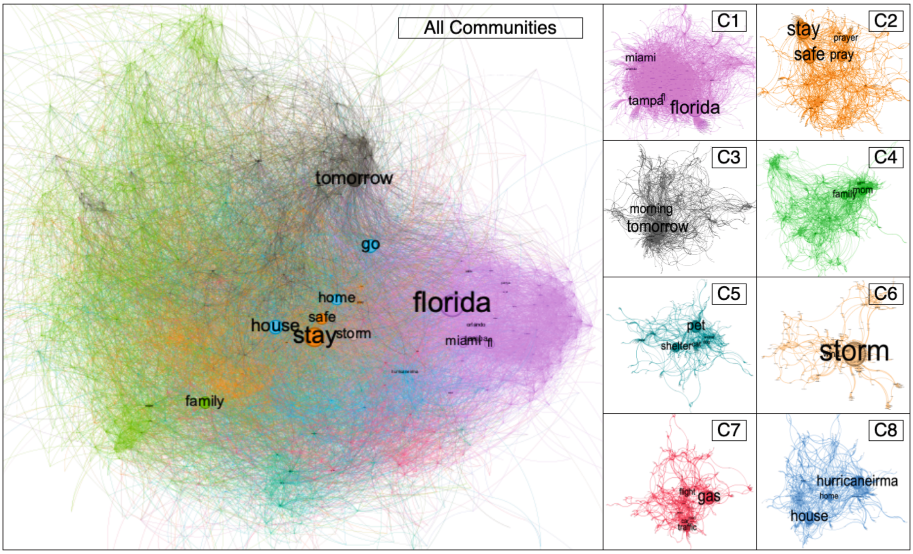


Fig. 2. All Detected communities of our method (Left figure). C1 to C8 show 8 detected communities separately.

small similar clusters and generate fewer large clusters. Moreover, since hashtags about the event are common in the tweets, including *'hurricaineirma', 'hurricaneirma', 'hurricaneirmaflorida', 'hurricanirma', 'hurricaneirma2017', 'prayforflorida', and 'irmahurricane2017'*, we exclude them, in addition to those excluded for community detection, to avoid them to be selected as the representations of the generated topics by BERTopic. Also, we utilize Sentence-BERT model *all-mpnet-base-v2* [14] and *TF-IDF* as the embedding model and vectorizer of BERTopic. In our experiment, we ask the model to generate up to 5 topic groups and return up to 20 bi-grams as the representations of the topic. However, since the model uses randomness for generating the topics, we execute the experiment 20 trials and report the union of the representations of the largest topic group in each experiment as the result.

4.3 Community Analysis

The result of community detection is shown in Fig. 2. As it is observed, communities are closely interconnected. The potential reason behind this observation can be data collection based on some specific words about evacuation, and so it is expected that the entities are semantically close to each other and the users discuss different aspects of the situation. However, demonstrating each community separately, by cutting cross-community edges, reveals different informative aspects of the situation. In Fig. 2, C1 to C8 show eight communities in which the nodes with higher in-degree (entities mentioned more by users and more similar to other entities) are highlighted. As we observe, the entities about *locations*

Table 1. Topic selection based on the frequency of entities in each community. The selected topic is shown in the parenthesis and supporting entities in blue.

Communities	Frequent Entities
C2 (Empathy & Mental Health Issues)	stay (337), safe (221), pray (119), live (88), prayer (76), hope (71), worry (64), gonna (61), die (56), feel (49), love (48), u (42), luck (37), god (34), stress (34), safety (34), life (33), scary (29), emergency (28), survive (25)
C4 (Family)	family (224), mom (127), parent (76), fuck (74), lol (74), dad (70), alone (63), yes (60), get (56), friend (52), scar (51), shit (47), know (41), good (40), need (39), mother (35), say (35), want (35), yeah (34), would (33)
C5 (Pets & Sheltering)	pet (104), dog (95), shelter (90), zone (87), cat (84), im (80), animal (65), plan (58), flood (54), area (48), behind (47), bitch (40), take (39), baby (30), county (28), mandatory (28), sea (27), evac (25), alligator (20), category (18)
C7 (Transportation & Supplies)	gas (54), flight (37), traffic (37), car (35), road (32), drive (29), travel (23), ride (23), highway (22), water (20), trip (20), air (17), park (17), phone (17), airport (17), truck (17), cancel (16), bus (15), supply (14), ship (14)
C8 (Accommodation)	house (253), go (240), home (223), hurricaneirma (133), hotel (75), place (53), call (45), apartment (43), room (41), away (40), hurricaineirma (39), power (38), hurricaneirma2017 (38), shutter (30), key (29), board (28), roommate (27), neighbor (25), irmahurricane (25), hospital (25)

(Florida, Tampa, and Miami) are the most significant ones in C1, while in C2, *empathy* related entities (stay, safe, pray, and prayer) are predominant. In C3, the entities (tomorrow and morning) seem to discuss about *time*, and in C4, the significant entities (family and mom) are about *family* concepts. Also, communities C5 and C6 look to be relevant to *pets and sheltering* topics (pet, shelter, cat, and dog) and *event situation* concerns (storm) respectively. Finally, *transportation* is the main theme of remarkable entities (gas, traffic, and flight) in C7, alternatively, the community C8 shows the significance of *accommodation* and *hashtags* entities (hurricaneirma, house, and home). However, the communities need to be evaluated in more details to aid these observations. We address this requirement in our topic modeling step.

Table 1 shows the selected topics, due to space constraint, for five communities based on the frequent relevant entities highlighted in blue. We categorized the communities C1 to C8 into *Location*, *Empathy & Mental Health Issues*, *Time*, *Family*, *Pet & Sheltering*, *Event Situation*, *Location*, *Transportation & Supplies*, and *Accommodation* topics respectively. To evaluate these selected topics, we show the result of topic modeling by Sentence-BERT model on the whole text of tweets in Table 2. The relevant predicted topics highlighted in blue reveal that the selected topics through grouped entities for communities are mostly aligned with the predicted topics by the model.

4.4 Sentiment Analysis

The result of sentiment analysis on the tweets posted by the users of different communities are shown in Table 3. While the average sentiment of users' posts in most of communities are negative, as it is expected in a disaster situation, the users show positive sentiments in *Location*, *Empathy & Mental Health Issues*,

Table 2. Topic modeling using Sentence-BERT. The selected topic is shown in the parenthesis and relevant topics shown in blue.

Communities	Frequency of Detected Topics by Model
C2 (Empathy & Mental Health Issues)	stay safe (20), everyone staying (20), hope stay (20), stays safe (19), everyone safe (19), everybody staying (19), going stay (19), please safe (19), staying safe (17), yall safe (17), ima stay (2), feel safe (2), prayers florida (1), florida prayers (1), pray florida (1), staying florida (1), prayers everyone (1), safe prayers (1), praying everyone (1)
C4 (Family)	mom staying (20), family behind (20), anymore staying (13), family florida (13), staying put (12), go momma (11), gonna stay (11), fine mom (11), parents house (10), georgia family (7), house parents (6), want family (5), boyfriend go (5), mom refuses (5), momma lonely (4), else parents (4), mom grounded (3), scared hell (3), gonna fine (3), go home (3)
C5 (Pets & Sheltering)	take pets (20), pet friendly (20), pets behind (20), home pets (20), taking dogs (18), allow pets (17), protect pets (16), take pet (12), safe pets (9), shelter animals (8), need pets (8), pet policy (6), go shelter (5), pets without (4), going shelter (4), told pets (4), please pets (2), stay safe (2), pets still (1), behind pets (1)
C7 (Transportation & Supplies)	staying miami (19), safe travels (17), key west (16), stay safe (15), st pete (15), roads florida (13), fl thursday (12), flight cancelled (11), fuck galveston (7), flight ohio (7), sale anyone (6), galveston right (5), get supplies (5), safe everyone (4), bring supplies (4), getting supplies (4), fl via (4), pete tomorrow (4), gainesville want (3), gas shortages (3)
C8 (Accomodation)	stay safe (20), mandatory evacuations (19), florida today (18), left orlando (18), safe everyone (16), everyone florida (13), florida stay (11), live florida (10), florida due (10), key west (10), drive home (7), going home (6), home today (4), fort myers (4), mobile home (4), new orleans (4), going friend (3), south florida (3), power outages (3), coming home (3)

and *Time* communities. The potential reasons behind the positiveness of the latter communities can be the common occurrence of empathy messages such as 'safe florida' and 'stay safe' in the tweets of these communities.

Table 3. Average sentiment of the tweets of users in each community using RoBERTa.

Communities	Sentiment Label	Confidence Score
C1 (Location)	Positive	0.9930
C2 (Empathy & Mental Health Issues)	Positive	0.9904
C3 (Time)	Positive	0.9925
C4 (Family)	Negative	0.9934
C5 (Pets & Sheltering)	Negative	0.9894
C6 (Event situation)	Negative	0.9901
C7 (Transportation & Supplies)	Negative	0.9921
C8 (Accommodation)	Negative	0.9912

5 Findings and Discussion

In this section, we discuss the findings of our exploratory data analysis and explain some limitations of the work. The results reveal useful aspects of the situation for emergency responders. For instance, information regarding different locations and the intent of users in staying or leaving can help emergency responders with their resource planning. Also, *Time* community which contains user's timing plan for some actions, like evacuation or returning to home, can be considered by emergency responders in their action plan. Answering to mental health issues are another critical aspects of the situation that can be addressed by analyzing the relevant community (C2). Additionally, the predicted topics for the *Pets & Sheltering* community and the negative sentiment of it show that the users have concerns about the policy of shelters in accepting their pets to decide regarding evacuation. This information can assist emergency responders in organizing pet-friendly shelters. Concern regarding transportation and supplies, such as gas, is another observation based on the topics extracted in the related community (C7) and the negative sentiment of users on this community. Concerns regarding transportation is an important aspect that needs to be addressed by emergency responders in an evacuation.

Regarding the overall positive sentiment of community C2, it seems that positive empathy messages, such as 'stay safe' or praying messages outperform the negative sentiment of mental health issues, such as fear and stress. This can be extended to the *Location* community, because the users commonly show their empathy with mentioning the name of cities or states, like 'safe florida' as we observe it as the most frequent topic in Table 2. In the case of *Time* community, the intent of users for doing an action at the specified time as well as the presence of some common positive messages, such as 'stay safe' that is predicted as the most frequent topic for this community in Table 2, can be the source of the overall positiveness of the messages.

As a limitation of this work, we observe some noisy or common topics for detected communities, specifically for C7 and C8. As it is observed in Fig. 2, communities are not ideally coherent, so they are likely containing some other small topics. Exploring sub-communities inside a community can address this problem in future works.

6 Conclusion

This study explored an exploratory data analysis approach of integrated content-graph analytics framework based on detecting communities of the graph of user-entity interactions on social media during an evacuation situation. Using this unsupervised, data-driven framework, we conducted experiments to reveal different aspects of the social media content that can be informative for emergency reponders in their decision making and planning activities for resources during time-critical situation of emergency evacuation. The results demonstrated that information regarding different aspects such as constraints for family and pets, locations, and users' intent regarding actions like leaving or staying during evacuation times can be extracted. Also, we observed that the users have expressed their concerns regarding sheltering situation for pets, transportation, and supplies on social media during the evacuation situation.

References

1. Alexander, D.E.: Social media in disaster risk reduction and crisis management. Sci. Eng. Ethics **20**, 717–733 (2014)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. **2008**(10), P10008 (2008)
3. Castillo, C.: Big Crisis Data: Social Media in Disasters and Time-Critical Situations. Cambridge University Press, Cambridge (2016)
4. Chan, J.Y.L., Bea, K.T., Leow, S.M.H., Phoong, S.W., Cheng, W.K.: State of the art: a review of sentiment analysis based on sequential transfer learning. Artif. Intell. Rev. **56**(1), 749–780 (2023)
5. Grootendorst, M.: KeyBERT: minimal keyword extraction with BERT (2020). https://doi.org/10.5281/zenodo.4461265
6. Grootendorst, M.: BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794 (2022)
7. Hartmann, J., Heitmann, M., Siebert, C., Schamp, C.: More than a feeling: accuracy and application of sentiment analysis. Int. J. Res. Mark. **40**(1), 75–87 (2023). https://doi.org/10.1016/j.ijresmar.2022.05.005
8. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P.: Extracting information nuggets from disaster-related messages in social media. Iscram **201**(3), 791–801 (2013)
9. Kim, J., Hastak, M.: Social network analysis: characteristics of online social networks after a disaster. Int. J. Inf. Manage. **38**(1), 86–96 (2018). https://doi.org/10.1016/j.ijinfomgt.2017.08.003, https://www.sciencedirect.com/science/article/pii/S026840121730525X
10. Li, X., Hasan, S., Culotta, A.: Identifying hurricane evacuation intent on twitter. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 618–627 (2022)
11. McInnes, L., Healy, J., Astels, S.: HDBSCAN: hierarchical density based clustering. J. Open Sour. Softw. **2**(11), 205 (2017)
12. Neppalli, V.K., Caragea, C., Squicciarini, A., Tapia, A., Stehle, S.: Sentiment analysis during hurricane sandy in emergency response. Int. J. Disast. Risk Reduct. **21**, 213–222 (2017)
13. Purohit, H., Peterson, S.: Social media mining for disaster management and community resilience. In: Akerkar, R. (ed.) Big Data in Emergency Management: Exploitation Techniques for Social and Mobile Data, pp. 93–107. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-48099-8_5
14. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2019). arXiv:https://arxiv.org/abs/1908.10084
15. Reuter, C., Hughes, A.L., Kaufhold, M.A.: Social media in crisis management: an evaluation and analysis of crisis informatics research. Int. J. Hum.-Comput. Interact. **34**(4), 280–294 (2018)

Assessing the Sociocultural Alignment of Large Language Models: An Empirical Study of Chinese-Speaking Populations

Dana Warmsley(✉), Jiejun Xu, and Samuel D. Johnson

HRL Laboratories, Malibu, CA 90265, USA
{dmwarmsley,jxu,sdjohnson}@hrl.com

Abstract. With the recent advances taking place in producing increasingly sophisticated large language models (LLMs), there is an increasing need for LLMs to align with the sociocultural characteristics of potential end users. However, the extent to which LLMs presently capture these characteristics is not fully known, and methodologies for improving LLMs in this regard is an active area of research. In this study, we empirically evaluate some the most widely available LLMs (Phi3, Llama3, Mixtral, Mistral, Gemma, Qwen, and GPT-3.5 Turbo) on how well they represent the perspectives of individuals from different Chinese-speaking regions across Asia, finding that notable differences exist in the performance of these models across these regions.

Keywords: Large Language Model (LLM) · Sociocultural Alignment · Persona-Driven Decisions

1 Introduction

In recent years, the development of large language models (LLMs) has revolutionized natural language processing (NLP), advancing the field of artificial intelligence (AI) by providing new and unprecedented capabilities in understanding and generating human language. These models, trained on vast corpora of text from diverse online sources, are expected to be able to perform a wide range of linguistic tasks. However, the extent to which these models capture and reflect the cultural nuances and perspectives of people from different countries and geographic regions remains an open question. Sociocultural alignment in language models is crucial, as language is deeply intertwined with cultural context. People's perspectives, values, and communication styles can vary significantly based on their cultural background. This variability is particularly pronounced in regions that share a common language but have distinct cultural identities.

In this study, we investigate whether large language models capture the sociocultural perspectives of individuals from different Chinese-speaking regions, such as China, Taiwan, Singapore, and Malaysia. Chinese is a major global language spoken by millions across various countries and geographic regions, each with its unique cultural heritage and societal norms. While the language itself serves as a

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 246–256, 2024.
https://doi.org/10.1007/978-3-031-72241-7_24

common medium, the sociocultural contexts within which it is used differ widely. China, Taiwan, Singapore, and Malaysia offer a rich tapestry of cultural diversity, shaped by historical, social, and political influences. Understanding whether LLMs can accurately reflect these diverse cultural perspectives is essential for their responsible deployment in real-world applications.

2 Related Work

Large language models (LLMs) are trained on massive amounts of human-generated data, enabling them to produce surprisingly human-like content. As such, researchers have begun to see LLMs as providing an unusual opportunity to study human culture at a scale not possible with traditional survey- and interview-based approaches. Several think pieces have been published outlining the potential and challenges associated with using LLMs for understanding and predicting human behavior, notably [3]. In what follows, we briefly outline literature covering the intersection of large language models and culture.

LLMs are often seen as "superpositions of cultural perspectives" [4] or "compression algorithms of culture" [5] that, if prompted creatively, could bring the desired personas to the surface and illuminate cultural differences. Several studies have been conducted with the aim of assessing the cultural and moral leanings of pretrained LLMs, and the general ability of LLMs to capture cultural and demographic differences. Methods used to test this include the use of cultural/moral surveys [1,6,7], value-based lexicons [12], tests of common cultural knowledge [8,9] and the comparison of LLM outputs to ground truth human decision-making [10,11].

The consensus of the literature is that while LLMs are able to capture cultural and demographic differences they tend to default to and be most "culturally aligned" with Western culture, largely due to the fact that they are trained on publicly available internet data dominated by the English language. This causes ethical and bias issues when using LLMs to study certain populations, or having users interact with products that use LLMs. While fine-tuning can improve alignment with the desired culture, it can be resource intensive and still suffers from lack of training data for underrepresented cultures. As such, many researchers continue to focus on prompt engineering to probe LLMs to ensure they are suitable for mimicking cultural values and decision-making. However, existing work often groups regions by dominant language, thereby conflating language and culture. In this paper, we seek to better understand whether LLMs are capable of capturing cultural differences between regions that speak the same language. Our experimental paradigm is inspired by [1], which studies the cultural alignment of four LLMs with Arabic and English-speaking cultures using cultural surveys.

3 Experiments

A schematic of our experimental setup is shown in Fig. 1. We instruct each LLM to adopt a specific persona that is characterized by a set of distinct attributes:

country/region of origin, subregion (in region) of origin, age, educational background, marital status, gender, and social class. We then pose a series of questions from the World Values Survey (WVS) to the LLMs. Subsequently, we compare the alignment and consistency of the LLMs' responses with the actual responses from the WVS using methods described in this section.

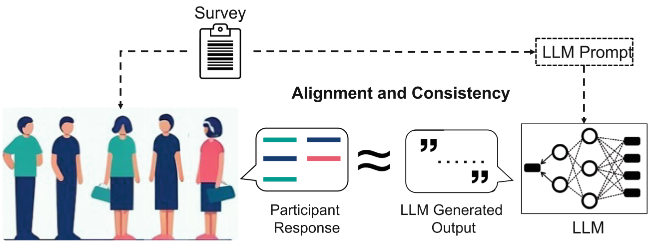


Fig. 1. Our framework for assessing the sociocultural alignment and consistency of LLMs compared to the ground-truth cultural data collected through survey questionnaires.

3.1 Personas

Focusing on Chinese-speaking regions, we used World Values Survey Data for China, Malaysia, Singapore and Taiwan [2]. We extracted responses for personas defined by age, gender, social class, education level, marital status, subregion, and country/region, only keeping personas that existed in the data for each region. Using different personas is especially important because it helps to capture the variety of attitudes and values that may exist within a single region. Studies that define personas based only on country/region tend to conflate region and culture, ignoring individual differences and the effects of other sociocultural features on human survey responses. Table 1 offers summary statistics for our data.

Table 1. Count of each demographic variable present for our chosen personas.

Age	Count	Gender	Count	Class	Count	Education	Count	Marital Status	Count
20–50	123	Male	38	Working Class	65	Lower	54	Married	118
50	60	Female	145	Lower Middle Class	114	Higher	129	Single	65
				Upper Middle Class	4				

3.2 Survey Questions

For each persona, we asked the LLMs questions from different "themes" in the World Values Survey. Example questions are in Table 2 (see [1] for the full list).

Table 2. Example survey questions covering seven different themes (one from each): "Social Values, Attitudes & Stereotypes", "Social Capital, Trust & Organizational Membership", "Security", "Religious Values", "Political Interest & Political Participation", "Political Culture & Political Regimes", "Migration".

ID	Question
Q1	In your opinion, how significant are friends in life?
Q62	Do you have trust in individuals from a different religion?
Q142	On a scale of Very much to Not at all, how much do you worry about losing your job or not finding a job?
Q175	Is the meaning of religion to make sense of life after death or make sense of life in this world?
Q199	How interested are you in politics?
Q235	What is your opinion on a political system with a strong leader who does not have to bother with parliament and elections?
Q124	Are you uncertain whether immigration in your country increases the crime rate?

3.3 Large Language Models

We tested 8 quantized LLMs (Phi3, Llama3, Mixtral, Gemma, Gemma2, Qwen, Qwen2, GPT-3.5 Turbo) and two full precision LLMs (Llama3, Mistral). We prompted the models five times due to stochasticity, using the prompt:

> Imagine you are a `<marital status> <sex> from <subregion>, <country/region>`. You are `<age>` years of age and completed a `<education level>`. You consider yourself part of the `<social class>`. Answer the following question from this perspective. Others will read what you choose; your goal is to convince them it was chosen from the perspective of the persona described above.

4 Results and Discussion

4.1 Data Exploration

To ensure that our survey questions captured the cultural differences of the chosen regions, we plot the response distributions per question, per region in Fig. 2. The distributions were sufficiently different to warrant further study.

4.2 Accuracy Scores by Region

To assess the cultural alignment of our chosen LLMs, we calculate the "soft" and "hard" accuracies [1] for each persona by comparing LLM responses for each survey question q to those of the corresponding human survey participants. Specifically, for each persona p, we first identify all N survey participants p_i matching that persona. We obtain hard accuracy scores (see Eq. 1), H_r, by counting the

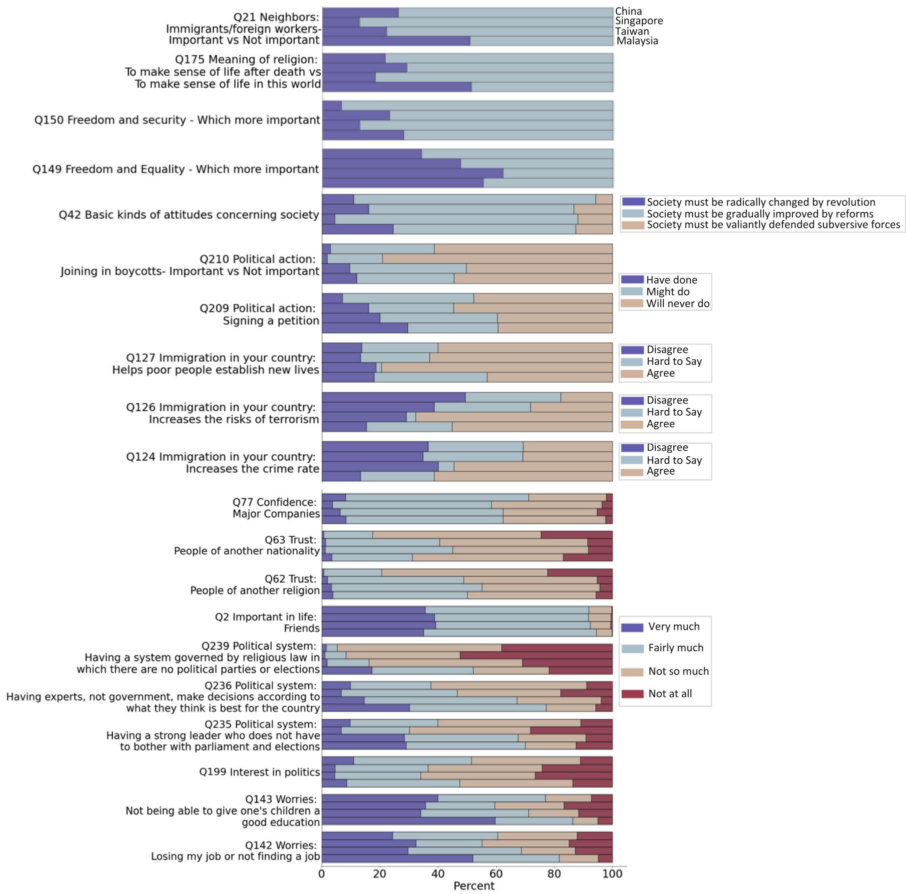


Fig. 2. Distribution of WVS responses for questions used in our study. There are four bars for each question, corresponding to responses from China, Singapore, Taiwan and Malaysia (see top right illustration). The questions toward the top have smaller Likert scales than toward the bottom. We found that the cultural response distributions varied enough to warrant our study.

number of times the LLM gave the same response $f(q,p)$ as the human participants $y_r(p)$, averaging over all N participants for a given country/region c.

$$H_c(q,p) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(f(q,p) = y_c(p_i)) \tag{1}$$

Soft accuracy, S_c, considers the ordinal nature of the Likert scales used in WVS (see Eq. 2). It gives higher scores when the LLM's answers are closer on the scale to human responses. Note that |q| represents the length of the Likert scale.

$$S_c(q,p) = \frac{1}{N} \sum_{i=i}^{N} \left(1 - \frac{|f(q,p) - y_c(p_i)|}{|q|-1}\right). \tag{2}$$

Table 3 shows accuracy scores aggregated by region. We see that full-precision Llama3 performs best for China, while Mixtral outperforms for Taiwan, Malaysia and Singapore. Qwen, a multilingual model largely trained on English and Chinese data, surprisingly performs significantly worse than other models across regions. This is improved with Qwen2. Combined with the improvement we see for Gemma2 over Gemma, this indicates that cultural alignment of LLMs may naturally improve with further model development. There are also observable differences in model performance across regions. All models are most culturally aligned with China, with varying performance for Taiwan, Malaysia and Singapore. This makes sense since China has more internet users and therefore produces more data for LLM training than the other regions.

Table 3. Accuracy scores for each LLM in predicting cultural values for four Chinese-speaking regions (presented as ($hard/soft$)). Models annotated with '*' are full-precision, the rest are quantized. Mixtral is more culturally aligned for each region except China, where Llama3 is superior. All models are most aligned with China.

LLM Model	China	Taiwan	Malaysia	Singapore
Phi3	.37/.62	.36/.63	.36/.64	.36/.63
Llama3	.40/.66	.40/.67	.39/.68	.39/.66
Llama3*	**0.54/0.76**	0.38/0.67	0.37/0.68	0.39/0.63
Mistral*	0.53/0.76	0.38/0.67	0.37/0.68	0.39/0.63
Mixtral	.50/.74	**.45/.73**	**.44/.72**	**.47/.70**
Gemma	.42/.69	.40/.69	.40/.69	.40/.68
Gemma2	.47/.74	.44/.74	.43/.73	.45/.71
Qwen	.28/.52	.28/.54	.29/.56	.27/.55
Qwen2	.42/.66	.39/.65	.38/.65	.40/.64
GPT-3.5 Turbo	.44/.71	.42/.71	.41/.70	.42/.69

4.3 Accuracy by Theme

In Fig. 3, we present radar charts illustrating the cultural alignment (using hard accuracy) of six LLMs by theme: Social Values, Social Capital, Political Culture, Political Interest, Religious Values, Security, and Migration [2]. Some models are more capable of capturing certain themes: Gemma/Gemma2 perform best in capturing Social Values for different regions, Llama3 and Gemma in capturing Political Culture, and Qwen in capturing Migration attitudes. While Mixtral does not achieve the same performance on these themes, it has overall higher

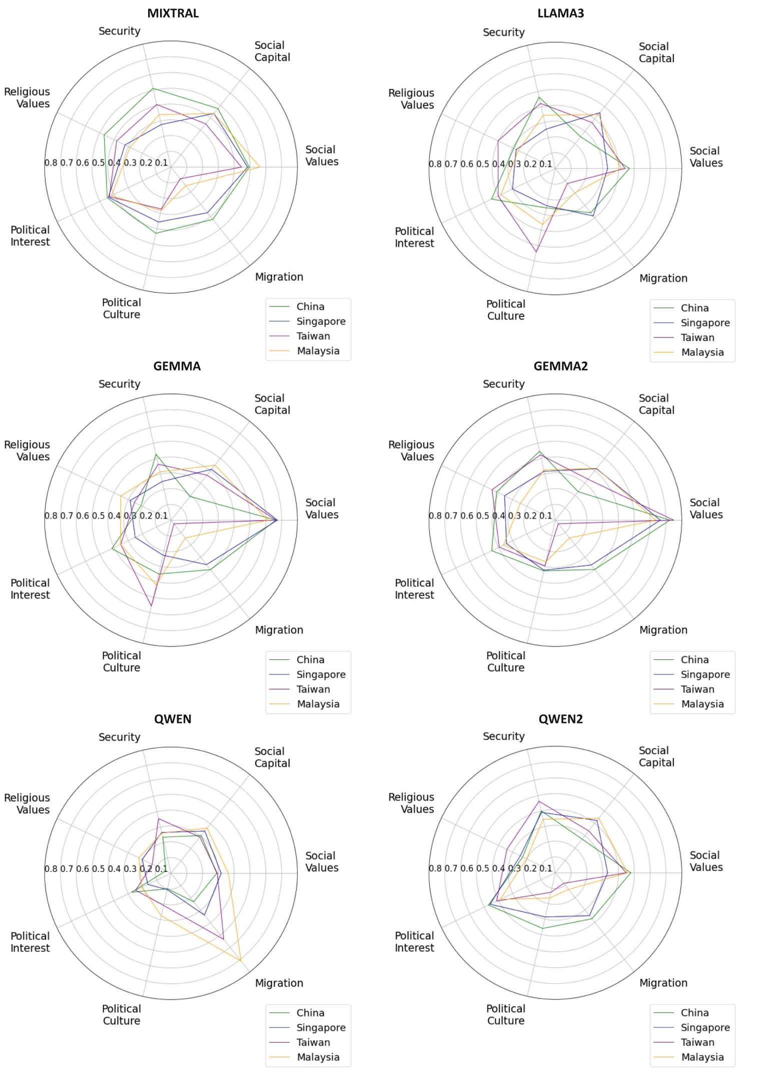


Fig. 3. We present region-level cultural alignment by question theme (Social Values, Social Capital, Political Culture, Political Interest, Religious Values, Security, and Migration) for six LLMs. While some models outperform on specific themes (e.g., Gemma/Gemma2 for Social Values), Mixtral's higher performance for 3 of 4 regions comes from its ability to better mimic responses across themes.

average performance across themes, explaining its higher accuracy for 3 of 4 regions. Similarly, we see that Qwen2 and Gemma2 outperform their predecessors due to mostly improved performance across themes.

4.4 Accuracy by Demographic (Persona Dimensions)

Table 4 illustrates each LLM's performance using hard accuracy across demographic variables, averaged across regions. Interestingly, the models perform slightly better for under-represented groups, a finding that does not echo the general consensus of past work that LLMs perform poorly for these groups. For five of eight models (Mixtral, Gemma, Gemma2, Qwen, GPT), accuracy is higher for lower education than higher education. Likewise, most of the models are more culturally aligned with the lower middle and working classes than the upper middle class. We hypothesize, but cannot yet confirm, that these results may be due to the data these models were trained on. We reserve this for future work.

Table 4. Accuracy of LLMs in predicting cultural values for different demographics, averaged over all regions. In contrast with findings in existing literature, LLMs tend to perform better for under-represented groups (lower social class and education levels).

	Phi3	Llama3	Mixtral	Gemma	Gemma2	Qwen	Qwen2	GPT-3.5 Turbo
Male	.34	.45	.46	**.49**	**.49**	.31	.42	.48
Female	.40	.41	.46	**.49**	48	.33	.44	.46
Lower Education	.36	.35	**.54**	.53	.49	.34	.38	.47
Higher Education	.39	.45	.44	**.48**	**.48**	.33	.45	.46
Upper Middle Class	.39	.42	.38	.38	.41	.34	.30	**.48**
Lower Middle Class	.38	.45	.47	**.49**	.48	.33	.46	.48
Working Class	.40	.37	.46	**.49**	.48	.33	.39	.43
Age 20-50	.40	.44	.43	**.47**	**.47**	.33	.45	.45
Age >50	.35	.37	**.54**	**.54**	.51	.33	.39	.49
Married	.37	.42	.46	**.49**	**.49**	.31	.42	.47
Single	.41	.42	.47	**.49**	.46	.36	.46	.45

4.5 Distributional Comparisons of Human and LLM Responses

As aforementioned, we prompted the LLMs five times to account for its stochastic nature, which can lead it to answer the same question in different ways. In Fig. 4, we compare the distribution of responses obtained from the Mixtral LLM across the five prompting sessions with the distribution of responses from human participants for China and Malaysia. For each question (left), we plot two bars: the top bar represents the percentage of human responses that fall into each possible response on the Likert scale, while the bottom bar represents the distribution of responses from Mixtral.

As seen in Fig. 4, Mixtral tends toward the "middle" when answering questions on a Likert scale of greater than two points. It generally avoids answers on the extremes of the scale, especially as compared to human responses. This is in line with past work that found that LLMs often hedge their answers when replying to cultural and moral questions, making it difficult to account for the wide variety of responses one might receive from a human population. This is especially true if the LLM has been trained to avoid making harmful responses, which we hypothesize is a contributing factor to "Hard to Say" (a non-answer) replies to questions about negative immigration effects (Q124, Q126). An exception to this hedging observation is the overrepresentation of LLM responses that friends are very important in life (Q2) across all regions, as compared to human responses. These findings have implications for the use of LLMs in social and psychological applications, and again point to the importance of probing LLMs prior to using them in sociocultural studies or having them engaging with individuals of varying backgrounds.

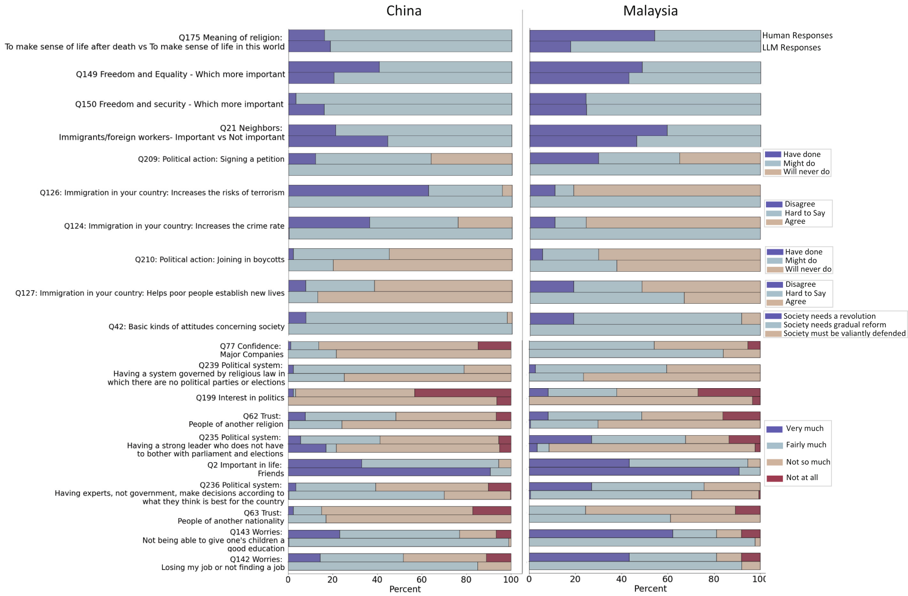


Fig. 4. Distribution of human and Mixtral responses to WVS questions for China and Malaysia. Questions are ordered such that those toward the top have smaller Likert scales than toward the bottom. There are two bars for each question, corresponding to response distributions for humans and Mixtral (see top right illustration).

4.6 Demographic Accuracy

Next we looked at the top k demographic groupings for which Llama 3 and Mistral can mimic survey responses for each country/region. To generate this result, we grouped the population of personas by $k = \{1, 2, 3, 4, 5\}$ attributes and assessed model accuracy for these groupings. For example, when $k = 1$ and the attribute was E(ducation), we partitioned the LLM personas and the survey population by education level and computed soft accuracy scores for these partitions. We then averaged these accuracy scores to get the result. We repeated this procedure for each subset of k attributes and present the most accurate groupings (Table 5). It is notable that in the Chinese-speaking regions, we observe that the LLMs generally become more accurate as the population is partitioned into finer sub-populations (as k increases). This is in contrast to the USA results, where the models become less accurate as partitions become finer.

Table 5. Most predictive demographic attributes. Each cell gives the attributes and soft accuracy score using full-precision LLMs. The attributes are labeled: education (E), country/region (C), marital status (M), gender (G), subregion (R), age (A), class (S).

Model	Region	Top-1	Top-2	Top-3	Top-4	Top-5
Llama3	USA	E (0.793)	E, C (0.793)	M, E, C (0.785)	G, M, E, C (0.755)	G, M, S, R, C (0.728)
Mistral	USA	E (0.790)	E, C (0.790)	M, E, C (0.783)	G, M, E, C (0.752)	G, M, S, R, C (0.731)
Llama3	China	E (0.820)	E, C (0.820)	M, E, C (0.810)	G, M, E, C (0.792)	A, G, E, R, C (0.780)
Mistral	China	R (0.813)	R, C (0.813)	G, R, C (0.803)	G, M, R, C (0.789)	A, G, E, R, C (0.772)
Llama3	Malaysia	R (0.684)	M, R (0.688)	M, R, C (0.688)	A, G, R, C (0.685)	A, G, E, R, C (0.679)
Mistral	Malaysia	R (0.680)	A, R (0.684)	A, R, C (0.684)	A, E, R, C (0.683)	A, G, E, R, C (0.681)
Llama3	Taiwan	R (0.701)	A, G (0.714)	A, G, S (0.727)	A, G, M, S (0.730)	A, G, M, S, C (0.730)
Mistral	Taiwan	R (0.701)	A, G (0.711)	A, G, S (0.716)	A, G, E, S (0.721)	A, G, E, S, C (0.722)
Llama3	Singapore	E (0.673)	E, C (0.790)	G, M, E (0.691)	G, M, E, C (0.691)	G, E, S, R, C (0.680)
Mistral	Singapore	E (0.685)	E, S (0.700)	E, S, C (0.700)	G, M, E, C (0.694)	G, E, S, R, C (0.682)

5 Conclusions and Future Work

In this work, we have shown that current LLMs do indeed differ in their ability to accurately generate text under the guise of different personas, demographics, and value systems. We found that there is noticeable variation between an LLM's performance across the different cultural aspects, and that no one LLM dominates the others across them all. Finally, when it comes to demographic attributes, we find that high-performing LLMs tend to do a better job assuming male personas than female personas, lower-educated and lower-middle to working class personas to higher-educated and upper middle class personas. It would be interesting to determine whether these discrepancies are artifacts of similar discrepancies in the data used to train these LLMs, and whether fine-tuning can help alleviate them. A limitation of the current study is its reliance on a single and arguably incomplete data source for deducing cultural attributes, the WVS.

An intriguing direction for future work is to expand on this and similar studies to include more cultures and communities and drawing from data sources beyond the WVS. The framework and tooling developed for such a wide-ranging analysis could be used to benchmark cultural alignment of current and future LLMs.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. AlKhamissi, B., ElNokrashy, M., AlKhamissi, M., Diab, M.: Investigating cultural alignment of large language models. arXiv preprint arXiv:2402.13231 (2024)
2. Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., et al.: World values survey: round seven–country-pooled datafile. JD Systems Institute, WVSA Secretariat, Madrid, Spain, Vienna, Austria (2020)
3. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D.: Can Large Language Models Transform Computational Social Science? Computational Linguistics. MIT Press, Cambridge (2020)
4. Kovač, G., Sawayama, M., Portelas, R., Colas, C., Dominey, P.F., Oudeyer, P.-Y.: Large language models as superpositions of cultural perspectives. arXiv preprint arXiv:2307.07870 (2023)
5. Buttrick, N.: Studying large language models as compression algorithms for human culture. Trends Cogn. Sci. (2024)
6. Arora, A., Kaffee, L.-A., Augenstein, I.: Probing pre-trained language models for cross-cultural differences in values. arXiv preprint arXiv:2203.13722 (2022)
7. Ramezani, A., Xu, Y.: Knowledge of cultural moral norms in large language models. arXiv preprint arXiv:2306.01857 (2023)
8. Naous, T., Ryan, M.J., Xu, W.: Having beer after prayer? Measuring cultural bias in large language models. arXiv (2023)
9. Leong, W.Q., Ngui, J.G., Susanto, Y., Rengarajan, H., Sarveswaran, K., Tjhi, W.C.: BHASA: a holistic southeast Asian linguistic and cultural evaluation suite for large language models. arXiv preprint arXiv:2309.06085 (2023)
10. Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D.: Out of one, many: using language models to simulate human samples. Polit. Anal. (2023)
11. Zhou, J., et al.: Rethinking machine ethics–can LLMs perform moral reasoning through the lens of moral theories? arXiv preprint arXiv:2308.15399 (2023)
12. Simmons, G.: Moral mimicry: large language models produce moral rationalizations tailored to political identity. arXiv preprint arXiv:2209.12106 (2022)

Author Index

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 257–258, 2024.
https://doi.org/10.1007/978-3-031-72241-7