

Lecture Notes in Electrical Engineering 964

Sarika Jain
Sven Groppe
Bharat K. Bhargava *Editors*

Semantic Intelligence

Select Proceedings of ISIC 2022

 Springer

Lecture Notes in Electrical Engineering

Volume 964

Series Editors

Leopoldo Angrisani, Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Naples, Italy
Marco Arteaga, Departament de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico
Bijaya Ketan Panigrahi, Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, Delhi, India
Samarjit Chakraborty, Fakultät für Elektrotechnik und Informationstechnik, TU München, Munich, Germany
Jiming Chen, Zhejiang University, Hangzhou, Zhejiang, China
Shanben Chen, Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
Tan Kay Chen, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore
Rüdiger Dillmann, Humanoids and Intelligent Systems Laboratory, Karlsruhe Institute for Technology, Karlsruhe, Germany
Haibin Duan, Beijing University of Aeronautics and Astronautics, Beijing, China
Gianluigi Ferrari, Università di Parma, Parma, Italy
Manuel Ferre, Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain
Sandra Hirche, Department of Electrical Engineering and Information Science, Technische Universität München, Munich, Germany
Faryar Jabbari, Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA
Limin Jia, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
Alaa Khamis, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt
Torsten Kroeger, Stanford University, Stanford, CA, USA
Yong Li, Hunan University, Changsha, Hunan, China
Qilian Liang, Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX, USA
Ferran Martín, Departament d'Enginyeria Electrònica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain
Tan Cher Ming, College of Engineering, Nanyang Technological University, Singapore, Singapore
Wolfgang Minker, Institute of Information Technology, University of Ulm, Ulm, Germany
Pradeep Misra, Department of Electrical Engineering, Wright State University, Dayton, OH, USA
Sebastian Möller, Quality and Usability Laboratory, TU Berlin, Berlin, Germany
Subhas Mukhopadhyay, School of Engineering and Advanced Technology, Massey University, Palmerston North, Manawatu-Wanganui, New Zealand
Cun-Zheng Ning, Electrical Engineering, Arizona State University, Tempe, AZ, USA
Toyooki Nishida, Graduate School of Informatics, Kyoto University, Kyoto, Japan
Luca Oneto, Department of Informatics, BioEngineering, Robotics and Systems Engineering, University of Genova, Genova, Genova, Italy
Federica Pascucci, Dipartimento di Ingegneria, Università degli Studi "Roma Tre", Rome, Italy
Yong Qin, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
Gan Woon Seng, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore
Joachim Speidel, Institute of Telecommunications, Universität Stuttgart, Stuttgart, Germany
Germano Veiga, Campus da FEUP, INESC Porto, Porto, Portugal
Haitao Wu, Academy of Opto-electronics, Chinese Academy of Sciences, Beijing, China
Walter Zamboni, DIEM—Università degli studi di Salerno, Fisciano, Salerno, Italy
Junjie James Zhang, Charlotte, NC, USA

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering—quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

Sarika Jain · Sven Groppe · Bharat K. Bhargava
Editors

Semantic Intelligence

Select Proceedings of ISIC 2022

 Springer

Editors

Sarika Jain
Department of Computer Applications
National Institute of Technology
Kurukshestra, India

Sven Groppe
Department of Computer Science
University of Lübeck
Lübeck, Germany

Bharat K. Bhargava
Department of Computer Sciences
Purdue University West Lafayette
West Lafayette, IN, USA

ISSN 1876-1100

ISSN 1876-1119 (electronic)

Lecture Notes in Electrical Engineering

ISBN 978-981-19-7125-9

ISBN 978-981-19-7126-6 (eBook)

<https://doi.org/10.1007/978-981-19-7126-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023, corrected publication 2023

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

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

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

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.

The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Organization

Advisory Committee



Rajkumar Buyya
CLOUDS Lab, School of Computing and Information Systems,
The University of Melbourne, Australia
rbuyya@unimelb.edu.au



Gurdeep Singh Hura
University of Maryland Eastern Shore, USA
gshura@umes.edu



Valentina Emilia Balas
Aurel Vlaicu University of Arad, Romania
balas@drbalas.ro

(continued)

(continued)



Bharat K. Bhargava
Purdue University, Indiana, United States
bbshail@purdue.edu

General Chairs



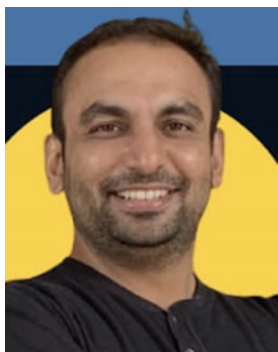
Sarika Jain
National Institute of Technology, Kurukshetra, Haryana, India
jasarika@nitkr.ac.in
<https://sites.google.com/view/nitkkrsarikajain>



Sven Groppe
University of Lübeck, Germany
groppe@ifis.uni-luebeck.de
<http://www.ifis.uni-luebeck.de/~groppe/>

Track PC Chairs

Valentina Janev
The Mihajlo Pupin Institute, Belgrade, Serbia
Valentina.Janev@instituteupin.com



Manas Gaur
Artificial Intelligence Institute, University of South Carolina,
Columbia, United States
MGAUR@email.sc.edu
<https://manasgaur.github.io/>



Asha Subramanian
Founder and CEO, Semantic Web India, Bengaluru, India
asha@semanticwebindia.com
<https://www.semanticwebindia.com/Aboutus.html>

Organizing Chairs

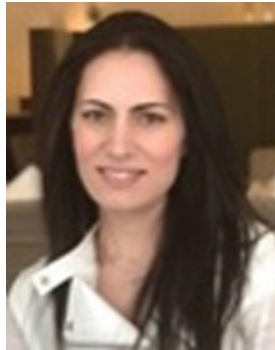


Atef Mohamed (Shalan)
Allen E. Paulson College of Engineering and Computing,
Georgia Southern University, United States
amohamed@georgiasouthern.edu



Hossain Shahriar
Associate Professor, BSIT/BASIT Coordinator, College of
Computing and Software Engineering; Director of Research
and Graduate Education, Institute for Cybersecurity Workforce
Development, Marietta, GA 30060, United States
hshahria@kennesaw.edu
<https://msit.kennesaw.edu/current-students/faculty.php>

Workshops and Special Sessions Chairs



Pelin Angin
Middle East Technical University, Ankara, Turkey
pangin@ceng.metu.edu.tr
<https://user.ceng.metu.edu.tr/~pangin/>

(continued)

(continued)



Prateek Agrawal
Associate Professor, Lovely Professional University, India
prateek061186@gmail.com
<http://www.itec.aau.at/~prateek/>

Publication Chairs



Jyotir Moy Chatterjee
Assistant Professor (IT) Lord Buddha Education Foundation
(Asia Pacific University of Technology and Innovation),
Kathmandu, Nepal
jyotirchatterjee@gmail.com
<https://sites.google.com/view/jyotirmoychatterjee>



Sachi Nandan Mohanty
Department of Computer science & Engineering, ICFAITech,
ICFAI Foundation for Higher Education, Hyderabad, India
sachinandan09@gmail.com
<http://drsachinandan.com/>

Publicity Chairs



Md Rafiqul Islam
University of Technology Sydney, Australia
MdRafiqul.Islam@uts.edu.au
<https://rafiqulislamcse24.wixsite.com/rafiqulcse>

Website Chairs



Mohamed Hamada
Senior Associate Professor The University of Aizu, Japan
Hamada@u-aizu.ac.jp
<http://www.u-aizu.ac.jp/~hamada/>



Ahmed A. Elngar
Assistant Professor, Faculty of Computers & Artificial
Intelligence, Beni-Suef University, Egypt
eIngar_7@yahoo.co.uk
<https://bsu-eg.academia.edu/AhmedElngar>

Technical Program Committee

Kumar Abhishek, NIT Patna
 Moussa Aboubakar, LIST, Communicating Systems Laboratory, France
 Prateek Agrawal, University of Klagenfurt, Austria
 Nesar Ahmad, AMU, Aligarh
 Tameem Ahmad, Aligarh Muslim University, Aligarh
 Syed Ahmed, Aligarh Muslim University
 Pelin Angin, Middle East Technical University, Turkey
 Valentina Emilia Balas, Aurel Vlaicu University of Arad, Romania
 Shajulin Benedict, IIIT Kottayam, India
 Bharat K. Bhargava, Purdue University, United States
 Amit Bhatia, Senior Principal Research Scientist, BAE Systems Inc, FAST Labs, NC, USA
 Suparna Biswas, Maulana Abul Kalam Azad University of Technology, West Bengal, India
 Rajkumar Buyya, The University of Melbourne, Australia
 Joel Luís Carbonera, UFRGS (Federal University of Rio Grande do Sul), Brazil
 Jyotir Moy Chatterjee, Lord Buddha Education Foundation, Kathmandu, Nepal
 Pethuru Raj, Reliance Jio Infocomm Ltd., Tamil Nadu, India
 Chandreyee Chowdhury, Jadavpur University, India
 Neama Abdulaziz Dahan, Sana'a University, Yemen
 Olawande Daramola, Cape Peninsula University of Technology, South Africa
 Gerard Deepak, Sr. Research Scholar, NIT Tiruchirappalli, India
 Ajantha Devi, AP3 Solutions, Chennai, India
 Abderrahim El Qadi, Mohammed V University in Rabat Morocco
 Moawia Elfaki Yahia Eldow, University of Khartoum, Sudan
 Ahmed Elngar, Beni-Suef University, Egypt
 Nafees Farooqui, Dehradun Institute of Technology, Uttarakhand, India
 Xiao-Zhi Gao, University of Eastern Finland, Finland
 Shankey Garg, National Institute of Technology Raipur, India
 Vinay Gautam, Chitkara University, India
 Sachin Sharma, Manav Rachna International Institute of Research and Studies, Faridabad, India
 Hiranmay Ghosh, Ex-Advisor, TCS Research
 Prakash Gopalakrishnan, Amrita Vishwa Vidyapeetham
 Vikas Goyal, Education Department, Haryana, India
 Sven Groppe, University of Lubeck, Germany
 Le Gruenwald, The University of Oklahoma, USA
 Charu Gupta, Indraprastha University, India
 Kapil Gupta, NIT Kurukshetra
 Mohamed Hamada, Aizu University, Japan
 Xavierlal J Mattam, Sacred Heart College, Kerala, India
 Priti Jagwani, Ram Lal Anand College
 Amit Jain, Sir Padampat Singhania University, Udaipur, India

Sarika Jain, National Institute of Technology, Kurukshetra
 Valentina Janev, The Mihajlo Pupin Institute, Belgrade, Serbia
 Yiming Ji, Georgia Southern University, USA
 Filbert H. Juwono, Lecturer, Department of. Electrical and Computer Engineering,
 Curtin University Malaysia
 Arpan Kar, Indian Institute of Technology Delhi, India
 Subodh Kesharwani, IGNOU, Delhi, India
 Laszlo T. Koczy, Budapest University of Technology and Economics, Hungary
 Petr Kremen, Czech Technical University in Prague, Czechia
 Naresh Kumar, IIT Roorkee, India
 Kamlesh Kumari, UIET, Kurukshetra University, India
 Naziha Laaz, Université Ibn Tofail, Kenitra, Morocco
 Shilpa S. Laddha, Govt. College of Engg., Aurangabad, Maharashtra, India
 Vishal Lama, Amdocs, Pune, India
 Aadil Ahmad Lawaye, BGSB University, India
 Dac-Nhuong Le, Haiphong University, Haiphong, Vietnam
 Ravi Lourdusamy, Sacred Heart College (Autonomous), Tirupattur, Vellore, Tamil
 Nadu, India
 Yang Lu, University of Kent, UK
 M. Niranjanamurthy, M. S. Ramaiah Institute of Technology, Bengaluru
 Aaisha Makkar, Thapar University, India
 Sonika Malik, MSIT, New Delhi, India
 Monika Mangla, CSED, LTCoE, Navi Mumbai
 Ganapathy Mani, Purdue University, USA
 Nikolaos Mavridis, United Arab Emirates University, United Arab Emirates
 A. Medina-Santiago, National Institute of Astrophysics, Optics and Electronics,
 Mexico
 Nandana Mihindukulasoor, IBM Research AI, India
 Sanjay Misra, Covenant University, Ota, Nigeria
 Ruchi Mittal, Netaji Subhas Institute of Technology, India
 SachiNandan Mohanty, ICFAI Foundation for Higher Education, Hyderabad, India
 Michael Mrissa, University of Primorska, Slovenia
 San Murugesan, BRITE Professional Services, Sydney, Australia
 Gagandeep Singh Narula, Guru Gobind Singh Indraprastha University, India
 Shahrul Azman Noah, Universiti Kebangsaan Malaysia
 Kingsley A. Ogudo, University of Johannesburg, South Africa
 Shanmugaraja P, Sonal College of Technology, Salem, Tamil Nadu, India
 Jyoti Pareek, Professor in Computer Science, Gujarat University, India
 Nenad Petrovic, University of Nis, Serbia
 Prajoy Podder, Bangladesh University of Engineering and Technology, Bangladesh
 Shivika Prasanna, University of Missouri—Columbia, USA
 Iurii Prokopchuk, National Academy of Sciences of Ukraine, Ukraine
 Dana Rad, Aurel Vlaicu University of Arad, Romania
 Suja Radha, VIT University
 Ripal D Ranpara, Atmiya University, India

Regina Reine, Curtin University, Malaysia
Aleksei Rozhnov, Institute of Control Sciences, Moscow, Russia
Shridevi. S, Vellore Institute of Technology, India
Ayodeji Salau, AfeBabalola University, Ado-Ekiti, Nigeria
Neetu Sardana, Jaypee Institute of Information Technology, Noida, India
Salma Sassi, University of Jendouba, Tunisia
Suneeta Satpathy, BPUT, India
Sharad Saxena, Thapar University, India
K Kalaiselvi, Vels Institute of Science, Technology & Advanced Studies, Chennai, India
Fatmana Senturk, Pamukkale University, Turkey
Hossain Shahriar, Kennesaw State University, Georgia, USA
Atef Shalan, Georgia Southern University, Georgia, USA
Sachin Gengaje, WIT, Solapur
Sudhir Kumar Sharma, IITM Janakpuri, GGSIPU Delhi, India
Cogan Shimizu, DAGSI Fellow and Instructor, Data Semantics Lab, Wright State University, USA
Zee Ang Sim, Curtin University, Malaysia
Pranav K. Singh, Department of CSE, CIT Kokrajhar, India
Sushil Kumar Singh, Seoul National University of Science and Technology, South Korea
Vikram Singh, NIT Kurukshetra, India
Karen Smiley, Senior Technology Development Manager, BAE Systems Inc, FAST Labs, USA
Konstantinos Sofianos, Ionian University, Corfu, Greece
Rituraj Soni, Engineering College Bikaner, India
Sandeep Sood, Central University of Himachal Pradesh, Shahpur, Himachal Pradesh, India
Srinath Srinivasa, Web Science Lab, IIIT-Bangalore, Bengaluru, India
Sweta Srivastava, Amity University-Noida, India
Jon Stammers, University of Sheffield, England, United Kingdom
Asha Subramanian, Semantic Web India Private Limited, Bengaluru, India
Muhammad Imran Tariq, Superior University, Lahore, Pakistan
Mohseena Thaseen, NES Science College, Maharashtra, India
Sree Ganesh Thottempudi, University of Heidelberg and BBAW, Germany
Sanju Tiwari, Universidad Autonoma de Tamaulipas, Mexico
Ted Tschang, Singapore Management University, Singapore
Olegs Verhodubs, Training Center SIA DRMC, Latvia
Pawan Kumar Verma, GLA University, India
Benjamin Warnke, University of Lubeck, Germany
Wong Wei Kitt, Curtin University, Malaysia
Kai Wussow, SAP SE, Germany
Asmita Yadav, Jaypee Institute of Information Technology, Noida, India
Chandra Shekhar Yadav, STQC, MeitY, India
Anatolij Zabrovski, University of Klagenfurt, Austria

Acknowledgments

No one can whistle a symphony. It takes a whole orchestra to play it.

—H. E. Luccock

ISIC 2022 is the result of a larger group of people working together. The entire organizing committee had been extremely helpful in ensuring the success of ISIC 2022. The editors express their gratitude to the organizing team, which includes the track chairs, session chairs, technical program committee members, external reviewers, and also the authors. We are grateful to the many volunteers who worked tirelessly to ensure the event's success.

About This Book

Many ideas grow better when transplanted into another mind than the one where they sprang up.

—*Oliver Wendell Holmes*

Considering this quote as the sole motto of the International Semantic Intelligence Conference, this book constitutes the proceedings of the 2nd International Semantic Intelligence Conference (ISIC 2022), held at Georgia Southern University (Armstrong Campus), Savannah, United States from May 17 to 19, 2022. The ISIC is an international symposium that convenes to publish cutting-edge research results in intelligent applications for the Artificial Intelligence, Machine Learning, and Semantic Web communities.

Semantic Intelligence is rising as an important suite of technologies as a way forward with Artificial General Intelligence. Although the first version of ISIC 2021 started in the general domain of Artificial Intelligence and Machine Learning, we hope and envision that with years ISIC will succeed as a more focused Semantic Intelligence Conference.

ISIC 2022 has four top-shot invited researchers as advisory. The conference committee has diligently finalized the five keynote speakers. They are multi-diversified in nature across the whole world and are esteemed experts in their field. The main conference organization has 18 chair members whereas there are approximately 121 technical program committee members from various countries globally. ISIC 2022 also exhibits four pre-conference tutorials this year. The second edition of the conference also depicts a high geographical diversity from around 40 different countries as the members and more than 35% women as high gender diversity within its organization, similar to the last edition.

ISIC 2022 has been conducted in Online mode. Only high-quality manuscripts in the area of the conference are accepted for final publication by virtue of review and selection process. Every manuscript was reviewed by three to four reviewers and just like last year, this year also the acceptance rate was 50%. The volume comprises 22 manuscripts from 74 authors coming from 34 different universities/institutions over

10 different countries, namely, the United States, India, Germany, Serbia, Turkey, Nigeria, Italy, South Korea, Canada, and Morocco.

We invite proposals from universities and institutes to host the next edition of the International Semantic Intelligence Conference ISIC 2023.

Keynote Talks

ISIC 2022 witnessed five keynote talks in five different topics under the umbrella of Semantic Intelligence.

a. **Title:** Applying Knowledge Graphs for Data Analytics and Machine Learning

Speaker: Dr. Ernesto Jiménez-Ruiz (Lecturer, City, University of London)

Abstract: The application of knowledge graphs (KG) is going beyond the original vision of the Semantic Web and KGs are starting to play a key role to organize the enterprise, GLAM, and governmental data, and they are already the backbone in several bio-medical applications. Enterprises are also leveraging knowledge graphs to drive their products and make them more “intelligent.” The next steps in AI involve the creation of richer and smarter AI systems in regards to semantically sound, explainable, and reliable. Hybrid learning and reasoning systems combining subsymbolic and symbolic representations are gaining renewed attention, within both the Machine Learning (ML) and Knowledge Representation communities, to lead to the design and creation of such richer AI systems.

Video Link: <https://youtu.be/KSAKSOvCMHs>

Short Biography: Ernesto Jimenez Ruiz is a Lecturer in Artificial Intelligence at City, University of London affiliated to the Research Centers for Machine Learning and Artificial Intelligence. He is also a researcher in the Centre for Scalable Data Access (SIRIUS) at the University of Oslo, Norway. He previously held a Senior Research Associate position at The Alan Turing Institute in London (UK) and a Research Assistant position at the University of Oxford. His home university (Universitat Jaume I, Castellon, Spain) awarded a “Premio extraordinario de doctorado” (roughly translated as a Extraordinary Doctoral Award) to his doctoral thesis (Engineering category 2010–2011). His research has covered several areas, including bio-medical information processing and integration, ontology reuse, ontology versioning and evolution, ontology alignment. His current research interests focus on the application of Semantic Technology to Data Science workflows and the combination of Knowledge

Representation and Machine Learning techniques. My complete list of publications can be found here. The PDF of most of the articles are available online.

b. **Title:** Detect, Characterize, and Accommodate Novelities in AI systems.

Speaker: Bharat K. Bhargava (Purdue University, Indiana, USA)

Abstract: Novelities are surprises that an AI system encounters. It is easier for a human to detect novelties and adjust. An automated and autonomous system must learn about the characteristics and detect, understand, and adapt to novelty in not only the environment but in agents that interact with it. For example, in a game such as monopoly or chess, the rules of the game can be suddenly changed or players may change their objectives. Players may also collude with other players to achieve an outcome such as draw or extend the game beyond time limit, or make one of the players lose or win. Even though the military is trained to deal with different environments before deployments, it can encounter novelties that it must deal with. Example could be an enemy using a motorcycle on a narrow path in high mountains where jeep or trucks cannot follow.

Systems or agents do not need to react to every novelty. Some of them are just nuisance and do not affect the operations. Some novelties are transient and disappear and do not reappear. Some novelties are easy to detect and react to. Some novelties overlap with past novelties and system can easily adapt.

The context, timing, duration, extent, and duration of novelty must be considered in agent's adaption and accommodation. How to build AI/ML system that can adapt to fluid novelties in open world will be presented. We present scientific principles to quantify and characterize novelty in open-world domains. We identify measures and evaluation criterion for behavior of AI system when encountering novelties.

Novelties are found in many environments and agents must learn about them and accommodate them.

Video Link: <https://youtu.be/XWq8I-rv94U>

Short Biography: Bharat K. Bhargava is a Professor of the Department of Computer Science with a courtesy appointment in the School of Electrical & Computer Engineering at Purdue University. His recent research is on Intelligent Autonomous Systems and data analytics and machine learning. It includes cognitive autonomy, reflexivity, deep learning and knowledge discovery. His earlier work on Waxed Prune with MIT and NGC built a prototype for privacy preserving data dissemination in cross-domains. Currently he is leading the NGC REALM consortium. He has graduated the largest number of Ph.D. students in CS department at Purdue and is active in supporting/mentoring minority students. In 2003, he was inducted in the Purdue's Book of Great Teachers. In 2017, he received the Helen Schleman Gold Medallion Award for supporting women at Purdue and Focus award for advancing technology for differently abled students.

c. **Title:** Leveraging Artificial Intelligence and Machine Learning in Pandemics using COVID-19 as a Case Study

Speaker: Sven Groppe (University of Lübeck, Germany)

Abstract: The COVID-19 pandemic slows down or even often stopped the world's activities in economy, education, society, and other areas of our daily life, but was a huge trigger for research. Smart and hardworking scientists all over the world are still extending the knowledge about the COVID-19 virus and are contributing to various technologies to fight against the COVID-19 pandemic. Continuously newly occurring mutations of the original virus demand for still working on and improving the developed technologies against the pandemic.

This talk covers a short introduction into the effects of the COVID-19 pandemic by naming its winners and losers. Losers of the COVID-19 pandemic include infected humans (suffering more than necessary from overburdened health systems), economy (caused by lockdowns), students (having to catch up with missed topics due to closed schools), and society (suffering from canceled events). There are also some winners of the COVID-19 pandemic like vaccine developers (with increasing stock price performance), sellers of medical products (increasing their sales), and technologies used to overcome pandemics (the development of which is enormously triggered by funded research).

This talk tries to provide an overview to answer where computers can help in our fight against the pandemic. Many areas and technologies have been identified for this purpose. According to my opinion, the most important technology for a short-time reaction to the COVID-19 virus in medical research is sequencing a genome and analyzing it via supercomputers. One of the most prominent examples for other developed approaches are the predictions of incidence rates and other COVID-19 data (like hospitalization rates) considering COVID-19 confinements and other contexts by computer simulations and machine learning approaches. There is also a need for the management of physical contacts, e.g., at events and restaurants, and apps for personal contact tracking to warn a group of or single persons in the case they have been in contact with an infected person. In order to overcome security risks different approaches for contact tracking have been discussed and developed like mobile operator, location-based, and proximity-based contact tracing. Software within health systems has been improved or introduced, e.g., patient registration and status in hospitals, automatically recognizing COVID-19 patients from their computer tomography scans, and publicly available databases of confirmed COVID-19 cases and other COVID-19-related data, which are the basis for deeper analysis of the COVID-19 pandemic. On the basis of the achieved knowledge about the COVID-19 virus and the effects of the COVID-19 pandemic, a set of COVID-19 knowledge graphs have been released, which provide automatic means for answering related questions and help to structure the information flood of COVID-19-related data. Because of the enormous list of developed and used technologies related to COVID-19, this talk cannot dive into all of them, but will tackle the most important ones to learn for future pandemics.

Video Link: <https://youtu.be/GAnC0ktlJFU>

Short Biography: Sven Groppe is Professor at the University of Lübeck, Germany. He was a member of the DAWG W3C Working Group, which developed SPARQL. He was the project leader of the DFG project LUPOSDATE and two research projects on FPGA acceleration of relational and Semantic Web databases, and is a member of the Hardware Accelerator Research Program by Intel. He is currently the project leader of German Research Foundation projects on GPU accelerated database indices and on Semantic Internet of Things. Furthermore, he is leading a project about quantum computer accelerated database optimizations and he is project partner in a project about COVID-19 high-quality knowledge graphs, visualizations and analysis of the pandemic with 2 French partners. His research interests include Internet of Things, Semantic Web, query and rule processing and optimization, Big Data, Cloud Computing, peer-to-peer (P2P) networks, data visualization and analysis, and visual query languages.

He is the workshop organizer and chair of the Semantic Big Data workshop series (2016 to 2020) in conjunction with ACM SIGMOD. In 2021 and 2022 he organized the International Workshop on Big Data in Emergent Distributed Environments (BiDEDE) @ SIGMOD and the International Workshop of Internet-of-Things (VLIoT) in conjunction with VLDB since 2017. He is the general chair of the International Semantic Intelligence Conferences in 2021 and 2022.

d. **Title:** Responsible AI for National Security

Speaker: Amanda Muller (Artificial Intelligence Systems Engineer and Technical Fellow Northrop Grumman Mission Systems)

Abstract: Human-machine teaming is a critical consideration for ensuring the successful implementation of semantic technologies. Without consideration for the human element of an Artificial Intelligence or Machine-Learning-enabled system, performance will suffer, or worse—the system simply will not be used. AI ethical frameworks can be leveraged as an enabler of human-machine teaming by certifying that systems are developed in line with human values. Ethical frameworks such as the U.S. Department of Defense’s Five Ethical Principles of AI contain the necessary guidelines to ensure that AI systems are interpretable, governable, and usable by humans. However, there is no one-size-fits-all ethical framework—the right framework must be carefully selected based on the use case in question, the risk profile, and applicable laws and regulations. This presentation will examine the use of ethical frameworks as an enabler of human-machine teaming in AI, and the factors to consider when choosing the right one for a particular use case.

Video Link: <https://youtu.be/PKPhCtG3EDo>

Short Biography: Dr. Amanda Muller is a Consulting Artificial Intelligence (AI) Systems Engineer and Technical Fellow Emeritus based in Northern Virginia. Dr. Muller currently serves as the Responsible AI Lead for Northrop Grumman. In this role, she is responsible for coordinating the strategy, policy, and governance

efforts related to Artificial Intelligence across the Northrop Grumman enterprise. As a Mission Systems Technical Fellow Emeritus specializing in User Experience and Human-Systems Integration, she also serves as a subject matter expert on proposals, program reviews, and research efforts. Prior to her current role, Dr. Muller worked for Northrop Grumman Space Systems in Redondo Beach, California, as a Systems Engineer. She led the User Experience teams for several restricted space programs, conducting user research in operational environments around the world. Previously, Dr. Muller served as a Systems Engineer on State Health and Human Services programs, as a Human Factors Engineer in Aurora, Colorado, and as the Human-Systems Integration lead for airborne platforms in Melbourne, Florida. In addition to her program roles, Dr. Muller has been a mentor in the Mentoring the Technical Professional program for over seven years.

Dr. Muller's publications include a book chapter in *Emerging Trends in Systems Engineering Leadership: Practical Research from Women Leaders* (in press), and peer-reviewed articles in *Information Fusion*, *Journal of Defense Modeling and Simulation*, *WSEAS Transactions on Advances in Engineering Education*, and the *Annals of Biomedical Engineering*.

Dr. Muller holds a Ph.D. in Engineering from Wright State University in Dayton, Ohio, and B.S. and M.S. degrees in Biomedical Engineering from Worcester Polytechnic Institute in Worcester, Massachusetts. She also holds a graduate certificate in Design Thinking for Strategic Innovation from Stanford University. Dr. Muller is a Certified Systems Engineering Professional (INCOSE), Professional Scrum Master (Scrum.org), and is certified in Professional Scrum with User Experience (Scrum.org).

e. **Title:** Semantic Intelligence: The Next Step in AI

Speaker: Sarika Jain (National Institute of Technology Kurukshetra, India)

Abstract: Intelligent agents work autonomously by seeking necessary information, coordinating with each other, and taking necessary actions to make life simple for human beings. There are three information aspects for an intelligent agent: syntax (sentence construction, grammatical correctness), semantics (human-level interaction), and pragmatics (intention behind the communication). An intelligent agent is required to fuse heterogeneous sources of information together for which it should be equipped with both the data-driven (statistical) and knowledge-driven (symbolic) AI disciplines. We need a representation of our data that not only includes the data itself but where the interactions in it is a first-class citizen.

We have seen in the past decade that statistical models have revolutionized the world. Though the Statistical models have already proved themselves, they are not a Universal Solvent but only a tool as others. Deep learning is very good at learning in a static world and executing low-level patterns, provided it is fed with a lot of data. More deep, more intelligent, and, of course, more black. The question is "Is the AI of today Artificial Super Intelligence (ASI)/Artificial General Intelligence (AGI)/Artificial Narrow Intelligence (ANI)? Is the AI of today the AI that we are craving for?" In fact, today's artificial intelligence is weak AI. There are a number of

instances where DL has produced delusional and unrealistic results. Accuracy alone is not sufficient. We require exploring ways of opening the black box of statistical models. When DL researchers are asked to open the black box, this today implies less intelligent models to them (limited capability). In addition to increased performance, AGI aims to build trust.

Symbolic AI and statistical AI have to go together to achieve contextual computing. The symbolist approach is nowadays manifested as a knowledge graph that advanced statistics and machine learning can run on top of. The Hybrid Model combines machine intelligence with human intelligence to reach conclusions faster than possible by humans alone along with the explanations needed for trust in the decisions and results, while requiring far fewer data samples for training and conversing in natural language. The Hybrid Model is able to generalize and is excellent at perceiving, learning, and reasoning with minimal supervision. In addition, semantics have come a long way in enhancing explainability in AI systems.

Video Link: <https://youtu.be/r18vXwkt57Y>

Short Biography: Sarika Jain graduated from Jawaharlal Nehru University (India) in 2001. Her doctorate, awarded in 2011, is in the field of knowledge representation in Artificial Intelligence. She has served in the field of education for over 19 years and is currently in service at the National Institute of Technology Kurukshetra (Institute of National Importance), India.

Dr. Jain's major research interests include Artificial Intelligence, the Semantic Web, Ontological Engineering, and Knowledge Graphs. She has received grants from Defense Research and Development Organization, Department of Science and Technology, Council of Scientific and Industrial Research for research, and National Institute of Technology Kurukshetra for research projects; from All India Council for Technical Education (thrice) for FDPs; from DAAD RISE worldwide (thrice) for hosting research interns from Germany; and Ministry of Human Resource and Development (twice) for hosting a reputed international faculty and for FDP. She has published over one hundred peer-reviewed technical papers in books, journals, and conference proceedings. She has served as a General Chair, Workshop Chair, Program Committee Chair at many international conferences and workshops; and Reviewer for journals published by IEEE, Elsevier, and Springer.

She has held various administrative positions at the department and institute levels in her career. Among the awards and honors, she has received are the Best Paper Awards, Feb 2021 (two), Aug 2020, Aug 2017; and the Best Faculty Award, Sep 2019. Dr. Jain works in collaboration with various researchers across the globe, including in Germany, Australia, Malaysia, the United States, and Romania. She is a senior member of the IEEE, a member of ACM, and a Life Member of CSI.

Pre-conference Tutorials

ISIC 2022 witnessed three pre-conference tutorials under the umbrella of Semantic Intelligence. This series of tutorials was the fifth workshop on Semantic Intelligence in its series with the first four already held on different occasions. This workshop aims to establish the importance of using semantic data models by integrating the full potential of existing approaches, tools, techniques, methodology to provide situation awareness, and advisory support in a seamless manner among the participants.

Key features: The uniqueness of this workshop lies in several respects.

- This is the fifth workshop in its series and being well organized it discusses current research in knowledge-based systems and presents it in a way that non-experts in computer science may grasp.
- It provides comprehensive hands-on pedagogy in the latest approaches for developing and publishing Linked Data Applications on the web.
- This workshop will assist graduate and undergraduate students taking courses in Artificial Intelligence, Semantic Web, Knowledge Engineering, and Decision Support Systems.
- A novice in the field of computer science can learn how to use semantic web technologies for real-world challenges after taking this program.
- This workshop will be beneficial to a variety of users:
 - senior undergraduate and graduate students,
 - academicians and researchers, and
 - practitioners in all application domains.

1. Tutorial 1: Building Domain-Specific Linked Data Applications

Presenters: Sarika Jain, National Institute of Technology Kurukshetra, India; Pooja Harde, National Institute of Technology Kurukshetra, India; Ankush Bhist, University of Delhi, India; Nandana Mihindukulasooriya, MIT-IBM Watson AI Lab, Cambridge, USA

Duration: 3 Hours

Video Link:

Module 1—Semantic Web Vision as Motivation for Linked Data: <https://youtu.be/X9Y6snFFdUs>

Module 2—Hands-On Session for Knowledge Modelling: <https://youtu.be/vpOADh4qh5E>

Modules 3&4—Hand-On Sessions for RDF Creation and SPARQL Query: <https://youtu.be/8M5Xd9haF8k>

Description of the Tutorial: Traditional data techniques and platforms do not prove to be efficient because of issues concerning responsiveness, flexibility, performance, uncertainty, heterogeneity, scalability, accuracy, and more. This data is understandable by humans and is really not amenable for machine processing. Semantic Web and Linked Data technologies have been found as the most important ingredient in building artificially intelligent knowledge-based systems.

2. Tutorial 2: Knowledge Infused Reinforcement Learning for Social Good Applications

Presenters: Manas Gaur, Research Scientist, Artificial Intelligence Institute, University of South Carolina; Kaushik Roy, Ph.D. Student at Artificial Intelligence Institute, University of South Carolina

Duration: 2 Hours

Video Link: <https://youtu.be/A5SzrWBDxCY>

Description of the Tutorial: Reinforcement Learning (RL) is a popular framework to control a sequential decision-making process using rewards or reinforcement. Though optimizing a goal-directed reward is suitable for many real-world applications, the emergence of big data has led to highly data-driven and black-box algorithms. However, in social good applications, well-defined domain knowledge and procedural information are critical to human decision-making that should be incorporated in the reward. Furthermore, the significant discrepancy between black-box decision-making and human-like decision-making limits effective communication to facilitate the seamless incorporation of data-driven and human-provided rewards. In this study, we develop extit {Knowledge Infused Reinforcement Learning} (KiRL) that addresses the above challenges. We test our approach on benchmark datasets and real-world applications—specifically for Contagion Control and Mental Health Triaging. We illustrate the qualitative and quantitative efficiency of transparent, explainable methods that provide knowledge-guided, safe, and transparent mechanisms for effective interaction between human domain experts, users, and RL algorithms. Thus, this tutorial will establish the usefulness of KiRL as a much-needed technological assistance tool for real-world social good applications.

3. Tutorial 3: Knowledge Base Question Answering

Presenters: Nandana Mihindukulasooriya, MIT-IBM Watson AI Lab, Cambridge, USA

Duration: 1 Hour

Video Link: <https://youtu.be/2BknPDxaGUE>

Description of the Tutorial: Knowledge Base Question Answering (KBQA) is the task of providing precise answers to natural language questions using facts in a knowledge base and it has been an important research topic since the early days of AI. KBQA systems have been using many different approaches from rule-based expert systems to end-to-end deep-learning-based systems and more recently Neuro-Symbolic approaches. KBQA also has several closely related subtasks such as entity linking, relation linking, and answer-type prediction. In this talk, we will discuss the KBQA task, its challenges, different reasoning types needed for question answering, and different subtasks involved in KBQA. We will also look at different benchmarks for evaluating the KBQA task.

The Best Paper Awards

Like last year, this year also the International Semantic Intelligence Conference witnessed high-quality submissions. The conference committee selected three best paper awards this year after following a rigorous criteria. The selection was done at three levels with filtering being done at every level. The Track PC Chairs looked into the grading and the comments provided by the session chairs for every paper. The review comments received during the review process were considered for the second level of filtration. The Track PC Chairs themselves did the third level of filtration. The final major criteria set was the works pertaining to the theme of the conference, i.e., SEMANTIC INTELLIGENCE.

1. **Session Chairs (Presentations made):**

- Valentina Janev, Institut Mihajlo Pupin, Serbia;
- Archana Patel, Eastern International University, Vietnam;
- Fatmana Şentürk, Pamukkale University, Turkey;
- Sachi Nandan Mohanty, ICFAI Foundation for Higher Education, Hyderabad, India;
- Asha Subramanian, Semantic Web India, Bengaluru;
- Prakash Gopalakrishnan, Amrita Vishwavidyapeetham, Bengaluru;
- Filbert H. Juwono, Curtin University, Malaysia; and
- Kapil Gupta, National Institute of Technology Kurukshetra, India.

2. **Reviewers Comments.** The Technical Program Committee (TPC) is already announced on the ISIC 2022 website. The Program Committee is very diversified across the globe and the maximum percentage is the Artificial Intelligence researchers.

3. **Track PC Chairs (Overall):** The Track PC Chairs are already announced on the ISIC 2022 website. All the three Track PC Chairs are active researchers in Semantic Intelligence.

Professor Bharat K. Bhargava, one of the advisory committee members, announced the Best Paper Awards during the Valedictory session. Here is the list of papers that received the Best Paper Award during ISIC 2022:

Sr. no.	Paper title	Authors
1	Knowledge-based Extraction of Cause-Effect Relations from Biomedical Text	Sachin Pawar, Ravina More, Girish Palshikar, Pushpak Bhattacharyya, and Vasudeva Varma
2	Towards a Solution for an Energy Knowledge Graph	Dušan Popadić, Enrique Iglesias, Ahmad Sakor, Valentina Janev, and Maria-Esther Vidal
3	NyOn: A Multilingual Modular Legal Ontology for Representing Court Judgments	Sarika Jain, Pooja Harde, and Nandana Mihindukulasooriya

Contents

The Research Track

Toward a Solution for an Energy Knowledge Graph	3
Dušan Popadić, Enrique Iglesias, Ahmad Sakor, Valentina Janev, and Maria-Esther Vidal	
Web-Based Visualization and Analysis Framework for Graph Data	13
Fatmana Şentürk, Mehmet Ali Bilici, Sezercan Tanışman, and Vecdi Aytaç	
Web Service Credibility Evaluation Methods in Different Application Domains	29
Atef Shalan, Jaciel E. Reyes, Hayden Wimmer, Sarika Jain, and Mohamed Hefny	
Semantic Web Ontology for Botnet Classification	43
Omotola Adekanmbi, Hayden Wimmer, and Atef Shalan	
Design and Performance Evaluation of a Multi-patient Health Monitoring System	55
Samson Olasunkanmi Adigun, Ayodeji Olalekan Salau, and Fatima Chiamaka Ujunwa	
Discovering Novelty via Transfer Learning	67
Shafkat Islam and Bharat K. Bhargava	
An Ontology for Social Media Data Analysis	77
Sarika Jain, Sumit Dalal, and Mayank Dave	
The Coronavirus Disease Ontology (CovidO)	89
Sumit Sharma and Sarika Jain	
An Ethnolinguistic Research Agenda for Intelligent Autonomous Systems	105
Bharat K. Bhargava, Sarika Jain, and Abhisek Sharma	

QuantumRNG, A Random Number Generator Using One Qubit 119
Dara Ekanth, Bheemanathy Saketh Chandra, and Meena Belwal

Sign Language Detection Using Machine Learning 135
P. Ilanchezhian, I. Amit Kumar Singh, M. Balaji, A. Manoj Kumar,
and S. Muhamad Yaseen

**Scrutinize and Discover of Image of Freshwater Taken by Faraway
Realizing Using FFNN and ConvNet Mechanisms** 145
D. Komalavalli, P. Ilanchezhian, A. Diwakar, K. Gayathri,
T. S. Indhuja, and R. V. Devadharshini

**Knowledge-based Extraction of Cause–Effect Relations
from Biomedical Text** 157
Sachin Pawar, Ravina More, Girish K. Palshikar,
Pushpak Bhattacharyya, and Vasudeva Varma

**NyOn: A Multilingual Modular Legal Ontology for Representing
Court Judgements** 175
Sarika Jain, Pooja Harde, and Nandana Mihindukulasooriya

The Applications and Deployment Track

**Technologies for AI-Driven Fashion Social Networking Service
with E-Commerce** 187
Jinseok Seol, Seongjae Kim, Sungchan Park, Holim Lim,
Hyunsoo Na, Eunyoung Park, Dohee Jung, Soyoun Park,
Kangwoo Lee, and Sang-goo Lee

**Deep Learning-Based Classification of Customer Communications
of a German Utility Company** 205
Jinghua Groppe, René Schlichting, Sven Groppe, and Ralf Möller

The Trends and Perspectives Track

Short Analysis of the Impact of COVID-19 Ontologies 225
Sven Groppe, Sanju Tiwari, Hanieh Khorashadizadeh,
Jinghua Groppe, Tobias Groth, Farah Benamara, and Soror Sahri

**Demystifying Semantic Intelligence for Enabling Intelligent
Applications** 241
Sarika Jain

**Sentiment Analysis of Public Health Concerns of Tokyo 2020
Olympics Using LSTM** 255
Ayodeji Olalekan Salau, Temiloluwa Oluwatomisin Omojola,
and Wasii Adeyemi Oke

Internet of Nano and Bio-Nano Things: A Review 265
Şeyda Şentürk, İbrahim Kök, and Fatmana Şentürk

**iTelos—Case Studies in Building Domain-Specific Knowledge
Graphs** 277
Simone Bocca, Mauro Dragoni, and Fausto Giunchiglia

**Comparative Study of Image Encryption and Image Steganography
Using Cryptographic Algorithms and Image Evaluation Metrics** 297
Surya Teja Chavali, Charan Tej Kandavalli, T. M. Sugash,
and G. Prakash

**Correction to: QuantumRNG, A Random Number Generator
Using One Qubit** C1
Dara Ekanth, Bheemanathy Saketh Chandra, and Meena Belwal

About the Editors

Sarika Jain has served in the academic field for over 19 years and is currently associated with the National Institute of Technology Kurukshetra, India. Dr. Jain has authored over 150 publications to her credit. Her research interests include knowledge management and analytics, ontological engineering, knowledge graphs, and intelligent systems. Dr. Jain has been a principal investigator of sponsored research projects and has worked collaborating with various researchers across the globe. She serves as a reviewer for journals published by reputed publishers. She is a senior member of the IEEE, a member of ACM, and a life member of the CSI.

Sven Groppe is a professor at the University of Lübeck, Germany. His publication record contains over 100 publications, including the book *Data Management and Query Processing in Semantic Web Databases*. He was a member of the DAWG W3C Working Group about SPARQL. He was the project leader of the DFG project LUPOSDATE, an open-source semantic web database, and of two research projects in the area of FPGA acceleration of relational and Semantic web databases. He is also leading a DFG project on GPU and APU acceleration of main-memory database indexes and a DFG project about semantic Internet of things. His research interests include artificial intelligence, databases, semantic web, and (post) cloud computing.

Bharat K. Bhargava is a professor at Purdue University, USA and is researching security and privacy issues in distributed systems. His recent work is on intelligent autonomous systems, data analytics, and machine learning. He is the recipient of seven best paper awards at various international computer science conferences. He is a fellow of the IEEE and the IETE. He has received various awards including the IEEE Technical Achievement Award, the charter Gold Core Member distinction, and the Outstanding Instructor Award. He has graduated with the largest number of Ph.D. students in the CS department and is active in supporting/mentoring minority students. Prof. Bhargava is the founder of the IEEE Symposium on Reliable and Distributed Systems, the IEEE conference on Digital Library, and the ACM Conference on Information and Knowledge Management.

The Research Track

Toward a Solution for an Energy Knowledge Graph



Dušan Popadić, Enrique Iglesias , Ahmad Sakor , Valentina Janev ,
and Maria-Esther Vidal 

Abstract Data integration demands the development of data management techniques to efficiently overcome interoperability issues and provide a harmonized view of both data and their meaning (i.e., metadata). This paper addresses the challenges of energy data management and integration and proposes a process of creating a knowledge graph, motivated by the needs of the stakeholders from Serbia and related to the integration of a large number of different renewable energy sources (RES) with the proprietary SCADA system of the Institute Mihajlo Pupin. The Energy Knowledge Graph (KG) has been built by reusing the energy-based semantic data model and the SDM-RDFizer, an open-source tool and interpreter of the W3C Recommendations Standard R2RML and its RDF Mapping Language (RML) extension. The data connectors implemented by the SDM-RDFizer plan the execution of the mapping rules and loading of the dataset to an RDF triple store to speed up the process of knowledge base creation. The Energy KG has been deployed on a Smart Grid Architecture Model (SGAM)—compliant platform hosted at the Institute Mihajlo Pupin.

Keywords Energy · Knowledge graph · Mapping rules · Application · Services

D. Popadić

School of Electrical Engineering, University of Belgrade, Belgrade, Serbia

e-mail: dusan.popadic@pupin.rs

D. Popadić · V. Janev (✉)

Mihajlo Pupin Institute, University of Belgrade, Belgrade, Serbia

e-mail: valentina.janev@pupin.rs

E. Iglesias · A. Sakor · M.-E. Vidal

L3S Research Centre, Leibniz University of Hannover, Hanover, Germany

e-mail: iglesias@l3s.de

M.-E. Vidal

e-mail: vidal@l3s.de

M.-E. Vidal

TIB-Leibniz Information for Centre for Science and Technology, Hanover, Germany

1 Introduction

After the announcement of the Google Knowledge Graph [1] in 2012, semantic technologies and knowledge graphs (KGs) gained on popularity and have been applied in many domains as many companies explore the technologies for competitive advantages, especially in the domain of integration of distributed resources over the Internet, e.g., for facilitating product/service discovery [2], managing business registers and company data [3], managing drug data [4], emergency management [5], or managing knowledge in the energy sector [6]. In this paper, the authors assess the applicability of the technologies for the energy sector.

1.1 *The European Electricity System*

The increased volume of data generated from distributed renewable data sources creates data integration and processing challenges in modern electrical systems. Therefore, there is a need to develop computational methods for ingesting, managing, and analyzing big data. More importantly, considering the bidirectional flow of information and energy in Smart Grids, knowledge needs to be extracted from this data, to uncover actionable insights. Hence, the future energy infrastructure will be based on intelligent power electronics, smart meters, context-aware devices, IoT, and AI-driven services. Interoperability problems caused by currently fragmented applications will be overcome in the new generation of grids, thus enabling data exchange between different players in the energy sector. For instance, the EU data strategy envisages the establishment of energy data spaces based on semantic web technologies and W3C standards. The information model proposed in the context of the International Data Space includes exemplary data models for describing datasets and services metadata needed to facilitate information search, service matching, and data exchange.

1.2 *Example Case Study*

The recently adopted EU energy-related strategies create opportunities to modernize the energy system, making it competitive and environmentally sustainable. Herein, we will use the example of the Serbian electricity system (see Fig. 1).

Because the national electricity infrastructure is not isolated, interoperability should be ensured at different levels (i.e., legislation, functional, syntactic, and semantic) and in different parts of the energy value chain, i.e., electricity generation, transmission, and consumption. The PUPIN SCADA system has been deployed at many parts of the national electricity grid. The system monitors and controls energy production, distribution, and usage with different objectives, including improving

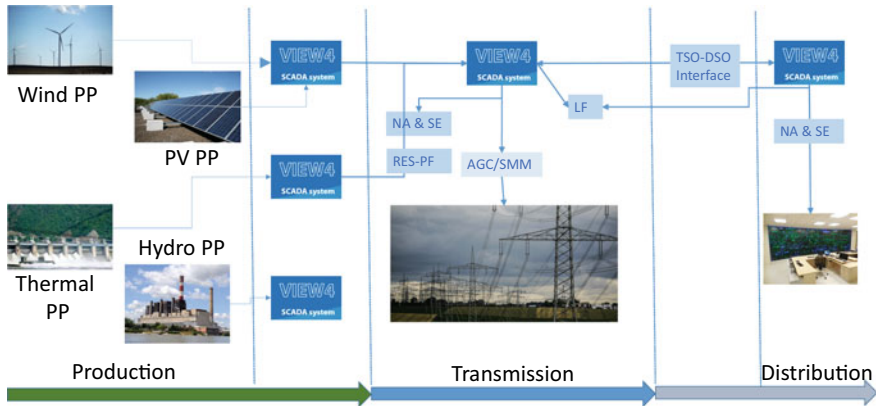


Fig. 1 The Serbian electricity system and the deployment of the VIEW4 SCADA system

energy efficiency, increasing flexibility and renewable generation share, and reducing energy costs. Therefore, the goal of the case study is to provide an innovative energy management service layer on top of existing SCADA based on reusable semantic models or knowledge graphs. They will facilitate the integration of data silos and their fine-grain semantic description; further, they will provide a common understanding of the energy domain based on existing vocabularies.

This paper comprises four additional sections. Section 2 presents the methodology followed by designing the pipeline for knowledge graph creation depicted in Sect. 3. The knowledge graph exploration tools are described in Sect. 4.

2 Achieving Semantic Interoperability

Interoperability and the possibility of building cross-border and cross-sector services are the focus of many initiatives in Europe, see, for instance, ISA² [7]. The high-level vision of the European Union for 2030 is to create a single internal market through a standardized law system transposed in the national legislation of all member states and a single European data [6] space for data exchange. In order to drive data-based innovations, standardization [8] should be applied, for instance, using metadata schemata, data representation formats, license terms for data and services, data integration [9], and data exchange approaches.

2.1 Research Questions

The Institute Mihajlo Pupin (PUPIN) currently hosts several SGAM (smart grid architecture model [10]) compliant service-oriented, cloud-based platforms that serve

for testing different energy-specific scenarios, see also [11]. Data exchange with external components (e.g., edge computers) is based on an adaptable gateway built upon OGEMA (open gateway energy management) framework. The data exchange within the broader EU energy ecosystem is still under elaboration. For instance, in the PLATOON project framework, the platform shall be integrated with the PLATOON marketplace based on the Industrial Data Space concept, i.e., using the IDS information model and Linked Data principles [12]. The Semantic Web community has developed more than 700+semantic vocabularies (see the LOV repository, <https://lov.linkeddata.es>). The aim is to analyze the standard schemas (i.e., vocabularies/ontologies) promoted by the community and adopt them for the targeted SCADA knowledge graph and services/applications. The following research questions guide our research:

- RQ1—Which are the concepts and properties that characterize the energy domain, and which ontologies cover the needs for modeling the electricity value chain and ensure uniform access to data collected with the proprietary SCADA system?
- RQ2—How to build a knowledge graph that will enable the development of services to support future energy marketplaces?

2.2 Selection of Semantic Models

In order to implement the “no-vendor lock-in” principle and ensure that future services will integrate smoothly with different legacy and proprietary solutions [13], the knowledge graph layer shall be based on open standards and open APIs. Therefore, in our research, one of the first steps toward developing the knowledge graph is the analysis of existing semantic models already in use, such as CIM, SAREF, SEAS [14], and DCAT, defined as follows:

- CIM—Common Information Model (CIM, https://ontology.tno.nl/IEC_CIM/), a standard developed by the electric power industry that has been officially adopted by the International Electro technical Commission (IEC); it comprises concepts (e.g., classes or relations) for software applications to exchange information about electrical networks.
- SAREF—Smart Appliances REference ontology (SAREF, <https://saref.etsi.org/saref4ener/v1.1.2/>). It is modular ontology for the Internet of Things domain; it integrates a family of vocabularies to represent smart cities, buildings, energy, agriculture, food, and environmental. SAREF4ENER is an extension for the energy domain; it includes majority of classes of interest for smart energy management.
- SEAS—Ontology developed in the framework of the Smart Energy-Aware Systems (SEAS, <https://w3id.org/seas/>) project with the aim of designing a global ecosystem of services and smart things collectively capable of ensuring the stability and the energy efficiency of future energy grids. SEAS includes features of interest and their properties, evaluation of features, smart and microgrids, smart homes, electrical cars, electrical market, and weather forecast.

- DCAT—The Data Catalog Vocabulary (DCAT, <https://www.w3.org/TR/vocab-dcat-2>) provides a common understanding of the classes and properties that describe a catalog of datasets and data services. DCAT is expressed in RDF and provides unified representation of catalog properties in a way that is understandable by humans, and also by machines. DCAT includes also classes from other vocabularies, e.g., *foaf:Agent*, *skos:Concept*, or *skos:ConceptSchema*.

2.3 Methodology

The work has been divided into the following phases:

- *Requirement Analysis* phase: The authors defined different business questions that we would like to answer with the knowledge graph.
- *Design* phase: Relevant concepts are selected for modeling. Then, data connectors toward the SCADA database and the messaging mechanisms are specified.
- *Specification* phase: The knowledge graph is specified in terms of RML rules.
- *KGs in Action* phase: The authors are involved in automating the semantic pipeline and developing exploration GUIs.

In this paper, the authors focus on the last two activities, namely, the implementation of the semantic pipeline and the exploration of the knowledge graph (Fig. 2).

3 Knowledge Graph Creation—The Semantic Pipeline

This section describes the process of knowledge graph creation and highlights the main challenges tackled in the work reported in this article. The process (injection,

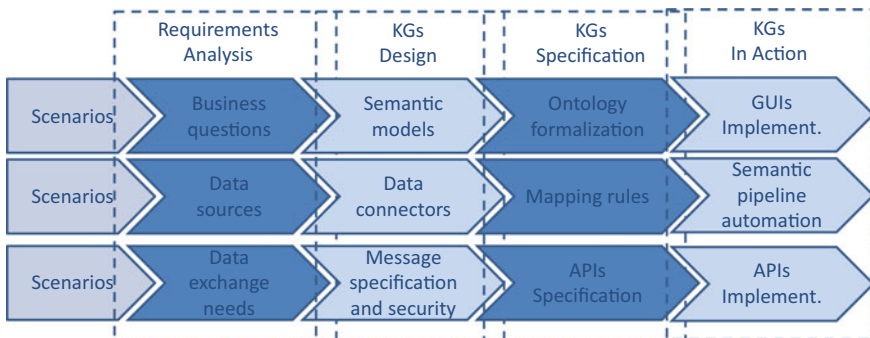


Fig. 2 Four-step methodology

transformation, and integration), also known as a semantic pipeline, is given in Fig. 3. There are two types of knowledge graph creation strategies:

- The Materialized Knowledge Graph Creation Process, i.e., data warehousing where the data are loaded and stored in an RDF format in a physical database, e.g., Virtuoso RDF triplestore.
- The Virtual Knowledge Graph Creation Process (i.e., Data Lake) where the data remains in the sources (in raw format) and is accessed as needed during query time.

We follow the first approach in order to experiment with (1) mechanisms for efficient search and visualization of energy data at different levels of granularity and (2) provide mechanisms for explainability and interpretability of results of the analytical services. The correspondences among energy data sources and semantic models are described using two mapping languages R2RML and RML, namely, the Relational to RDF Mapping Language (R2RML) [15] and the RDF Mapping Language (RML) [16]. As a result of the execution of R2RML and RML mapping rules, a knowledge graph expressed in RDF is created. Mapping rules are expressed as triples maps. Each triples map refers to a single logical source which can be SQL table or view or data gathered by executing SQL query against the input database. In our case, the mapping rules are applied to transform static data about plants, generation units, and weather stations, see Appendix. This data includes geographical location, control area membership, and similar data that are not changed frequently. Following examples of the mapping rules focus on PV plants. Since some of the data already exists in a MySQL database, this data is converted to RDF format using the RML-complaint engine, SDM-RDFizer [17]; it executes R2RML and RML mapping rules and transforms raw data in various formats: CSV, JSON, RDB, and XML, into an RDF graph knowledge graph. SDM-RDFizer resorts to data structures and physical

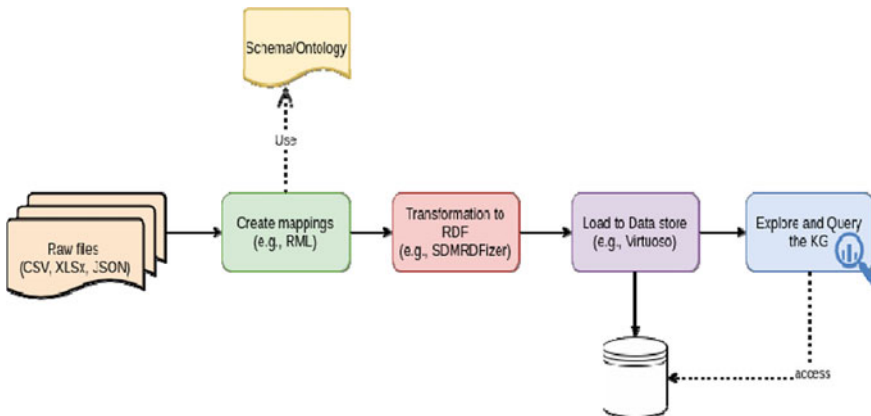


Fig. 3 Semantic pipeline

Table 1 The SCADA KG statistics

Statistics	Value
Total number of RDF triples	18,278,850
Number of classes	83
Number of distinct properties	156
Number of class/subclass pairs	12
Number of different timestamps for timestamped data	1,108,298

operators to scale up to large datasets, physical operators, and efficiently execute pipelines of knowledge graph creation.

Apart from static information about power plants and the grid, measured values from power plants are also collected. The data collected through the SCADA system is available in real time through a MySQL database; it includes power production forecast, power production measurements, and weather information (e.g., air temperature, wind direction, and solar panel temperature).

The SCADA knowledge graph (KG) is created as a result of execution of the mapping rules on top of the MySQL. By the time of this submission, the SCADA KG comprises more than 18M RDF triples, with instances of 83 classes. These classes are described in terms of 156 properties and more than 1M timestamps. Table 1 reports on the characteristics of the current version of the SCADA KG.

4 Knowledge Graph Exploitation

This section presents the services implemented on top of the SCADA KG; they allow for the exploration of the integrated data and their descriptions with the energy semantic data models. SPARQL, the W3C recommendation query language, is utilized to express basic queries against the SCADA KG.

4.1 Energy Analytics Dashboard

Since SCADA KG shall work in synergy with various AI-based analytic services and help users to understand results, a visualization tool (EAD—energy analytics dashboard) has been developed. The tool allows fetching data from arbitrary SPARQL end points and supports different analysis/visualization options.

EAD is a data visualization tool that works on top of the SCADA KG. It allows the users to select the data of interest, compare time series (i.e., forecasted load and actual load at that time), and visualize summary statistics on the geographical map. It has been implemented as a web application using JavaScript programming language

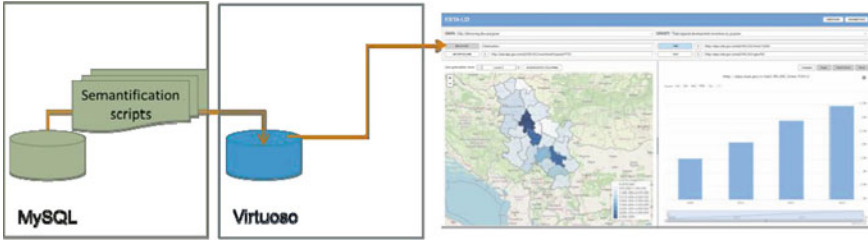


Fig. 4 Semantic pipeline and KG exploration

with help of JQuery library. It uses the Highcharts library for visualization (<https://github.com/highcharts/highcharts>) and the Leaflet library for interacting with geo data (<https://leafletjs.com/>). Figure 4 depicts the dashboard and its connection with the pipeline of knowledge graph creation is described in Sect. 3.

4.2 Alignment with EU Initiatives

In order to inherently address interoperability, the CEN-CENELEC-ETSI Smart Grid Reference Architecture [10] framework defines five interoperability layers, where the information layer specifies the business context and the semantic understanding. Hence, in future energy smart grids, the technologies described herein are not optional, but mandatory. Currently, under development represents different energy services marketplaces that in their core include components such as vocabulary management tools and dataset/service registries. In case the production of all PV plants in Serbia can be reached via a SPARQL query, with one click, we can answer the following question “*Show the total energy produced by PV plants in Serbia*”.

```
SELECT DISTINCT ?solararray SUM(?value) as ?totalPower
WHERE {
    ?solararray a seas:SolarArray .
    ?solararray art:country <https://projekat-artemis.rs/Country/RS>.
    ?panel seas:isMemberOf ?solararray .
    ?panel a seas:SolarPanel .
    ?panel seas:producedElectricPower ?activePowerProperty .
}
```

5 Conclusions

The reusability of energy services is limited due to different representations of data used by different stakeholders in the energy value chain. Therefore, this paper proposes an approach for building a knowledge graph enabling semantic interoperability. The semantic data models from the energy sector and the internal SCADA information model are currently used as an information hub materialized in a knowledge graph. It provides the basis for developing and integrating services in the Energy Data Spaces. Additionally, this layer provides the basis for the explainability of machine learning services/analytical applications installed in the smart ecosystem.

The future work includes activities that will connect the PUPIN platform with the PLATOON marketplace, thus creating opportunities for broader exploitation of the PUPIN analytical services.

Acknowledgements This research has received funding from EU H2020 Research Program (GA No. 872592, GA No. 952140) and the Republic of Serbia (MPN, No. 451-03-9/2021-14/200034; Innov. Fund, Artemis, No. 6527051).

Appendix

```
@base <https://projekat-artemis.rs/> .
<#ARTEMIS_DB> a d2rq:Database;
<#PUPIN_PVPlantMapping> a rr:TriplesMap; rml:logicalSource [
    rml:source <#ARTEMIS_DB>; rr:sqlVersion rr:SQL2008; rml:query """

SELECT DISTINCT

    plants.id AS plant_id, plants.name AS plant_name, weater_locations.lat AS lat,
    weather_locations.lon AS lon, weather_locations.city AS city, assets.asset_name AS asset_name,
    country.country_code AS ccode,eic_functions.eic_type_function_acronym AS eic_func_acronym,
    organization.organization_short_name AS organization_short_name, organization.organization_name
    AS organization_name,assets.id AS asset_id

FROM

    `plants`

JOIN weather_locations ON plants.weather_location_id = weather_locations.id

JOIN assets ON plants.asset_id = assets.id

JOIN organization ON assets.organization_id = organization.id

WHERE

""""
];
```

References

1. Noy N, Gao Y, Jain A, Narayanan A, Patterson A, Taylor J (2019) Industry-scale knowledge graphs: lessons and challenges. *Commun ACM* 62(8):36–43
2. Jain S (2020) Exploiting knowledge graphs for facilitating product/service discovery (2020). [arXiv:2010.05213](https://arxiv.org/abs/2010.05213)
3. Roman D, Alexiev V, Paniagua J, Elvesæter B, Marius von Zernichow B, Soyulu A, Simeonov B, Taggart C (2022) The euBusinessGraph ontology: a lightweight ontology for harmonizing basic company information. *Semant Web J* 13(1):41–68. IOS Press
4. Lackshen G, Janev V, Vraneš S (2021) Arabic linked drug dataset consolidating and publishing. *Comput Sci Inf Syst* 18(3):729–748. <https://doi.org/10.2298/CSIS200510047L>
5. Jain S, Meyer V (2018) Evaluation and refinement of emergency situation ontology. *Int J Inf Educ Technol* 8(10):713–719
6. Janev V, Vidal ME, Endris K, Pujić D (2021) Managing knowledge in energy data spaces. In: Companion proceedings of the web conference 2021 (WWW '21 Companion), April 19–23, Ljubljana, Slovenia. ACM, New York, NY, USA, 11 pp. <https://doi.org/10.1145/3442442.3453541>
7. European Commission: interoperability solutions for public administrations, businesses, and citizens. <https://ec.europa.eu/isa2/>. Last accessed 21 Mar 2022
8. European Commission: the rolling plan on ICT standardisation (2020). <https://ec.europa.eu/digital-single-market/en/news/rolling-plan-ict-standardisation>. Last accessed 21 Mar 2022
9. European Commission: European data strategy, COM (2020) 66 final. <https://ec.europa.eu/digital-single-market/en/policies/75981/3489>. Last accessed 21 Mar 2022
10. CEN-CENELEC-ETS: smart grid reference architecture, November 2012
11. Pujić D, Jelić M, Tomasević N, Batic M (2020) Chapter 10 case study from the energy domain. In: Janev V, Graux D, Jabeen H, Sallinger E (eds) *Knowledge graphs and big data processing*. Lecture notes in computer science, vol 12072. Springer International Publishing, pp 1–208
12. Berners-Lee T, Design issues: linked data. <http://www.w3.org/DesignIssues/LinkedData.html>. Last accessed 21 Mar 2022
13. Cuenca J, Larrinaga F, Curry E (2020) DABGEO: a reusable and usable global energy ontology for the energy domain. In: *Web semantics: science, services and agents on the world wide web*, vol 61–62. <https://doi.org/10.1016/j.websem.2020.100550>
14. Lefrançois M, Kalaoja J, Ghariani T, Zimmermann A (2017) The SEAS knowledge model. ITEA2 12004 smart energy aware systems deliverable 2.2
15. Das S, Sundara S, Cyganiak R (2012) R2RML: RDB to RDF mapping language. In: Working group recommendation, World Wide Web Consortium (W3C)
16. Dimou A, Vander-Sande M, Colpaert P, Verborgh R, Mannens E, Van de Walle R (2014) RML: a generic language for integrated RDF mappings of heterogeneous data. In: *Proceedings of the 7th workshop on linked data on the web*, CEUR workshop proceedings. CEUS
17. Iglesias E, Jozashoori S, Chaves-Fraga CD, Vidal M (2020) SDM-RDFizer: an RML interpreter for the efficient creation of RDF knowledge graphs. In: *The 29th ACM international conference on information and knowledge*. ACM

Web-Based Visualization and Analysis Framework for Graph Data



Fatmana Şentürk , Mehmet Ali Bilici , Sezercan Tanışman ,
and Vecdi Aytaç 

Abstract Graph theory has many applications in computer science. Graphs are involved in the mathematical modeling of problems belonging to very popular computer and data science, such as data mining, data clustering, computer network, image segmentation, etc. Graph theory is used in computer science to solve problems modeled as graphs. However, many applications do not have an interface that displays and edits graphs instantly. Moreover, these applications are not supported with a web interface. In this study, a web interface has been developed that support graph theory concepts, display more than one graph simultaneously, have a user-friendly interface, and include algorithms to analyze graphs. Also, we present a web service for these graph analyzing algorithms.

1 Introduction

Graph theory has several applications in computer science. Graphs are involved in the mathematical modeling of problems belonging to very popular computer science research topics, especially data mining, clustering, computer network, segmentation of an image, or capturing images. For instance, a data structure can be designed as a tree of vertices and edges. Similarly, when it is desired to create a model of network topologies, graph concepts can be used. In addition, the concept of graph painting, which is one of the graph processing methods, can be used for resource allocation and programming, which is important for many institutions and organizations.

F. Şentürk (✉)

Computer Engineering Department, Pamukkale University, Denizli, Turkey
e-mail: fatmanas@pau.edu.tr

M. A. Bilici · S. Tanışman · V. Aytaç
Computer Engineering Department, Ege University, İzmir, Turkey
e-mail: mehmet.ali.bilici@ege.edu.tr

S. Tanışman
e-mail: sezercan.tanisman@ege.edu.tr

V. Aytaç
e-mail: vecdi.aytac@ege.edu.tr

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
S. Jain et al. (eds.), *Semantic Intelligence*, Lecture Notes in Electrical Engineering 964,
https://doi.org/10.1007/978-981-19-7126-6_2

Diagrams can appear as UML class, activity, use case, business process, work-flow charts, or any diagram [1]. This variety of representation shows that the diagrams can be used in the modeling phase for any problem or in the solution step of the modeled problem.

The development of graph-based algorithms to find a way out of problems that are modeled as graphs highlights the importance of graph theory in computer applications. These algorithms generally combine the problems of daily life, theoretical graph concepts and computer science to reach the result. However, the vast majority of these algorithms focus only on the solution of the problem and either overlook or ignore the visualization of the graph in which the problem is modeled.

In addition, several computer languages exist to implement and enable graph theory operations. Examples of graph theoretical languages are SPANTREE, GASP, GRASPE, ...etc [2, 3]. Although these languages support graph theory concepts, they do not have an interface that displays graphs and allows them to be edited instantly. Moreover, none of these languages support the XML-based file format GraphML, nor does it have a web interface that allows editing of diagrams. These languages usually keep graph structures in the form of lists or matrices. This form of storage makes the diagram almost completely incomprehensible.

In addition, some of today's current problems include very large data. Graph algorithms can be used to solve these problems, but this becomes very difficult when graph sizes increase. These huge graphs cannot be processed especially by computers with low RAM capacity and low processing power. A service-based architecture is needed to eliminate physical constraints such as RAM and processor and to perform certain operations.

Considering the existing shortcomings, in this study, a framework that supports graph theory concepts, can display more than one graph at the same time, has a user-friendly interface, includes algorithms that will enable various analyzes on the graph, and web service has been developed. The following operations can be performed via the developed framework:

- Users can create any graph,
- The graph created with the interface can be stored as a GraphML file and the stored files can be opened and processed again
- Operations such as zooming, labeling, and weighting the edges can be done on the generated graph,
- Graphs taken from the GraphML extension file or created via the web interface can be combined
- All algorithms given in Sect. 5 can be operated on the generated diagram. In this way, both complex problems can be solved and the graphs modeled for these problems can be visualized.

The remainder of the paper is laid out as follows: The fundamental concepts of the graph are discussed in Sect. 2. In Sect. 3, the existing studies in the literature are presented. In Sect. 4, the technical details of the developed Framework are given, and in Sect. 5, the graph algorithms presented by the Framework are given. In the last section, the results obtained and future studies are included.

2 Preliminary

It will be useful to know some basic definitions before moving on to the features of the developed framework. For this reason, this section is reserved for identification.

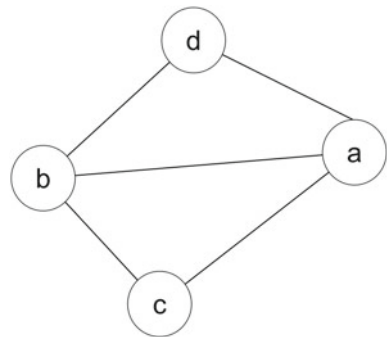
V is the non-empty set of vertices, and E is the set of edges connecting these vertices; $G = (V, E)$ is the definition of a G graph. Each edge connects one or two vertices [4]. For example, graph G with the set of vertices; $V = \{a, b, c, d\}$ and a set of edges; $E = \{\{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}\}$ is given in Fig. 1. In order to avoid errors in the expression part, two different vertices in a graph can be shown as u and v , and the edge between these two vertices can be shown as uv .

A G graph is considered a simple graph if it contains just one edge connecting two vertices and no edge connecting a vertex to itself [4]. This graph is considered a non-simple graph if there are more than one edge connecting two vertices, or if a vertex is related to itself via another vertex.

Let G be an undirected graph. If the vertex set defined in the graph G is V and the edge set formed by the edges between these vertices. It is denoted by E . $|V|$ and $|E|$ indicate the number of vertices and edges, respectively. If there is a link between i and j vertices indicated as $i \leftrightarrow j$, and these vertices are called adjacent [5]. In a G graph, if there is a link between i and j vertices; that is from i to j and from j to i , this kind of graph is known as an undirected graph. In a directional graph, there is a direction expression on the edges. If there is a link from i to j , this link is unidirectional.

The path is called a finite sequence of edges followed from one of the vertices in a graph to reach another vertex. The number of edges followed to arrive at the vertex to be reached is also called the path length [4]. For example, $G = (V, E)$ formed by the edges connecting two adjacent vertices in graph $v_0v_1, v_1v_2, \dots, v_{k-1}v_k \in E$, the finite edge array of the form gives the path between v_0 and v_k . In a G graph, if u and v are vertices and there is a path from u to v , u , and v are called connected. If the path length between u and v is 1, these vertices are called adjacent. If a vertex has an edge or a set of edges connecting to itself, this edge or a set of edges is called a loop.

Fig. 1 Graph G



If all vertices in a graph have an edge on all other vertices, that graph is a complete graph. The K_n represents a complete graph, with n denoting the number of vertices in the graph. If there is exactly one edge connecting each pair of vertices in a complete graph, it is referred to as a simple graph. Sometimes, two graphs can exist in similar forms. In this case, we can say that these two graphs are isomorphic [4]. For isomorphic graphs: Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are simple graphs; If there is a one-to-one function f for V_1 and V_2 and a and b are adjacent in G_1 ; If $f(a)$ and $f(b)$ are adjacent in G_2 , these two graphs are isomorphic.

Some problems require a different representation of graphs. One of them is the creation of a planar view. The ability to draw any edge in a G graph in a plane without intersecting with another edge is called planarity, and the graph is called a planar graph.

A simple connected graph without a circuit is called a tree. Some graphs consist of more than one self-connected but not completely connected part of the graph. If the connected parts that do not contain a simple circuit are in the form of trees, the structure formed by these connected parts is called a forest.

Let G be a connected graph. Let's subtract a vertex set V' from G . By doing this, if the graph G becomes disconnected, this set of vertices V' is called the cut vertices set. If an E' edge set is removed from graph G to make the graph disconnected, that edge set is called a cut edges set [4].

For any network graph, the measurement of the resistance of the graph against the deterioration that occurs at certain vertices or certain edges of the graph is called vulnerability measurement [6].

The coloring of G graph nodes ensures that no two neighboring nodes have the same color [7].

3 Literature Background

Graph theory and its applications are used to solve many problems in daily life. For this reason, many applications and special programming languages have been developed to supplement graph theory concepts and facilitate the solution of problems. Furthermore, these developed languages enable users to express or represent graph operations in a compact and intuitive manner. Some graph-theoretic languages are examined by Shirinivas et al. [2]. Although the graph-theoretic languages examined in this study support graph theory concepts, most of them do not have an interface or web interface that displays and organizes graphs instantly. In most of these languages, graphs are kept in the form of lists or matrices. Therefore, the intelligibility of the graph is very difficult, including for the people modeling the graph. In addition, none of these languages support the new XML-based file format, GraphML.

The use of computer systems to represent graphs, as well as other parts of graph theory, such as graph-theoretic data structures such as list and matrix structures, are discussed [3]. The most popular applications of graph theory in computer science

were also explored in this paper. However, it did not propose a method and application for a new graph representation and analysis.

There has also been work on processing large-scale graphs [8]. New theories have been presented to investigate the structure and dynamics of large-scale networks quantitatively. However, considering the difficulties of visualizing large-scale graphs, fragmentation of graphs is proposed as a solution. Algorithms and methods with different usage areas are investigated to visualize large and complex graphs.

Visualization of graphs is especially important for large and complex graphs [9]. For this purpose, different applications have been developed in order to express the diagrams in a more understandable. In other words, different software has been developed to draw complex, overlapping edges and vertices in a graph in more clear. These developed applications are grouped by the four common properties of partitions, attributes, time, and space [10]. In particular, algorithms were developed to reduce complexity in graph visualization by using techniques such as compression and summarization [5].

Among the applications developed for graph algorithms, graphs are also visualized for different purposes such as showing the paths in the graph and finding frequently used sub-parts in the graph. These applications can be grouped as visualization and exploration techniques (graph sampling, graph filtering, graph partitioning, graph clustering), global and local views, sub-graph mining tools, hybrid graph visualization [11]. However, a web interface and web service that offers all of these applications together have not been encountered.

Gene datasets are visualized with the application called KeyPathwayMinerWeb. The genes in the data sets are colored according to their frequency of use, and the information of intended any specific gene is shown in detail. With this application, operations such as visualization, data merging, and enrichment of data can be done [12]. However, this application can perform only gene dataset-specific operations. There are no basic graph algorithms such as obtaining a planar view of graphs or vulnerability of graphs.

Graph algorithms can be applied to many fields and one of these is to model the ground data as graphs using the Prov-N data set. Vertex similarities can be calculated and two graphs can be merged, with the developed application [13].

GRAD [1] is a Java library developed for quick and easy drawing. This library offers a different number of graph drawing algorithms to be used, including algorithms that have not yet been implemented in Java, in addition to the existing drawing algorithms. GRAD allows the application of an automatically selected graph layout algorithm in accordance with the structure of the graph drawn by the users. That is, if the drawn graph is in the tree structure, GRAD automatically draws this graph as a tree. Also, GRAD offers algorithms related to planarity, shortest path (Dijkstra), and cycle structure. However, it does not provide features such as minimum color, bipartite, n paths.

Graphiti, on the other hand, is a platform that allows modeling any network model as a graph, allowing users to simply define relationships or sets of relationships in the network. Graphiti allows the edges between two vertices to be more than one and each of these edges can be defined as different properties [14]. However, there

are no algorithms on the proposed system that can make network analysis such as vulnerability analysis. JUNG [15] is a library developed in Java. It provides a standardized and flexible vocabulary for modeling, analyzing, and visualizing data that is represented as a graph or network. JGraphX [16] is a visualization tool that supports the vertex-edge structure. It provides support for different applications such as flow charts, organization charts, workflow modeling, UML diagram, electronic circuit design. Prefuse [17] is a java based interactive information visualization application. Prefuse provides programmers with components quickly to create and customize working visualizations [18]. Although these libraries are impressive libraries, they offer basic algorithms such as depth-first search (DFS), Dijkstra [1]. However, these libraries do not support algorithms such as graph coloring and vulnerability.

Toytree is a Python library for visualizing and manipulating tree-based data structures. All shapes created with Toytree can be exported to PNG, PDF, and SVG formats [19]. However, since Toytree is a library written in python language, it does not have an interface suitable for end-users. In addition, it is not possible to run algorithms for other graph types, since operations can only be performed on the tree.

DynaGraph is an interactive web application. With this application, it can visualize multiple datasets, fast and easily. However, the interaction of DynaGraph with the user is limited, for example, it cannot handle two graphs interactively. Also, algorithms that can be applied to graphs are limited [20].

Considering all these existing libraries and applications, there is no application that users can edit on, supports the graphML data format, and offers different algorithms such as bipartite, graph coloring, vulnerability (detailed in Sect. 5) on a single platform. So, we have developed a web interface that offers many different graph algorithms, where end-users can draw a graph and manipulate graphs. In addition, we have developed a web service to use these graph algorithms for other programmers.

4 Proposed Architecture

In this paper, a web application and a web service have been proposed that supports the visualization of graphs and graph algorithms.

The proposed framework in this study consists of two different modules: web interface and web service. Through the first of these, the web interface, users can create their own directed or undirected graphs and operate different graph algorithms on the system on these graphs. At the same time, they can view the results produced by these algorithms through this web interface. Users can upload their previously produced graphs files (txt, XML, GraphML format) to the web interface, and add/remove vertices and edges on these diagrams. They can download a graph they designed in the web interface to their computers in GraphML [21] format, which is a language developed for storing graphs similar to XML file types. Also, through the developed interface, users can transfer the file in GraphML format. The system has requested this file once time and graph algorithms have been able to reduce the

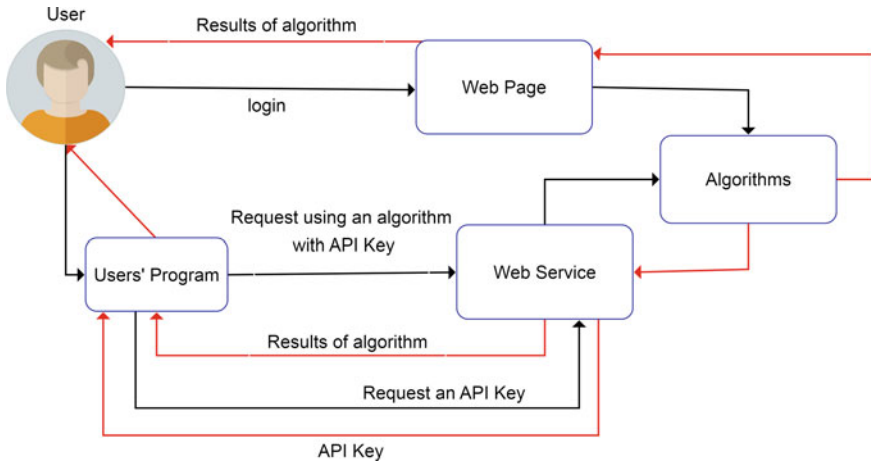


Fig. 2 A general work flow of the web application and web services

server’s response time and bandwidth usage by repetitively processing using this file that is uploaded to the system once.

The second part of the proposed framework is a web service module. Especially for software developers, there is not a comprehensive library where graphs can be displayed visually and graph algorithms can be operated on graphs. This developed web service enables programmers to both visualize graphs and operate more complex algorithms such as bipartite test, planarity view of a graph, vulnerability of a graph, shortest path algorithms, etc. For the use of the web service, it is needed to request an API key from our system. While using web services, it should be sent users’ own graphs along with API Key. The selected graph algorithm is executed on the server and the result of the operation(shortest path, view of graph, etc.) is returned to the users. The general workflow of our developed system is given in Fig. 2

For the developed interface, WebGL 1.0,¹ customized for the HTML5 platform of the open-source, platform-independent OpenGL ES 2.0 graphics library developed by the Khronos group was used. Thanks to its high performance, WebGL enables large-scale graphs to be easily supported by the interface.

Graph algorithms are implemented on the ASP.NET web platform with C# 6.0 language on .NET Framework 4.5. All graph algorithms are defined as RESTful API and have the ability to work without any interface. Microsoft SQL Server is used as the database server for the developed Web API, and also, EntityFramework 6.0 is used for accessing and communicating with the database.

¹ WebGL , <https://www.khronos.org/webgl/>, Accessed: 26.01.2022.

5 Methods

In the proposed framework, not only computer science problems, but also problems related to other basic sciences such as chemistry, biology, and medicine can be solved thanks to the algorithms. For example, although the concept of Closeness centrality seems like a theoretical graph parameter, it can be applied in a wide range from social networks to cancer research [22]. Therefore, a web interface and service have been developed for the problems that can be solved by modeling as a graph in other disciplines.

5.1 Layout Algorithms

Generating the view of a graph is of great importance in terms of visually representing the problems. When the view of a graph is created, users can more easily calculate or envision some operations such as finding connected vertices or finding the adjacent vertices, finding paths of length n, etc. For this reason, it should be visualized a graph, created or opened, on the web application.

In the web interface, vertices and edges can be added to a graph, users can give weight to these added edges, and also users can label vertices. Users can zoom in/zoom out, or scroll the vertices and edges of the graph. An example view of the web interface is given in Fig. 3. Also, two or more graphs can be copied and merged with each other. In the web interface, graphs can be displayed irregularly as they are drawn, or they can be viewed as a grid or planar according to the structure of the graph.

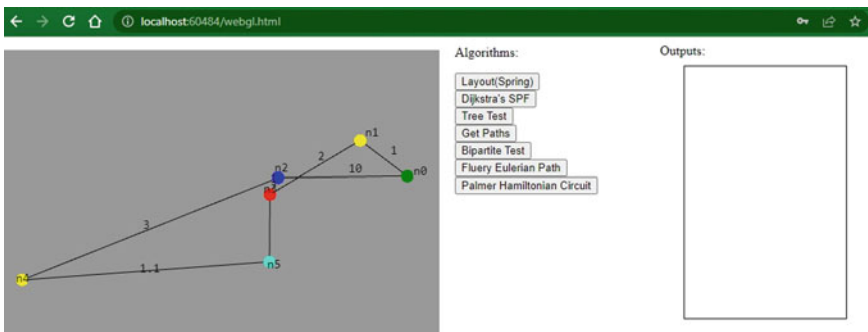


Fig. 3 The view of a drawn graph by using the web application

5.1.1 Grid View

In the grid algorithm, the vertices and edges are placed on the screen rectangularly to give a grid image. The algorithm resizes so that every component (vertices and edges) of the graph is the same size and draws the graph to the screen. Every graph that users work on can be displayed as grids through the developed framework. An example view of the web interface after grid algorithm, are executed is given in Fig. 4.

5.1.2 Planar View

A planar view of a graph is drawing that graph without overlapping its edges. In other words, while obtaining the view of the graph, the different edges should be drawn without intersecting with each other at any point except vertices. If the drawn graph by users is planar, a planar view of the graph can be obtained via this method. Before the planar view of the graph is drawn in the framework, the algorithm is tested whether the graph is planar or not. If this planarity test [23] is positive, the planar view of the graph is drawn on the screen.

5.2 Findings Paths and Cycle

A path can be defined as the edge or set of edges that connect two vertices on a graph. Finding paths in a graph can be used especially in solving different problems such as vehicle routing problems, routing optimization problems, finding network similarities, etc. For this purpose, the following algorithms for finding paths on the

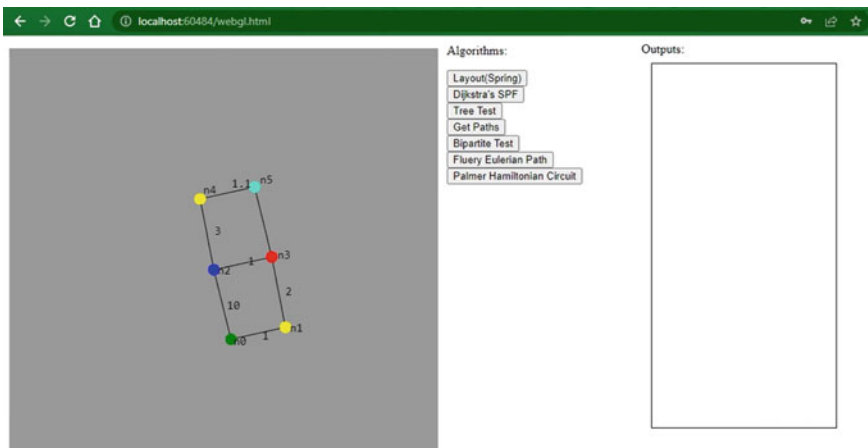


Fig. 4 The view of a graph after grid algorithms

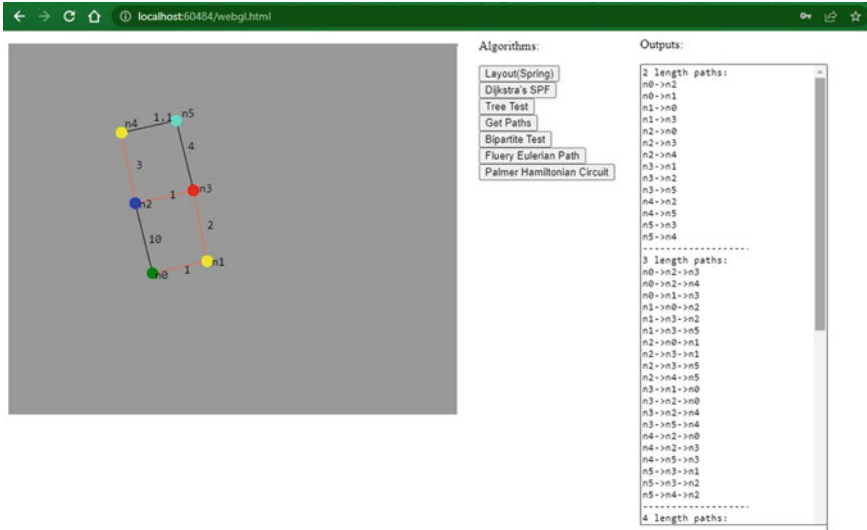


Fig. 5 An example view after the n-paths algorithm is executed

graphs can be executed on the web interface. An example view of the web interface after the n-paths algorithm is executed is given in Fig. 5.

5.2.1 Backtracking Algorithm

Backtracking is a recursive algorithm that tries to reach the final solution step by step by adding a new vertex to the solution set at each step. The algorithm basically works like this, any vertex is added to the candidate solution set, if the added vertex does not lead to the final solution, it is deleted from the candidate solution set. The next vertex, which is not in the candidate solution set, is included in the candidate solution set. In this way, a final solution set is created by trying all possible combinations. In particular, this algorithm can be used to find the Hamilton circuit. If it is possible to return to the starting vertex by starting from any vertex on the graph and visiting all vertices only once, it is called a Hamiltonian Circuit. The path is used while moving on the vertices is called the Hamiltonian Path [4]. The Hamiltonian circuit can be used to create the routes of cargo distribution companies.

5.2.2 Depth First Search

Depth-first search (DFS) algorithms are an algorithm used to search on graphs. Starting from a selected initial vertex, one goes up to the deepest vertex that can be moved on the graph. When there is no edge to move forward, the algorithm goes

back to the vertex on the graph [24] and moves to another vertex to find the deepest. The DFS algorithm can be used to find the furthest distance to travel on a route.

5.2.3 Fleury’s Algorithm

If it is possible to return to the initial vertex by traversing all the edges on the graph once, this is called the Eulerian Circuit, and the route used while doing this process is called the Eulerian Path. Fleury’s algorithm [25] can be used when finding the Eulerian path and circuit in a graph. For example, the Eulerian path can be used in bioinformatics to reconstruct DNA sequences [26].

5.2.4 Finding Shortest Path

In various applications, such as identifying directions on a map, networking, or telecommunication routing, it is important to discover the shortest route (shortest path) from one vertex to another vertex in a graph. The shortest route may be found using the Dijkstra’s algorithm [27]. For example, for graph G given in Fig. 6, the shortest path algorithm between n_0 and n_4 can be found by traversing the nodes $n_0, n_1, n_3, n_2,$ and n_4 , respectively. This node list can also be seen in the web interface, with the edge costs.

5.2.5 Findings Cycle

In addition to all these algorithms, Ore’s theorem [28] can be used to find the Hamiltonian circuit and the algorithm [29] developed by Hierholzer, and Wiener can be used to find the Eulerian circuit. Finding the cycle basis in a graph can be accomplished

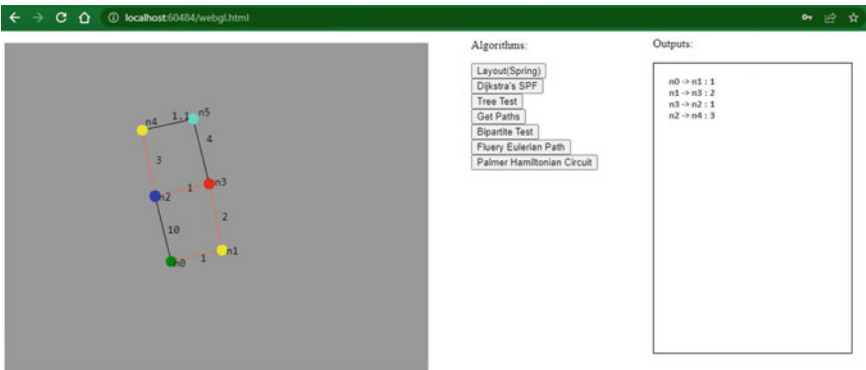


Fig. 6 An example view after executed shortest path algorithm

with the Paton algorithm [30] while finding simple circuits in a graph is found by the algorithm developed by Tiernan [31]. Also, these all methods are implemented in the framework.

5.3 *Bipartite*

Bipartite graphs are a special case for graphs. If all non-adjacent vertices in a G graph can be divided into two different sets, this graph is called a bipartite graph. For example, if a problem in which m jobs must be assigned to n employees is modeled as a graph, the resulting graph will be a Bipartite graph. In addition, maximum flow problems in a network can be solved by bipartite graphs.

In order to find out whether a graph is a bipartite graph or not, the graph coloring algorithm can be used, but it must be used in only two-color for graph coloring. According to this approach, if the graph can be painted with two colors, it is bipartite. If it cannot be painted, this diagram is not bipartite.

5.4 *Isomorphism*

It is called isomorphic graphs when two graphs show similar properties to each other. That is, graphs that are similar in items(vertices, edges, count of these components, etc.) of the structure are isomorphic. There are many ways to show that two graphs are isomorphic to each other [4]. It is often easier to show that two graphs are not isomorphic to each other than to check for all variants to check for isomorphism. For this reason, the following qualities are primarily considered to check isomorphism:

- Vertices number in the graphs: The number of vertices must be equal to each other, in isomorphic graphs.
- Edges number in the graphs: The number of edges in the graphs must be equal for two graphs that are isomorphic to each other. If these edge numbers in the graphs are different from one, they do not obey the isomorphism rules because the representation and appearance of the graphs will be different from each other.
- Finding the vertex degrees: While controlling the isomorphic conditions, in some cases, it may not be enough to have the vertex numbers and edge numbers equal in both graphs. For this reason, it is necessary to control each vertex degree in the graphs. So, the vertex degrees in a graph are calculated for each vertex and a vertex corresponding to this vertex is searched in the other graph. When searching for vertex, the only condition to be considered is that if a vertex has been paired with another vertex before, that vertex should not be used again. If for each vertex in the graph, there is a corresponding vertex to the other graph, these two graphs can be isomorphic.

- If the adjacency matrices of both graphs are similar, these two graphs can be isomorphic.

These two graphs are not isomorphic if any of the above-given qualifications are different.

5.5 *Minimum Graph Colour*

The painting of a graph's vertices is known as graph coloring. While doing this coloring process, it is aimed to paint the graph using minimum color without assigning the same color to two adjacent vertices. Some problems can be solved by graph coloring methods. These problems can be described as the scheduling of exams such that no individual is required to participate in two exams simultaneously, and the storage of chemicals such that no two mutually dangerous chemicals, etc [7].

5.6 *Tree Test*

Trees are a special representation of graphs and a data structure very often used in computer science. For example, people's family relationships can be modeled as a tree. Just as with the isomorphism check, tests are done to check if a graph is a tree. These tests follow:

- The number of edges in the graph must be 1 less than the number of vertices.
- Graph must be connected.
- The graph must not contain any cycles.

All the methods described in this section can be used to solve different problems and we offer all these methods as a web service. In addition to these methods, It can be calculated in many different graph parameters such as connectivity [4], binding number [32–34], closeness centrality [35, 36], vulnerability [37, 38].

6 **Conclusion**

In this study, a GraphML-supported, also XML-supported, user-friendly, web interface and a web service framework have been developed. The developed web-based framework of highlights is a graph analysis tool that can be easily accessed by every programmer and user. Any user can access this framework from their computer, tablet, or even mobile phone.

We believe that such a web interface can be easily used by many researchers, whether they are computer scientists or not, who can model the data set as graphs.

Therefore, it will be an application that will serve not only computer sciences but also other branches of science. For example, a medical scientist uploads a graph modeling the drug-drug interaction to this application and asks, "Is there a relationship between the two drugs?", and to seek the answer via the web interface. If there is any path, the scientist will conclude that it is not appropriate to take these two drugs at the same time. The developed application offers a more comprehensive algorithm package than existing applications and architectures. In this way, it is expected that there will be a wide range of users as it includes different analyzes that users may need.

In the developed framework, in addition to the Web interface, there is also a web service available to programmers. With this web service, it may be possible for other institutions and organizations to use the graph algorithms they need. In this way, programmers will be able to take and use some algorithms that take time to write and test and they will save time. This web service can be used via an API Key, thus, it can be used only for the person/institution, and user restrictions can be set.

The usability of the application in terms of the user interface will be tested with different devices such as phones, tablets, and computers. The interaction of the web service with different platforms and languages, limits such as time and memory will be investigated. In addition to the visualization of many graph data, this developed the platform can also be used for ontologies.

Acknowledgements This study was supported by Ege University Scientific Research Projects Coordination Unit (Project number: FGA-2019-20589).

References

1. Vaderna R, Dejanović I, Milosavljević G (2016) Grad: a new graph drawing and analysis library. In: 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), pp 1597–1602. IEEE
2. Shirinivas S, Vetrivel S, Elango N (2010) Applications of graph theory in computer science an overview. *Int J Eng Sci Technol* 2(9):4610–4621
3. Riaz F, Ali KM (2011) Applications of graph theory in computer science. In: 2011 Third International Conference on Computational Intelligence, Communication Systems and Networks, pp 142–145. IEEE
4. Rosen KH, Krithivasan K (2012) Discrete mathematics and its applications: with combinatorics and graph theory. Tata McGraw-Hill Education
5. Hu Y, Shi L (2015) Visualizing large graphs. *Wiley Interdiscip Rev: Comput Stat* 7(2):115–136
6. Ammann P, Wijesekera D, Kaushik S (2002) Scalable, graph-based network vulnerability analysis. In: Proceedings of the 9th ACM conference on computer and communications security, pp 217–224
7. Leighton FT (1979) A graph coloring algorithm for large scheduling problems. *J Res Natl Bur Stand* 84(6):489–506
8. Chin SP, Reilly E, Lu L (2012) Finding structures in large-scale graphs. In: *Cyber Sensing 2012*, vol 8408. International Society for Optics and Photonics, p 840805
9. Han D, Pan J, Zhao X, Chen W (2021) Netv. js: a web-based library for high-efficiency visualization of large-scale graphs and networks. *Vis Inf* 5(1):61–66

10. Hadlak S, Schumann H, Schulz HJ (2015) A survey of multi-faceted graph visualization. In: Eurographics Conference on Visualization (EuroVis), vol 33. The Eurographics Association Cagliari, Italy , pp 1–20
11. Pienta R, Abello J, Kahng M, Chau DH (2015) Scalable graph exploration and visualization: Sensemaking challenges and opportunities. In: 2015 International Conference on Big Data and Smart Computing (BIGCOMP), pp 271–278. IEEE
12. List M, Alcaraz N, Dissing-Hansen M, Ditzel HJ, Mollenhauer J, Baumbach J (2016) Key-pathwayminerweb: online multi-omics network enrichment. *Nucleic Acids Res* 44(W1):W98–W104
13. Kohwalter T, Oliveira T, Freire J, Clua E, Murta L (2016) Prov viewer: a graph-based visualization tool for interactive exploration of provenance data. In: International provenance and annotation workshop. Springer, pp 71–82
14. Srinivasan A, Park H, Endert A, Basole RC (2017) Graphiti: Interactive specification of attribute-based edges for network modeling and visualization. *IEEE Trans Vis Comput Graph* 24(1):226–235
15. JUNG: The java universal network/graphframework.jung.sourceforge.net. Accessed 26 Jan 2022
16. JGraphX: <https://github.com/jgraph/jgraphx>. Accessed 26 Jan 2022
17. Prefuse: <https://github.com/prefuse/Prefuse>. Accessed 26 Jan 2022
18. Heer J, Card SK, Landay JA (2005) Prefuse: a toolkit for interactive information visualization. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 421–430
19. Eaton DA (2020) Toytree: a minimalist tree visualization and manipulation library for python. *Methods Ecol Evol* 11(1):187–191
20. Arts JC, Corsten FD, Hu YT, Papandroudou S, Warzynska DA, Burch M, Dynagraph: Visualizing dynamic graph data
21. Brandes U, Eiglsperger M, Herman I, Himsolt M, Marshall MS (2001) Graphml progress report structural layer proposal. In: International symposium on graph drawing. Springer, pp 501–512 (2001)
22. Taba ST, Brennan PC, Lewis S (2019) Dynamics of breast imaging research: a global scoping review and sino-australian comparison case study. *PLoS One* 14(1):e0210,256
23. Appel KI, Haken W (1989) Every planar map is four colorable. *Am Math Soc* 98 (1989)
24. Tarjan R (1972) Depth-first search and linear graph algorithms. *SIAM J Comput* 1(2):146–160
25. Fleury M (1883) Deux problemes de geometrie de situation. *Journal de Mathematiques Elementaires* 2(2):257–261
26. Pevzner PA, Tang H, Waterman MS (2001) An eulerian path approach to dna fragment assembly. *Proc Natl Acad Sci* 98(17):9748–9753
27. Dijkstra EW (1959) A note on two problems in connexion with graphs. *Numerische Mathematik* 1(1):269–271
28. Palmer E (1997) The hidden algorithm of ore’s theorem on hamiltonian cycles. *Comput Math Appl* 34(11):113–119
29. Hierholzer C, Wiener C (1873) Über die möglichkeit, einen linienzug ohne wiederholung und ohne unterbrechung zu umfahren. *Mathematische Annalen* 6(1):30–32
30. Paton K (1969) An algorithm for finding a fundamental set of cycles of a graph. *Commun ACM* 12(9):514–518
31. Tiernan JC (1970) An efficient search algorithm to find the elementary circuits of a graph. *Commun ACM* 13(12):722–726
32. Woodall D (1973) The binding number of a graph and its anderson number. *J Comb Theory, Ser B* 15(3):225–255
33. Aytac V, Berberler ZN (2017) Binding number and wheel related graphs. *Int J Found Comput Sci* 28(01):29–38
34. Cunningham WH (1990) Computing the binding number of a graph. *Discret Appl Math* 27(3):283–285
35. Borgatti SP (2005) Centrality and network flow. *Soc Netw* 27(1):55–71
36. Aytac V, Turaci T (2018) Closeness centrality in some splitting networks. *Comput Sci* 26(3):78

37. Barefoot CA, Entringer R, Swart H (1987) Vulnerability in graphs-a comparative survey. *J Combin Math Combin Comput* 1(38):13–22
38. Aytaç V (2005) Vulnerability in graphs: the neighbour-integrity of line graphs. *Int J Comput Math* 82(1):35–40

Web Service Credibility Evaluation Methods in Different Application Domains



Atef Shalan, Jaciél E. Reyes, Hayden Wimmer, Sarika Jain,
and Mohamed Hefny

Abstract Websites are growing in number and size at an explosive rate. This has created an enormous field of service competitors, which can cause a real challenge for information seekers and inexperienced users. As a result, these users require public information to help their decision about service selection. This aid comes in the form of clear and accessible credibility measures. Numerous research articles have studied the concept of Web service credibility evaluation to support user choices. This paper studies the existing methods and frameworks for evaluating web service credibility. We contrast these methods concerning their applications, underlying approach, evaluation factors, and primary research outcomes. Our study also provides a classification of these methods based on the underlying approach, application domain, purpose of the measure, and measurement automation. This study will set the stage for future research on automating web credibility measurements and dissemination methodology.

Keywords Webservice credibility · Credibility measurement · Automation · Service dissemination

A. Shalan (✉) · J. E. Reyes · H. Wimmer
Georgia Southern University, Statesboro, GA, USA
e-mail: amohamed@georgiasouthern.edu

J. E. Reyes
e-mail: ab40372@georgiasouthern.edu

H. Wimmer
e-mail: hwimmer@georgiasouthern.edu

S. Jain
National Institute of Technology Kurukshetra, Kurukshetra, India
e-mail: jasarika@nitkk.ac.in

M. Hefny
Queen's University, Kingston, ON, Canada
e-mail: hefny@cs.queensu.ca

1 Introduction

The Internet has changed how people interact with each other, shop, and even access their entertainment. As a result, the channels of information are now traveling freely on a worldwide scale. This leads to information becoming the most valuable resource on the Internet. The question then arises, “How can we be sure that what we are looking at can be trusted?”. False information on the Internet is nothing new. Regulating information on the Internet is an important task. Large and popular online e-commerce sites like Amazon and eBay are used by many to purchase various appliances, collectibles, and other miscellaneous items. Relative to the customer, vendors are not known and are sometimes anonymous. Often, customers may receive incorrect or faulty items and then be left at the mercy of the seller’s and the vendor’s sales and return policy that does not always resolve the customer’s problem. Some transparency and less ambiguity about these vendors would help billions of customers make better decisions about their purchases.

While the previous example is an all-too-common occurrence, false online medical information can lead to even more critical damages. As stated by Freeman and Spyridakis [1]: “people do not always carefully evaluate all the information they encounter”. This can be incredibly dangerous, especially when people take the information at face value. Credibility is defined as “the quality or power of inspiring belief” [2]. In the web technology domain, we define web credibility as the degree of public trust stimulated by user interaction with web services. This definition provides an abstraction of web credibility measures considered by most of the research work surveyed in this paper.

The work proposed in this paper discusses the different techniques utilized to better vet and ensure web credibility. We study the existing methods, application environments, underlying parameters, and research goals, among many other factors, of each research work. We then classify the current techniques and illustrate different models in different application domains. The main categories at the top of our taxonomy are design-based web, content-based, and user-based service credibility measures. Each type considers various factors that play a role in how information from a service measure is extracted, evaluated, and then provided to the end user.

While we briefly introduced the top-level categories of this survey in [3], this work contributes to the literature by detailing the state-of-the-art mechanisms of web service credibility measures. This facilitates information to Internet users about how to make decisions about their services and how to interpret web information on different websites. The work also provides a taxonomy of the credibility measures of web services and their main classes, subclasses, factors, domains of application, level of automation, and several other aspects. This work also helps set the stage for future research that facilitates web credibility measurement and automates disseminating these measures to the Internet user community.

The remaining part of this paper is organized in three sections as follows. First, the article will provide some motivational background and an overview of web credibility approaches in Sect. 2. Section 3 explains our classification criteria and compares them

with previous and similar work. It then describes the class levels of our classification and the subcategories in each class. It also provides details of the primary types of web credibility measurement methods and analyzes their underlying parameters and variants. Section 4 will conclude our paper and discuss future research.

2 Overview of Web Service Credibility

With the amount of information growing on the Internet, it has become increasingly important to receive factual data and good services. As the largest source of information and services, the web is full of redundant, mutated, fake, incorrect, and malicious content competing with the authentic information source. Inauthentic or low-quality services can cause harm to all levels of stakeholders and can also lead to real-world consequences. This section provides a high-level discussion and overview of web credibility. We then discuss the primary research directions in this domain.

2.1 Overview of Web Credibility

Since the term “web-credibility” was proposed by the Stanford Persuasive Technology Lab in 1998, numerous researchers have been investigating methods to quantify and facilitate its measures toward the average Internet user. Many researchers have focused their efforts to study the concept of web content evaluation and support user choice. Due to malicious and untrusted services on the web, tracking web credibility impairments is continuously in demand, and the need for service credibility quantification and dissemination methods is becoming a significant concern in our cyberworld.

Researchers have exerted several research approaches to assure credibility during the design and operation of web services [4–9]. These techniques focus on web service design as the main feature for establishing user trust and, consequently, web credibility. Examples of these design aspects include functionality [5], simplicity [6], interactivity [7], navigation aid [8], privacy policy [8], personalization capabilities [5], customer service [5], and accessibility [9]. Although these techniques target web credibility as the primary goal, they focus on the design quality of web services and assume that the credibility of these services is a direct product of design quality. The main drawback of these approaches is the undervaluation of the data contained within these sites. Thus, they ignore the impact of authenticity, integrity, authorship, and ownership of information, products, and services, among other values, on web credibility measures.

Such a limitation in design-based web credibility evaluation methods is addressed in content-based web credibility approaches [1, 4, 10–18]. These methods focus on the data content of web services via the discovery of some features [1, 16, 17] or by performing some verification on contained data objects in these web services [10,

11]. Content-based approaches are generally suitable for addressing many underlying data-related qualities of web credibility, such as accuracy, authenticity, authority, etc. They are also effective with automated models to assure such web service qualities. These techniques do not incorporate the level of user satisfaction in web credibility measures. More importantly, they don't consider allowing users to report violations discovered on the Internet.

User-based approaches [1, 15, 19–27] incorporate crowd inputs in the web credibility measures by capturing user opinion [24], ranking [21], recommendation [19, 25], review [27], or a questionnaire [1, 26]. These techniques evaluate web credibility as a matter of reputation [15, 22, 23] or popularity [20] and thus enable users to exchange their experience with web services regardless of their design and content. Incorporating user experiences in an evaluation is an important factor for the Internet user community to avoid the fear of the unknown. By integrating design-based and content-based web credibility measures, user-based approaches can result in high trust for reliable services while keeping users away from unreliable services. This triad of technique categories enables a natural selection of existing web services based on their quality, authenticity, and community satisfaction.

2.2 *Related Work*

This section mainly describes the previous web credibility efforts focused on surveying, classifying, or comparing existing web credibility approaches. In the following paragraphs, we highlight the work of Shah et al. [2], Sbaffi et al. [28], Olteanu et al. [29], and Kim et al. [30].

The work in [2] examines web credibility techniques at length. More specifically, they divided web credibility evaluations into two categories: those completed by computers and those conducted by users. Several methods were then tested based on the techniques applied. Credibility evaluations completed by users included a checklist, cognitive, contextual, motivation-centered, social, and heuristic approaches. Computer techniques were divided into scaffolding, visual aspects, credibility seals, credibility ratings, and digital signatures. The work found that utilizing a hybrid model of both user and computer-based techniques is essential in making accurate judgments of web credibility.

In [28], Sbaffi et al. seek to identify techniques to expand web credibility. However, while they aim to use a hybrid model, they emphasize a user-based model. They also use the Content Credibility Corpus (C3), which is considered the largest credibility database available for research, as part of their study. The work also cites Fogg's Prominence interpretation theory, which states that credibility occurs when users notice web content cues, perceive, and interpret them.

While [29] did not emphasize any specific web credibility techniques, it focused on a type of information frequently searched for by users. Web-based health information, also known as WHI, is increasingly important for everyday users. However,

it is essential to realize that users must thoroughly evaluate the sources for credibility and trustworthiness. This study suggests that website design features such as clear layout, interaction, and owner's authority positively affect service credibility. However, advertising wound up having a negative impact. Content features such as the author's authority, ease of use, and content were all determined to positively affect the trust or credibility of information.

This research [30] evaluates web credibility and cites examples of how content can be produced by fact-checking. Examples cited in this work include earthquake kidnapping and attack rumors. The paper notes that, since the information on the web grows at an explosive rate, it is not entirely feasible for users to be the sole judges of web credibility. The work recommends a generalized and fully automated credibility measurement framework for web services by using supervised machine learning algorithms.

3 Taxonomy of Web Credibility Measures

This section introduces our taxonomy by first describing the classification criteria and the taxonomy structure. We then overview the main classes in this taxonomy and more details about the class hierarchy and main subclasses.

3.1 *Classification Criteria*

The main goal of this taxonomy is to pinpoint the existing work on web credibility measures and the methodology utilized to automate this measurement process at the general Web service level. We contrast the current techniques concerning their contribution to an easy-to-use, generalized, automated, and publicly available web credibility measurement method. Thus, our criteria for classifying the existing literature are based on the following perspectives: source of measurement, application domain, automation phase, and measurement goal. In the following sub-subsections, we shed some light on the main categories of these perspectives and the existing research work based on these perspectives. Figure 1 shows our taxonomy structure as a hierarchy of three primary levels: Level 1 is the classification perspectives. Levels 2 and 3 are the classes and subclasses of web credibility measures for each perspective in Level 1.

Source of measurement. The source of a web credibility technique measurement indicates the original elements used to generate the web credibility measure. In the existing research, we found that many researchers rely on web service design to evaluate the credibility measure [4–9]. Some other researchers consider the data contents of web services or websites as the main factor of credibility measures [1, 4, 10–18]. With the continuous increase of Internet users, a research trend that

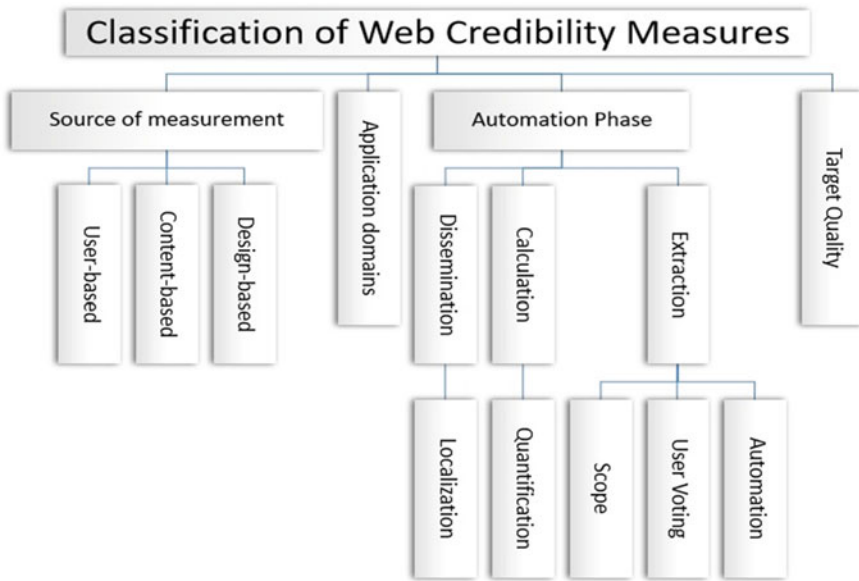


Fig. 1 Taxonomy structure of existing web credibility measurement techniques

emerged recently evaluates credibility based on user evaluation of these web services [1, 15, 19–27].

Application domains. Some research on credibility evaluation provides solutions at the application level for specific domains. Examples of these domain are health systems [4, 31], news media [10, 11], social media [14, 17], blogosphere [13, 20], e-commerce sites [27], service users [32], question and answering systems [33], and education [18, 26]. Specific domain models cannot be used in different areas, and they disregard general aspects of web content and system dynamics. Thus, many researchers provide generalized models for web pages [19, 24, 25, 34] and web services [12, 21–23] at the abstract level. Our taxonomy aims to discover the abstract factors of web services that impact web credibility.

Automation phase. Our classification of the existing web credibility measures based on the automation phase describes the contribution of work to the complete and high-level cycle of measurement project: extraction of measurements factors [11, 13, 23, 32], calculation of credibility measures [12, 14, 17], and dissemination of credibility measures [16, 35, 36] to the Internet user community. In each of these phases, we distinguish a few essential features.

The extraction phase of web credibility measures is either conducted manually [23, 24] or by using an automation mechanism [10, 11, 13]. In the extraction phase, we also distinguish whether the evaluation function is closed [4, 6, 13, 14] or open [10–12] over the domain of the source web service being evaluated. Closed evaluation means a website considers its service within itself. Open-source evaluation allows measuring the credibility of any web service on the web regardless of whether a

closed measurement exists or not, i.e., external evaluation. The third characteristic of the measurement function in this phase is whether the incorporated factors are limited [4, 6, 13, 14] or generalized [10, 11, 19]. Limited evaluation is focused on a specific set of factors or questions, while generalized inputs are not limited to any questionnaire or checklist. Generalized inputs enable reporting fraud activities, copyright violations, inauthentic information, authorship, ownership of products, etc.

The resultant measure type also characterizes the calculation phase, whether it is a quantity [11, 13, 23] or quality [6, 12, 15] measure. The third phase, credibility measure dissemination, is characterized by the target domain of dissemination: local [31, 35] or public [16, 37]. That is, whether the resultant measure is made available for only the local visitors of a website/service or it is made available through accessible dissemination channels to all Internet users.

Measurement goal. In our classification of web credibility measurements, we consider the research goals of such measurement in each work. These goals include authenticity [18], accuracy [34], usability [4, 9, 33, 38], security [17, 37], reliability [22, 39], design quality [1, 5], data integrity [10, 11, 40], and others [23, 41]. The following subsection provides details of our taxonomy and its main levels and classes.

3.2 *Taxonomy Classes of Web Credibility Measures*

Here we describe the classes and subclasses of existing web credibility techniques and illustrate the scope of these classes by providing details on their methods, factors, domains, goals, and other information.

First, Chart 1 divides the existing web credibility measures based on the source of measurement into three categories. Design-based approaches incorporate design aspects to evaluate web credibility measures. Content-based techniques focus on verifying the credibility of the content within the service. Finally, crowd-input-based involves the individual users and enables them to determine what is credible based on their overall input. The work share of traditional design-based approaches is the least among the three classes. Most recent approaches focus on data and users, and thus we find that most web credibility measures utilize user-based or content-based methods.

A pie chart of the number of articles available for each application domain is shown in Chart 2. Most research works target a general scope of the web page and web service applications. Environment-specific techniques follow, starting with health systems, news media, and social media.

As shown in Chart 3, most web credibility techniques attempt some sort of automation to obtain their measures. However, only 26 out of 36 techniques provide methods for extracting the calculation elements from the web service. Most of these techniques (21 out of 26) attempt to calculate a quantitative measure of web credibility. Surprisingly, only 2 out of the 36 articles addressing the automation method discuss how to disseminate these measures back to users. That is, the vast majority of web credibility work does not directly inform Internet users about their measures.

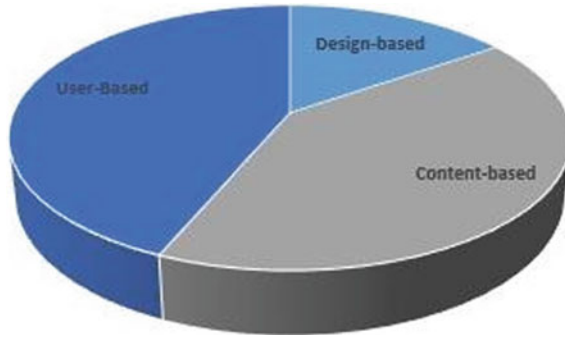


Chart 1 Number of articles in each source of measurement class

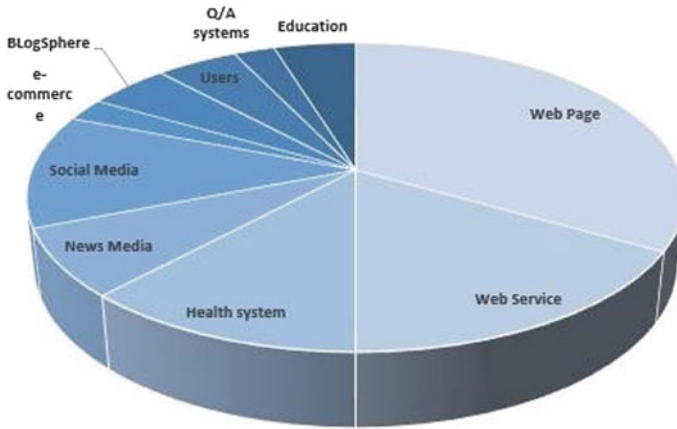


Chart 2 Number of articles in each application domain

Chart 3 A stacked-column chart showing the number of articles addressing the different automation phases

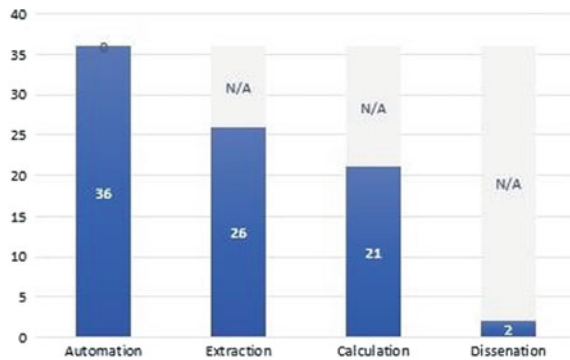




Chart 4 Number of articles addressing each research goal. Chart type: pie chart

Chart 4 shows that usability is the primary goal of the web credibility research work and data integrity comes second. The pie chart also shows the portion of research work focusing on other purposes such as authenticity, accuracy, security, and reliability.

Chart 5 details how the different web credibility techniques are focused on different types of services. It clearly illustrates the ratios of other methods in each application domain and the ratio of domains targeted by each different mechanism. For example, we can read from the chart that about half of the research work addressing health systems uses design-based approaches. Also, some minor user-based techniques are focused on health systems. We can similarly read the chart for each application domain. We also can read the chart from the left side by partitioning each class of techniques based on the target application domains. For example, we can see that design-based approaches only focus on health systems and general web page applications with an almost similar percentage.

Table 1 shows the main goals of web credibility measures in different applications. It also clusters these goals based on the source of measurement: service design, web content, and Internet user. Table 1 is wealthy with information, and it can help derive a lot of information. We can, for example, notice that the design-based web credibility measures usually aim for improving system usability and design quality. Content-based techniques focus primarily on data integrity, security, data quality, while user-based strategies concentrate on service quality and usability. Most contemporary applications (e.g., social media, e-commerce, new media) are targeted by content-based and user-based web credibility measures.

The sunburst in Chart 6 displays further details about the current status of credibility measures research automation. The first two layers of the sunburst are similar to levels 2 and 3 of the hierarchical taxonomy structure in Fig. 1. Additionally, the sunburst clarifies the percentages of research work addressing every automation phase. Thus, Chart 6 highlights the number of techniques that consider manual

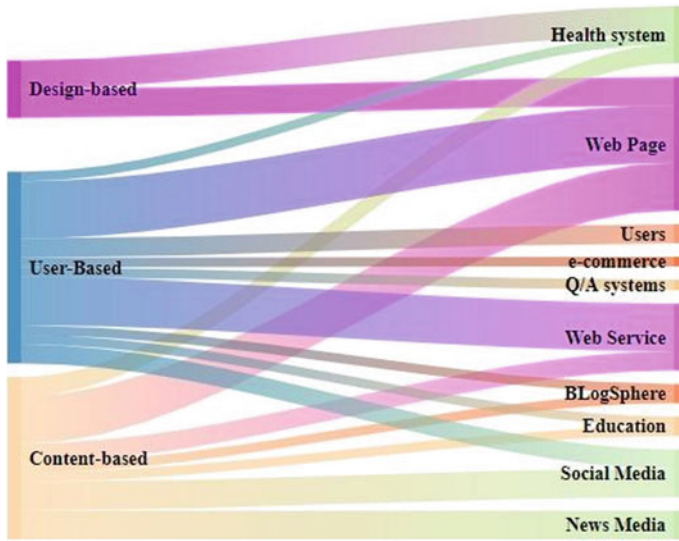


Chart 5 Mapping the ratios of web credibility methods with their targeted application domains. Left axis: source of measurement category, right axis: application domains

credibility evaluation (labeled no-automation). It also clarifies the small fraction of research work focusing on the web credibility measure dissemination to Internet users. The chart shows more details about the extraction methods in its third and fourth layers. About half of the work considers automating the extraction of web service parameters. Whether the extraction is automatic or manual, most of the parameter's extraction is closed to the web service stakeholders and not available to Internet users. That chart shows the ratio of generalized versus limited extraction parameters for each portion. For example, the chart shows that most web credibility measure parameters for manual extraction are closed to the service stakeholders and only use a limited and specific number of parameters.

4 Conclusion

Since the late 1990s, numerous research work has focused on quantification and facilitating web credibility measures for Internet users. This paper studies the existing techniques and frameworks for evaluating web service credibility. We also contrast these methods concerning their applications, underlying approach, evaluation factors, primary research outcomes, and other perspectives. Our study also provides a classification of these methods based on the underlying approach, application domain, source, purpose, and measurement automation.

Table 1 The source of measurement classifies the goals of web credibility measures in different applications domains. Table structure: rows are application domains, columns are the source of measurement, and cells are research goals

	Design based	Content based	User based
Web page	Usability [15, 16] Deign quality [5, 8, 42]	Data integrity [40] Security [43]	Accuracy [34] Usability [19, 35]
Web service		Reliability [12, 39]	Reliability [22] Service quality [23] Usability [38]
Online health information	Usability [4, 9] Design quality [7]	Design quality [1]	Service quality [31]
News media		Data integrity [11] Authenticity [44]	
Blogsphere		Usability [13]	Usability [20]
Q&A systems		Accuracy [33]	
Social media		Security [17] Accuracy [14, 38]	Opinion formation [41]
E-commerce			Usability [27]
Education		Authenticity [18]	Authenticity [26]

This article helps pinpoint the web credibility measures existing in the research literature within a larger map of web services, design aspects, and user needs. It also studies the methodology utilized to automate this measurement process at the general web service level. Our taxonomy contrasts these techniques concerning their contribution to an easy-to-use, generalized, automated, and publicly available web credibility measurement method. We survey above 40 articles and provide a multilevel classification hierarchy based on the source of measurement, application domain, automation phase, and measurement goal. Our taxonomy identifies several classes and subclasses and discusses these classes’ conceptual and implementation details.

While the existing techniques integrate numerous features to serve Internet users and help them select the appropriate services with a high confidence level, many limitations still exist. The main limitations of the existing methods of web credibility measures include the following. Most current web credibility measures are tightly coupled to specific web services. Very few mechanisms allow evaluating some scope of general services that are not set for evaluation by their stakeholders. Most of the

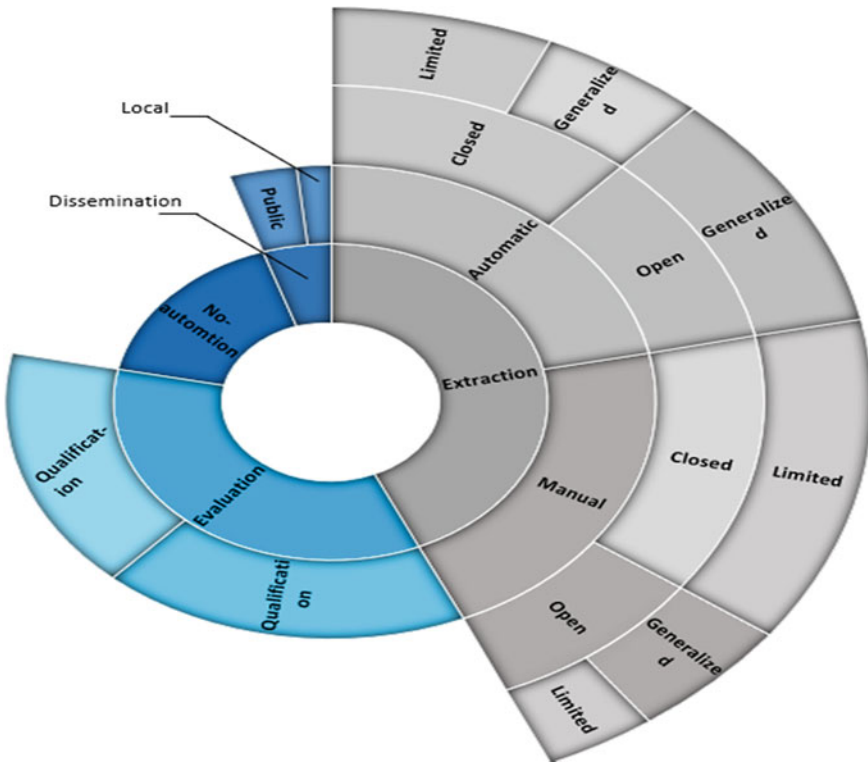


Chart 6 Automation phases and automation attributes of web credibility work

existing techniques utilize a limited number of parameters for assessing web credibility measures and do not allow open inputs from Internet users. Thus, their evaluations are restrictively limited to their stakeholder viewpoints. Few works consider crowdsourcing incorporation in web credibility measures restricted to the limitations mentioned above. However, most of these mechanisms are not automated and lack public dissemination to general Internet users through available and accessible access channels. Our future work focuses on designing a crowdsourcing framework for web credibility measures that allow evaluating any available service on the web, taking input from general Internet users, and then making results available through accessible channels.

References

1. Spyridakis KS, Freeman JH (2009) Effect of contact information on the credibility of online health information. *IEEE Trans Prof Commun* 52(2):152–166
2. Shah AA, Hamid S, Ravana SD, Ismail MA (2015) Web credibility assessment: affecting

- factors and assessment techniques. *Inf Res* 20(1):1–28
3. Reyes J, Shalan A, Shahriar H, Rahman MA, Jain S (2021) A classification of web service credibility measures. In: Proceedings of the IEEE computer society conference on computers, software and applications (COMPSAC 2020), pp 1399–1400, April 2021
 4. Sillencea E, Briggs P, Harris PR, Fishwick L (2007) How do patients evaluate and make use of online health information? *Soc Sci Med J* 64(9):1853–1862
 5. Fogg B, Soohoo C, Danielson D, Marable L, Stanford J, Tauber E (2003) How do users evaluate the credibility of web sites? In: Conference on designing for user experiences, San Francisco
 6. Thielsch MT, Blotenberg I, Jaron R (2013) User evaluation of websites: from first impression to the recommendation. *Interact Comput* 26(1):89–102
 7. Walther JB, Wang Z, Loh T (2004) The effect of top-level domains and advertisements on health website credibility. *J Med* 6(3):1–10
 8. Hong T (2005) The influence of structural and message features on website credibility. *J Am Soc Inf Sci Technol* 57(1):114–127
 9. Fisher J, Burstein F, Lynch K, Lazarenko K (2008) Usability + usefulness = trust: an exploratory study of Australian health. *Internet Res J* 18(5):477–498
 10. Jin Z, Cao J, Zhang Y, Zhou J, Tian Q (2017) Novel visual and statistical image features for microblogs news verification. *IEEE Trans Multimed* 19(3):598–608
 11. Chandrathlake R, Ranathunga L, Wijethunge S, Wijerathne P, Ishara D (2018) A semantic similarity measure based news posts validation on social media. In: 3rd International conference on information technology research (ICITR)
 12. Oskooei MA, Daud SM (2014) Quality of service (QoS) model for web service selection. In: International conference on computer, communications, and control technology, Langkawi, Malaysia
 13. Ulicny B, Kokar M, Matheus C (2010) Metrics for monitoring a social-political blogosphere—a Malaysian case study. *IEEE Internet Comput* 14(2):34–44
 14. BinSultan Al-Eidan RM, Al-Khalifa HS, Al-Salman AS (2010) Measuring the credibility of Arabic text content in Twitter. In: Fifth international conference on digital information management (ICDIM), Thunder Bay, ON, Canada
 15. Mahmood S, Ghani A, Daud A, Shamshirband S (2019) Reputation-based approach toward web content credibility analysis. *IEEE Access* 7:139957–139969
 16. Brewster Kahle BG (2021) Alexa rank. Alexa Internet, Inc. [Online]. <https://www.alexa.com/topsites>. Accessed 1 July 2021
 17. Nurrahmi H, Nurjanah D (2018) Indonesian Twitter cyberbullying detection using text classification and user credibility. In: International conference on information and communications technology (ICOIACT), Yogyakarta, Indonesia
 18. Rowley J, Johnson F, Sbaffi L (2015) Students’ trust judgments in online health information seeking. *SAGE Health Inform J* 316–327
 19. Deja D, Nielek R, Lin X, Wierzbicki A (2014) Hybrid algorithm for precise recommendation from almost infinite set of websites. In: IEEE/WIC/ACM International joint conferences on web intelligence (WI) and intelligent agent technologies (IAT). Warsaw, Poland
 20. Gonçalves MA, Almeida JM, dos Santos LGP, Laender AHF, Almeida V (2010) On popularity in the blogosphere. *IEEE Internet Comput* 14(3):42–49
 21. Guimarães S, Silva A, Meira W Jr, Pereira A (2010) CredibilityRank—a framework for the design and evaluation of rank-based credibility models for web applications. In: IEEE/IFIP International conference on embedded and ubiquitous computing, Hong Kong, China
 22. Wang M, Wang G, Zhang Y, Li Z (2019) A high-reliability multi-faceted reputation evaluation mechanism for online services. *IEEE Trans Serv Comput* 12(6):836–850
 23. Xu J, Zheng Z, Lyu MR (2016) Web service personalized quality of service prediction via reputation-based matrix factorization. *IEEE Trans Reliab* 65(1):28–37
 24. Zhuang Y, Xu Z, Tang Y (2015) A credit scoring model based on Bayesian network and mutual information. In: 12th Web information system and application conference (WISA), Jinan, China
 25. Kowalik G, Wierzbicki A, Borzyszkowski T, Jaworski W (2016) Credibility as signal: predicting evaluations of credibility by a signal-based model. In: IEEE/WIC/ACM International conference on web intelligence (WI), Omaha, NE, USA

26. Tresso A, Tartaglia E (2002) An automatic evaluation system for technical education at the university level. *IEEE Trans Educ* 45(3):268–272
27. I'm ET, Tung PH, Oh MS, Lee JY, Gim S (2021) A study on the extraction of customer satisfaction factors based on the customer satisfaction model using text review and preview. In: 21st ACIS International Winter conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD-Winter), Ho Chi Minh City, Vietnam
28. Sbaifi L, Rowley J (2017) Trust and credibility in web-based health information: a review and agenda for future research. *J Med Internet Res* 19(6):1–17
29. Olteanu A, Peshterliev S, Liu X, Aberer K (2013) Web credibility features exploration and credibility prediction. Lecture notes in computer science (LNCS) book series, vol 7814, pp 557–568
30. Kim Y (2014) Trust in health information websites: a systematic literature review on the antecedents of trust. *HHealth Inf J* 22:355–369
31. Eysenbach G (2008) Credibility of health information and digital media: new perspectives and implications for youth. Foundation series on digital media and learning. The MIT Press, pp 123–154
32. Kamkarhaghighi M, Chepurna I, Aghababaei S, Makrehchi M (2016) Discovering credible Twitter users in stock market domain. In: IEEE/WIC/ACM International conference on web intelligence (WI), Omaha, NE, USA
33. Shah AA, Ravana SD, Hamid S, Ismail MA (2020) Web pages credibility scores for improving accuracy of answers in web-based question answering systems. *IEEE Access* 8:141456–141471
34. Diana F, Bahry S, Masrom M, Masrek MN (2016) Website credibility and user engagement: a theoretical integration. In: 4th International conference on user science and engineering (i-USER), Melaka, Malaysia
35. Asad Ali Shah and Sri Devi Ravana (2014) Evaluating information credibility of digital content using a hybrid approach. *Int J Inf Syst Eng* 2:92–99
36. Alrubaian M, Al-Qurishi M, Alamri A, Al-Rakhami M, Hassan MM, Fortino G (2019) Credibility in online social networks: a survey. *IEEE Access* 7:2828–2855
37. Sullivan D (2007) Search engine land, 26-04-2007 [Online]. <https://searchengineland.com/what-is-google-pagerank-a-guide-for-searchers-webmasters-11068>. Accessed 1 July 2021
38. Shen L, Li Y (2011) Study about influence factor of credibility dissemination of internet mouth. In: 2011 International conference on management and service science, Wuhan, China
39. Gao H, Duan Y, Miao H, Yin Y (2017) An approach to data consistency checking for the dynamic replacement of service process. *IEEE Access* 5:11700–11711
40. Lin G, Wang D, Bie Y, Lei M (2014) MTBAC: a mutual trust-based access control model in Cloud computing. *China Commun* 11(4):154–162
41. Das R, Kamruzzaman J, Karmakar G (2019) Opinion formation in online social networks: exploiting predisposition, interaction, and credibility. *IEEE Trans Comput Soc Syst* 6(3):554–566
42. Kakol M, Nielek R, Wierzbicki A (2017) Understanding and predicting web content credibility using the Content Credibility Corpus. *Inf Process Manag* 53(5):1043–1061
43. Massarczyk E, Winzer P (2019) Influence of the perceived data security, credibility, trust and confidence on the usage frequency of internet services and the provision of security measures. In: International symposium on performance evaluation of computer and telecommunication systems (SPECTS), Berlin, Germany
44. Esteves D, Reddy AJ, Chawla P, Lehmann J (2018) Belittling the source: trustworthiness indicators to obfuscate fake news on the web. In: First workshop on Fact Extraction and VERification (FEVER), Brussels, Belgium
45. Hassan NY, Gomaa WH, Khoriba GA, Haggag MH (2018) Supervised learning approach for Twitter credibility detection. In: 13th International conference on computer engineering and systems (ICCES), Cairo, Egypt

Semantic Web Ontology for Botnet Classification



Omotola Adekanmbi, Hayden Wimmer, and Atef Shalan

Abstract Botnets have become a vital security problem on the Internet as such attacks lead to fraud, spam, identity theft, and information leakage. No intelligent classification knowledge graph of Botnets has been created for integration into AI applications. We address this by integrating concepts from cybersecurity into AI. Using an ontology model, we designed concept classes, individuals, and object properties of botnet to construct a knowledge graph of botnet containing their classification, features, and attack type. Our technique extracts cybersecurity knowledge from various textual sources to populate our knowledge graph on botnets and their attack type. To construct our knowledge base, we use Web Ontology Language (OWL 2 DL) for knowledge representation and Resource Description Framework (RDF) as a standard model for metadata representation. The system then reasons over the knowledge graph that combines a variety of collaborative agents to derive improved results. We describe a proof-of-concept framework for our approach as well as demonstrate its capabilities by testing it against different attack types and botnet identification features. Our knowledge base will help researchers analyze botnet samples and understand the infection procedures of botnets. It will also help in measuring the potential risk and possible damages of botnets.

Keywords Botnet · Ontology · Semantic web

1 Introduction

With the increasing growth of the internet and the widespread connected devices, IoT devices are often becoming the target of network attacks because they can be

O. Adekanmbi · H. Wimmer · A. Shalan (✉)
Georgia Southern University, Statesboro, GA 30415, USA
e-mail: amohamed@georgiasouthern.edu

O. Adekanmbi
e-mail: oa02259@georgiasouthern.edu

H. Wimmer
e-mail: hwimmer@georgiasouthern.edu

unsecured and used as a Bot. Botnets are one of the major cyber-security threats facing web technologies today. These botnets constitute a platform for launching numerous cyberattacks. Examples of these attacks include Distributed Denial of Service (DDoS), malware distribution, phishing, and click fraud for ransomware and Fortnite attacks [1]. Hospitals, companies, etc., can also become victims of botnet attacks because they often use a command and control (C&C) architecture to coordinate simultaneous encryptions of files. Bots look for vulnerable, unpatched, and unprotected devices to compromise to the C&C architecture.

The detection of botnets is an important research topic because these botnets stay hidden until their Botmaster is aware to execute an attack or task. Spammers have increasingly exploited these bots to send phishing and spam emails [2, 3]. To combat these attacks, our work proposed a method that constructs a comprehensive classification on the ontology of botnets which extracts knowledge from various textual sources to populate our knowledge graph on botnets and their attack type. We extract knowledge from details about the botnets via various web sources such as cert.org, and classify botnets based on relations and features. Ontologies introduce machine processing capability to knowledge structures [4] and are essential for semantic web engineering [5]. The ultimate goal of this effort is to develop an ontology of the Botnet domain, expressed using the Semantic Web Language (OWL, RDF, SPARQL, and Semantic Web rule Language (SWRL)) as explained by [6], that will enable the use of Pellet or Hermit reasoner to reason over the knowledge graph which combines a variety of collaborative agents to derive improved results and classification botnets based on features and attacks. The botnet ontology maintain detailed information repository about the behavior of botnet, classification, features, and attack type. It provides an effective way for researchers to analyze botnet samples, and further their understand infection procedures of botnets, as well as measure the potential damage extent of botnets.

The remaining parts of the paper will be organized as follows. In Sect. 2, we gave a background study about the semantic web, ontologies, botnets, classification, and attack patterns. In Sect. 3, we propose our Ontology and Methods. In Sect. 4, we performed experiments by running queries. Finally, Sect. 5 presents conclusions.

2 Background

2.1 Ontology

Ontologies are frameworks for representing a shares conceptualization of knowledge across a domain and are often defined as a shared conceptualization of a domain [5]. Ontologies help describe the concepts within a domain through representing classes and properties of taxonomies. The described concepts may include class properties and attributes, in addition to the relationships among these classes which can determine how they interact with each other. A class instance can be defined as

an individual instantiated example of a specific class. They play a vital role in the Semantic Web vision where ontologies provide the semantic annotation of websites in a meaningful way for machine interpretation [7]. Ontologies can describe relationships, classes, and their high interconnectedness. This makes an ontology ideal for modeling high-quality, linked, and coherent data.

2.2 *Semantic Web*

Berners-Lee, et al. [8] explained that Semantic Web gives meaningful structure to the content of Web pages in machine-readable formats for agent-based computing to carry out sophisticated tasks. There are three foundational Semantic Web technologies: RDF, SPARQL, and OWL.

- OWL is Semantic Web language family that represents complex relationships among things and groups of things [9]. They add semantics to the schema and rely heavily on the reasoner OWL gives you a much larger vocabulary to play with, making it easy to say anything you might want to say about your data model.
- RDF is a method for describing metadata by using a standard data model. RDF is used to build knowledge graphs, an abasis machine-readable data repository containing many structured and unstructured data. RDF statement states things about its subject by linking it to an object.
- SPARQL is the standard query language and the data access protocol for RDF databases. It can efficiently extract information hidden in non-uniform data and stored in various formats and sources by navigating relationships in RDF graph data through graph pattern matching.
- SWRL is a proposed language that can be used to express rules and logic when combined with OWL in the Semantic Web. However, SWRL expressions require an SWRL-enabled reasoner like Pellet, Hermit, etc. A reasoner is a software capable of inferring logical consequences from axioms. It helps determine whether the ontology is consistent, identifies subsumption relationships between classes, and more. This rule-based reasoner increases the inference capability of ontology-based models, and it achieves significant contributions when semantic queries are done.
- SQWRL stands for Semantic Query Web Rule Language and it is used for querying the ontology. SQWRL is based on SQL and facilitates client queries [10].

2.3 *Botnet*

A botnet is a set of infected computing devices under the control of an attacker (the bot herder or the Botmaster) (Fig. 1). There's no minimum size for a group of infected computers to be called a botnet, and an individual computer in a botnet is generally called "bots" or "zombies". The Botmaster can use the command and control channel

to disseminate malicious commands to the bot army via the C&C server (as shown in Fig. 2). The C&C channel enables the Botmaster to remotely control the action of a large number of bots to conduct various illicit activities and are usually distributed via Web downloads and email attachments.

Botnets are often classified according to their communication protocols;

- Internet Relay Chat (IRC) botnet—infected machines with malware that can be controlled remotely via an IRC channel. The IRC protocol mainly allows communication and data dissemination among users of large social networks. Examples of bots that use this communication protocol are DorkBot, Gamebot, RageBot, Phorpiex, etc.

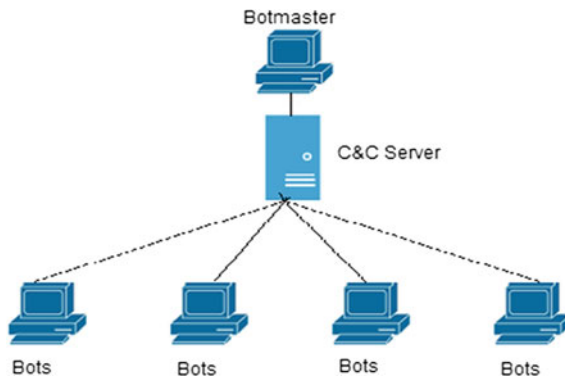


Fig. 1 Sample botnet diagram

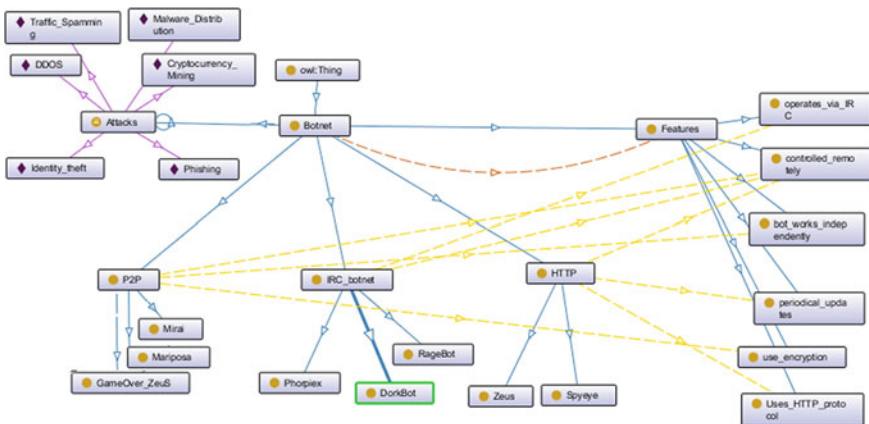


Fig. 2 A sample botnet ontology model showing classes, subclass, and relationships

- HTTP botnet—a web-based botnet that allows periodical dissemination of commands through the HTTP protocol. The herder of an HTTP botnet masks the malicious activities as regular HTTP traffic. Examples are Zeus and Spyeye.
- P2P (peer-to-peer)—a new generation of botnets in which different bots can share information and commands by coming to a direct contact with the C&C server. Relative to IRC and HTTP botnets, P2P botnets are harder to locate, monitor, and implement since they do not rely on one centralized server. Examples are Mirai, Mariposa, and GameoverZeus Botnets.

Examples of illegal activities that can exploit Botnets include DDoS, Identity Theft, and Traffic spamming.

- DDOS—An attack to crash a target server in which numerous bots send connection requests to the server to overwhelm it and prevent the operation of other legitimate requests to the server.
- Phishing: is a form of Social engineering where botnets are used to distribute malware via phishing emails.
- Identity Theft—An attack of stealing a victim’s identity by using botnets, e.g., identity theft botnets are keyloggers that can record the user password and send it to the bot herders during the login user operation.
- Traffic Spamming—an attack that allows actively injecting malicious code into the HTTP traffic or passively gathering user sensitive information.

3 Methods

We developed an ontology for the botnet classification based on RDF, OWL, SPARQL, and Python Framework. Using Protégé, we created an expressive botnet ontology by extracting important data about botnets from internet sources such as wikis, blogs, newspapers, magazines, social networking sites, and video-sharing sites. Examples of the extracted data include the following:

- Botnet indicators and detection methods,
- Botnet attributes and characteristics,
- Software vulnerabilities and security loopholes, and
- Attack tactics.

The ontology captures semantic information from threat reports into a shared repository structure that facilitates collecting, aggregating, and analyzing the captured data. However, a machine cannot infer this knowledge from the text alone. Our cognitive approach addresses this issue by integrating a standard reasoning technique that will detect inconsistencies during data sharing and infer new information from existing information. Modeling after the Pizza Ontology example, we created classes, individuals, and characteristics in the ontology.

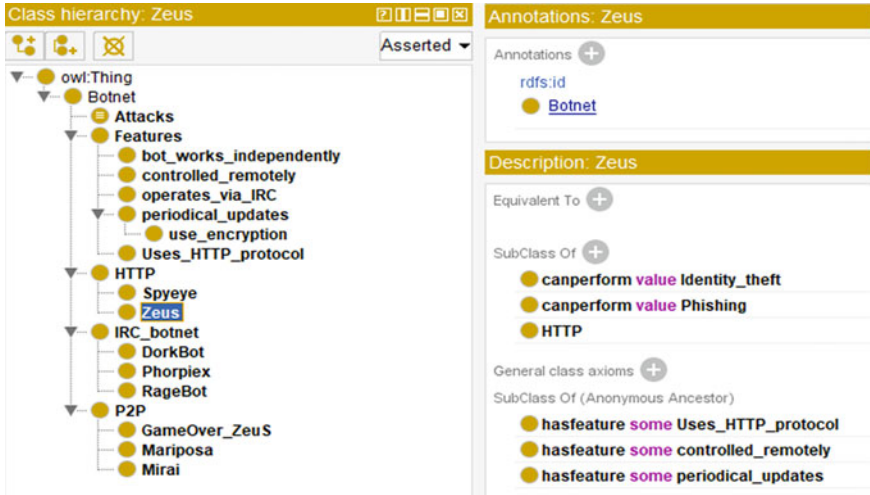


Fig. 3 Botnet class

3.1 Design of Botnet Class

As shown in Fig. 2, the botnet class defines the concepts about botnet categories, botnet features, and attack type. Botnets are often classified according to their communication protocols which are IRC, HTTP, and P2P. Each class represents a specific concept within the model. A class can have a subclass, e.g., Zeus is a subclass of HTTP Botnet. The attack class is the negative effect when a botnet affects a computer. It comprises several instances; DDOS, phishing, identity theft, malware distribution, traffic spamming, and cryptocurrency mining. The feature class provides valuable details about the features common to all types of botnets, as shown in Fig. 3.

Figure 4 shows the ontology of the Zeus Class, which is a subclass of the HTTP Botnet, and it inherits features from various classes as shown in Fig. 5, which gives a complete description of the Zeus Botnet showing the relationship between canperform and hasfeature.

3.2 Design of Object Property (Behavior Class)

Here, we define the main relationship representation properties between the concepts. The object properties represent the semantics of the sentences and connect the instances in the botnet classification. Property characteristics are defined by the domain and thereby enforce restrictions on classes and relationships. Figure 5 shows the class botnet with two properties hasfeature and canperform.

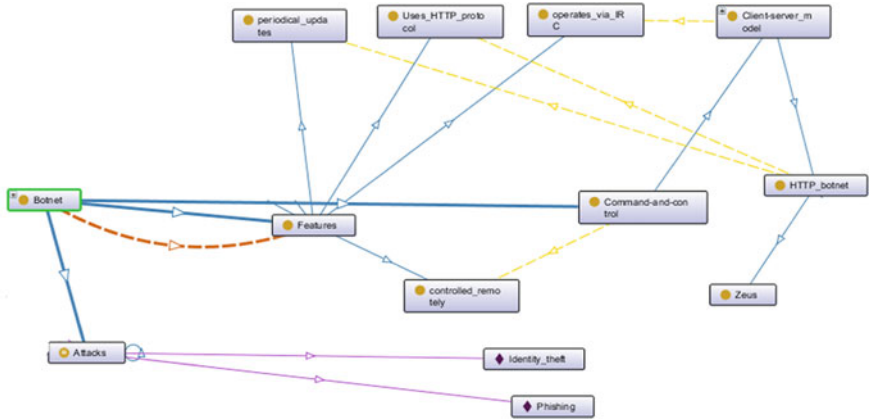


Fig. 4 Zeus class

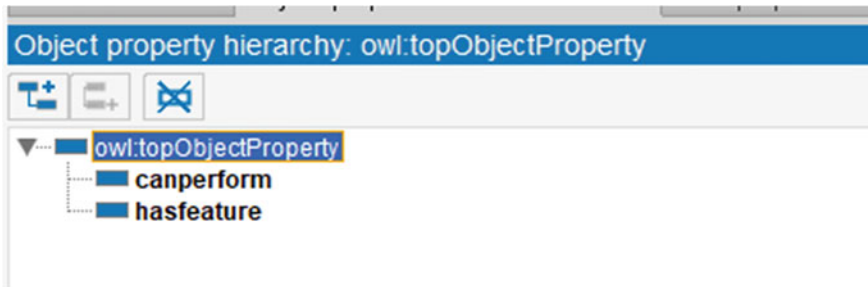


Fig. 5 Botnet object properties

- canperform: describes the types of attacks botnets can perform.
- Domain: Botnet
- Range: Attacks
- hasfeatures: connects botnets with their features.
- Domain: Botnet
- Range: Feature

3.3 Ontology Description of Botnet Individual

In the ontology model, the attack a botnet performs is individuals/instances belonging to the class Attack, including DDOS, phishing, identity theft, malware distribution, traffic spamming, and cryptocurrency. Using the canperform object property, we can create a relationship between botnet classification and the attack pattern. Figure 6 shows the individual phishing and how it is used.

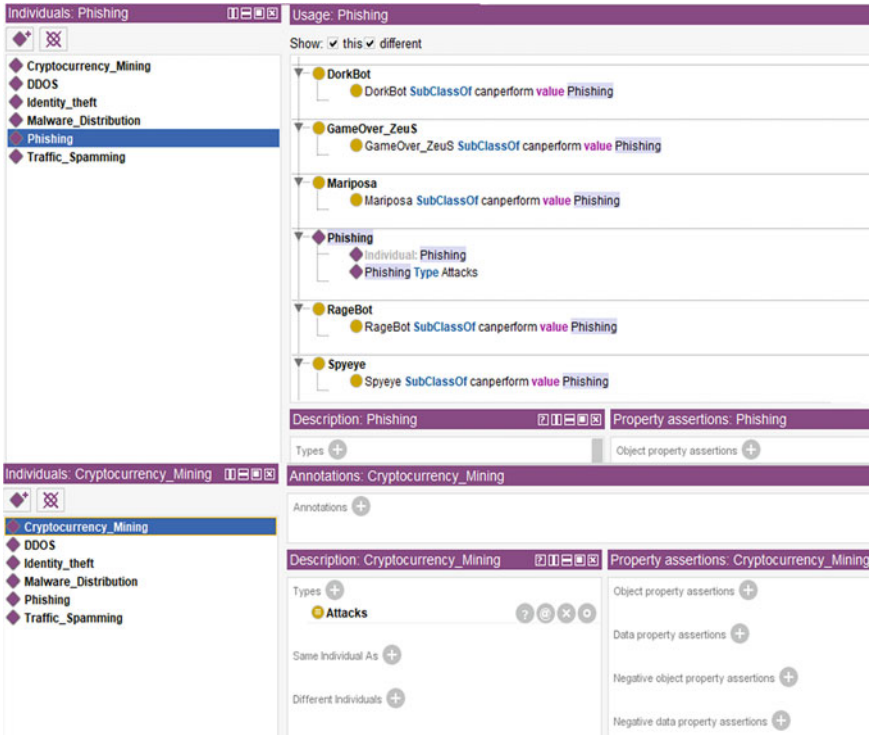


Fig. 6 Botnet individuals

As shown in Fig. 2, we created the ontology to show the classes and subclass (indicated using the blue lines) and individuals (denoted using the purple lines) as the attack type and gave them object properties (displayed using the yellow dotted lines).

4 Experiment and Results

This section evaluates botnets by using some SPARQL queries and running them on the ontology to answer a number of questions;

Retrieving botnet name related to an attack type: Botnet is queried to extract the names of the different botnets that can perform a phishing attack using the object property canperform (Fig. 7).

In Fig. 8, in lines 1–3, we defined the prefix at the top of the query so that we can abbreviate URIs [bonet: <http://www.semanticweb.org/ontologies/botnets#> (contains information on the botnet); RDFS: <http://www.w3.org/2000/01/rdf-schema#> (provides a mechanism for describing groups of related resources (RDF), and

```

1 PREFIX botnet: <http://www.semanticweb.org/ontologies/botnets#>
2 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3 PREFIX owl: <http://www.w3.org/2002/07/owl#>
4
5 SELECT ?Botnets
6 WHERE {
7   ?Botnets rdfs:subClassOf ?restriction .
8   ?restriction owl:onProperty botnet:canperform .
9   ?restriction owl:hasValue ?botnet:Phishing .
10
11 }

```

Zeus
 Spyeeye
 Mariposa
 RageBot
 GameOver_ZeuS
 DorkBot

Fig. 7 All botnet that can perform a phishing attack

the relation between botnet); OWL: <http://www.w3.org/2002/07/owl#> (for creating more detailed descriptions of resources.))] to make the query more readable. Lines 5–9 use the SELECT statement to query patterns Botnet WHERE rdfs is a subclass of a botnet with an OWL Object property canperform and has value phishing. In plain language, it searches for botnets that can perform phishing.

Retrieving information on botnet and their features: We extracted the names of the different botnets and their features using the Object property hasfeature.

Using SQWRL to list all the individuals within the ontology, as shown in Fig. 9, we use the Pellet or Hermit reasoner that supports SWRL to create rule S2, which queries the ontology OWL: Thing and selects (?i), the Individual or the Attack type, and displays them in ascending order.

4.1 Application of Ontology to Python

As Shown in Fig. 10, we applied our ontology to Python and Lines 1–3 import the RDFLIB, a pure Python package that provides the main types of RDF and their interfaces. The Python package provides a plugin interface for parsers, stores, and

```

1 PREFIX botnet: <http://www.semanticweb.org/ontologies/botnets#>
2 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3 PREFIX owl: <http://www.w3.org/2002/07/owl#>
4
5 SELECT ?Botnets ?Feature
6 WHERE {
7   ?Botnets rdfs:id botnet:Botnet .
8   ?restriction owl:onProperty botnet:hasfeature .
9   ?restriction owl:someValuesFrom ?Feature .
10  }ORDER BY ASC(?Feature)

```

Botnets	Feature
Spyeeye	Uses_HTTP_protocol
Mirai	Uses_HTTP_protocol
GameOver_ZeuS	Uses_HTTP_protocol
Zeus	Uses_HTTP_protocol
Mariposa	Uses_HTTP_protocol
Spyeeye	bot_works_independently
Mirai	bot_works_independently
GameOver_ZeuS	bot_works_independently
Zeus	bot_works_independently
Mariposa	bot_works_independently
Spyeeye	controlled_remotely

Fig. 8 List of botnets and their features

Active ontology	Entities	Individuals by class	Individual Hierarchy Tab	SWRLTab	OntoGraf	SPARQL Query	SQWRLTab
Name	Query						
S1	thor:scn/Attacks(?c) -> sqwr:select(?c)^ sqwr:orderBy(?c)						
S2	owl:Thing(?i) -> sqwr:select(?i)^ sqwr:orderBy(?i)						
SQWRL Queries OWL 2 RL S1 S2							
:Cryptocurrency_Mining :DDOS :Identity_theft :Malware_Distribution :Phishing :Traffic_Spamming							

Fig. 9 SQWRL query

serializers facilitated for other packages to implement and plug them into the Idlib package. RdfLib is the primary interface for working with RDF in Idlib Graph. Line 6 uses the parse command to read in the Botnet, OWL file. Lines 8–20 is the SPARQL query that extracts the list of botnets in which object property: canperform the attack

```
1 import rdflib
2
3 g = rdflib.Graph()
4
5 # ... add some triples to g somehow ...
6 g.parse("Botnet.owl")
7
8 qres = g.query(
9     """PREFIX botnet: <http://www.semanticweb.org/ontologies/botnets#>
10     PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
11     PREFIX owl: <http://www.w3.org/2002/07/owl#>
12
13     SELECT ?Botnets
14     WHERE {
15         ?Botnets rdfs:subClassOf ?restriction .
16         ?restriction owl:onProperty botnet:canperform .
17         ?restriction owl:hasValue botnet:Phishing .
18
19     }
20     """
21 )
22 for row in qres:
23     d = row['Botnets'].toPython()
24     print(d.split('#')[-1])
```

Run: Botnet x

- ↑ "C:\Users\HP\Desktop\M.SC IT CLASS\Data Science Methods\pythonProject\ve
- Mariposa
- ↓ DorkBot
- Zeus
- Spyeeye
- GameOver_ZeuS
- ↓ RageBot

Fig. 10 Botnet ontology in python

type: Phishing. Lines 22–24 print out the result of the query which matches with that of Fig. 7.

5 Conclusion

In this paper, we design botnet’s concept classes, individuals, and object properties and propose the methods for constructing the knowledge graph of botnet, classification, features, and attack type. Our technique extracts knowledge from various textual sources to populate our knowledge graph on botnets and their attack type. The


ultimate goal of this effort is to develop an ontology of the botnet domain, expressed using the Semantic Web Language (OWL, RDF, SPARQL, and SWRL) as explained by Tim Burners-Lee, that will enable the use of Pellet or Hermit reasoner to reason over the knowledge graph which combines a variety of collaborative agents to derive improved results and classification botnets based on features and attacks. The botnet ontology stores detailed behavior knowledge about botnet, classification, features, and attack type.

References

1. Chowdhury S et al (2017) Botnet detection using graph-based feature clustering. *J Big Data* 4(1):14
2. Levy E (2003) The making of a spam zombie army. Dissecting the Sobig worms. *IEEE Secur Priv* 1(4):58–59
3. Alparslan E, Karahoca A, Karahoca D (2012) BotNet detection: enhancing analysis by using data mining techniques. In: *Advances in data mining knowledge discovery and applications*, p 349
4. Jain S, Meyer V (2018) Evaluation and refinement of emergency situation ontology. *Int J Inf Educ Technol* 8(10):713–719
5. Patel A, Jain S, Shandilya SK (2018) Data of semantic web as unit of knowledge. *J Web Eng*
6. Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. *Sci Am* 284(5):34–43
7. Horrocks I (2008) Ontologies and the semantic web. *Commun ACM* 51(12):58–67
8. Berners-Lee T, Hendler J, Lassila O (2001) The semantic web—a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities (in English). In: *Scientific American*, p 34
9. McGuinness DL, Van Harmelen F (2004) OWL web ontology language overview. In: *W3C recommendation*, vol 10, no 10, p 2004
10. Roda F, Musulin E (2014) An ontology-based framework to support multivariate qualitative data analysis. In: *Computer aided chemical engineering*, vol 33. Elsevier, pp 1891–1896

Design and Performance Evaluation of a Multi-patient Health Monitoring System



Samson Olasunkanmi Adigun, Ayodeji Olalekan Salau ,
and Fatima Chiamaka Ujunwa

Abstract In today's society, we witness health conditions like lung failure, heart-related disorders, and cardiac failure spreading at an alarming rate, resulting in unexpected fatalities. As a result, there is an urgent need for these diseases to receive prompt medical attention. However, as the world's population grows, especially in developing countries, access to timely health care becomes more difficult due to limited medical facilities and the time required as medical practitioners must take care of several patients in a single day. As a result, the need for a remote health monitoring system is justified because it eliminates this problem. Previous researches have primarily focused on employing temperature and heart rate sensors to monitor temperature and heart rate. This study examines the use of temperature, heart rate, and carbon monoxide sensors to monitor patient health and detect harmful gases. To assess the disparities between the developed system and the standard equipment, we conducted a performance evaluation by comparing the values produced from the standard device found in a hospital to the values derived from the developed system. A statistical (t-test) was carried out to investigate if there was a significant difference between the standard measurement and the developed system measurement at a 5% significant level (alpha level). The validation results obtained show that the p-values of 0.88 and 0.28 were greater than 5% level of significance which indicates that there was no significant difference between the standard and developed system measurement at a 5% significant level.

Keywords Health monitoring · Temperature · Heart rate · Carbon monoxide · Gas detection

S. O. Adigun · A. O. Salau (✉) · F. C. Ujunwa
Department of Electrical/Electronics and Computer Engineering, Afe Babalola University,
Ado-Ekiti, Nigeria
e-mail: ayodejisalau98@gmail.com; ayodejisalau@abuad.edu.ng

S. O. Adigun
e-mail: adigunso@abuad.edu.ng

1 Introduction

The availability of medical services in rural areas is currently limited. As a result, most natives do not have access to these facilities or cannot afford to use them resulting in a complete disregard for any minor health issues that manifest in the early stages as variations in vital parameters such as heartbeat rate, body temperature, pulse, etc. [1]. Unnecessary squandering of earnings is also observed as the health situation worsens and lives are put in jeopardy. Aside from the demographic issue, there is a lack of social distancing, which can aid the spread of COVID-19. Health parameters can be monitored remotely from time to time by medical practitioners at any location using health monitoring systems, eliminating all of the problems associated with traditional hospital health care. A health monitoring system is an excellent weapon to use in the fight against COVID-19 because it prevents all physical contact between patients and doctors, which can spread the virus further [2]. With the increasing quantity of dangerous gases emitted which poses a threat to humanity, identification is critical, as it can help to reduce health hazards. This, together with health monitoring, will result in a large reduction in illness. The Internet of Things (IoT) concept has evolved over the years, allowing information to be gathered from a variety of sensors. The internet of things allows for remote monitoring of patients' vital signs without requiring personal contact with patients or clinicians. In this paper, we compared the values derived from a standard device found in hospitals to the system being developed to determine the discrepancies of the developed system from the standard equipment.

The rest of this paper is structured as follows. Section 2 presents the related works, while Sect. 3 introduces the proposed method. The experimental results and discussion are presented in Sects. 4 and 5 concludes the paper.

2 Related Works

Authors in [1] developed an IoT-based health monitoring system in which body temperature, pulse rate, and room humidity were all monitored, with sensors and displayed on an LCD. The sensor data is then sent to a medical server via the Internet.

Authors in [3] developed a system that used wireless sensors to collect patient's body temperature, pulse, and heartbeat rate which sends the information using a Wi-Fi module to an IoT cloud platform. The system tracks the physiological parameters at every 15 s interval. The limitations of the system were accuracy, cost utility, and the number of sensors. The authors presented a framework, which used an e-health sensor shield associated with a cloud platform to retrieve data from the sensors. The parameters measured by the sensors were airflow, glucometer, and patient positioning. The parameters were sent to the cloud platform via a gateway. The data stored in the cloud platform was accessible by an authorized personnel for further analysis and investigation of the correlation between parameters.

Authors in [4] presented a portable physiological health monitoring system that could continuously monitor the patient's pulse, temperature, and other specific room parameters. The monitoring and control system used Wi-Fi module-based remote communication to monitor patient's condition and store their information on a server. The stored data could be accessed using an IoT network, and doctors could diagnose diseases from afar based on values obtained.

Authors in [5] presented a system that monitors patient's health. Sensors were connected to the Arduino Uno to collect information. The microcontroller was linked to an LCD monitor and Wi-Fi network in order to send data to a web server. An alert is delivered to the patient if there are any sudden changes in the patient's heart rate or body temperature. The authors presented a simple, wearable continuous blood pressure, heart rate, and body temperature monitoring system that interfaces with an Android smartphone via Bluetooth. The device's major components are the IR transmitter, receiver, LM35, MPXV5050GP, data acquisition unit, microcontroller (i.e., Arduino), and Bluetooth. Bluetooth is utilized because it outperforms Zig-bee. In [6], the authors developed a system architecture in which sensors communicate with the Intel Edison platform which consists of three sensors that measure three basic vital signs: body temperature, pulse rate, and blood pressure. The sensor data was transferred to an IBM Bluemix for storage in the cloud.

Authors in [7] presented a health monitoring system that is able to monitor patient's vitals through an android application. The values were sent to a cloud platform and then to a database, from which the medical practitioner is able to make a diagnosis based on the values. In the case where the value is above a certain threshold, the medical practitioner will be alerted through a smartphone.

The authors in [8] monitored patient's vital signs using IoT and a microcontroller. They considered only one point of view, which is the ECG signal. A Raspberry Pi was used to collect data from wearable sensors and the acquired data was sent to a MySQL database. The developers have used GSM notification to deliver ready messages to medical healthcare personnel in emergency cases.

From the literature reviewed, it can be deduced that quite a number of the previous works focused on remote monitoring of mainly, temperature and pulse rate using either IoT, Bluetooth, or Zig-bee. However, the proposed system observes pulse rate and temperature, and sends the information via IoT for remote monitoring by the health personnel. Furthermore, it also functions as an environmental monitoring system and a hazardous gas detection system that detects harmful gases (e.g., CO) in the human body and alerts the user on detection. It could also act as a smoke detector just in case of a fire outbreak.

2.1 Health Monitoring Systems with Hazardous Gas Detection

Health monitoring systems comprise of sensors that detect signals that correspond to physiological parameters such as heart rate, temperature, blood pressure, etc. The health monitoring system also allows for such parameters to be monitored over the Internet by medical practitioners' through the concept of the Internet of Things (IoT). This plays a significant role in assisted living. The patients, especially the elderly patients, are administered routine check-ups to ensure sound health. Health issues such as cardiac failure, lung failure, and heart-related diseases can quickly be curtailed by using the system. Also, it is an excellent tool in the fight against COVID-19 as it eliminates the need for contact between doctors and patients [9]. Since there are hazardous gases everywhere at different locations, which can cause significant damage to the health of individuals, there is a need to be able to detect the presence of these gases [10]. The presence of hazardous gases is a significant threat to the lives of asthmatic patients as it can result in asthma attack and even death of the patient. The addition of hazardous gas detection to the conventional health monitoring system increases patients' chances of health safety. The sensor used can also detect the presence of smoke in case of fire outbreaks.

2.1.1 Gas Detector

Carbon monoxide (CO) gas has been considered in this work because it's one of the most dangerous gases around which poses a threat to human health (i.e., causes Asthma). CO is a gas that threatens lives, in nearly every home and sends more than 20,000 people to the hospital each year as a result of CO poisoning. Carbon monoxide at high concentrations kills in less than 5 min. At low quantities, it will take a longer time to affect the human body. When it exceeds the EPA concentration of 9 ppm for more than 8 h it adversely affects the health. The Occupational Health and Safety Administration's standard for healthy workers in the United States is 50 ppm. At levels considerably below 50 ppm, respiratory capacity declines and the risk of a heart attack increases. In households, the EPA standard of 9 ppm appears to be at a reasonable limit. The i4 Series CO/Smoke Detector and Interface Module is the first connected combined solution for conventional fire and security systems.

3 Methodology

The proposed system comprises of three primary sensors: temperature sensor, pulse/heart rate sensor, and gas sensor, which are used to detect temperature, pulse, and hazardous gases (e.g., CO), respectively. The signals obtained from these sensors are processed using a microcontroller (Arduino). After data acquisition using the

sensors, the data is sent remotely via a Wi-Fi module to a cloud platform (Thingspeak) for remote viewing by the required medical personnel through a web application via a device. As the system detects the presence of a gas, it triggers an alert sound from the buzzer, and the LED of the system starts to blink. Figure 1 shows the block diagram of the proposed system.

To evaluate the system, data values from 20 individuals were required. The device will monitor the value of the health parameters (temperature and heart rate) from the Thingspeak cloud platform and other records. After this is achieved, the standard device used in hospitals is also used to measure the values of the health parameters. Figure 2 shows the PBC design of the system.

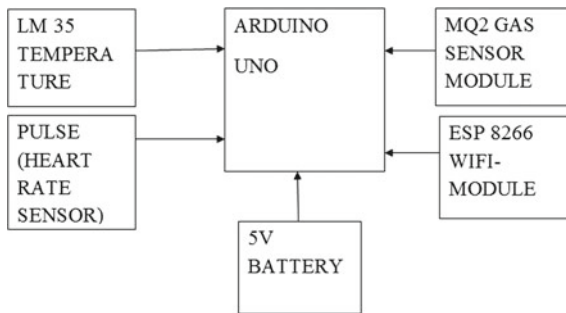


Fig. 1 Block diagram of the health monitoring system with hazardous gas detection

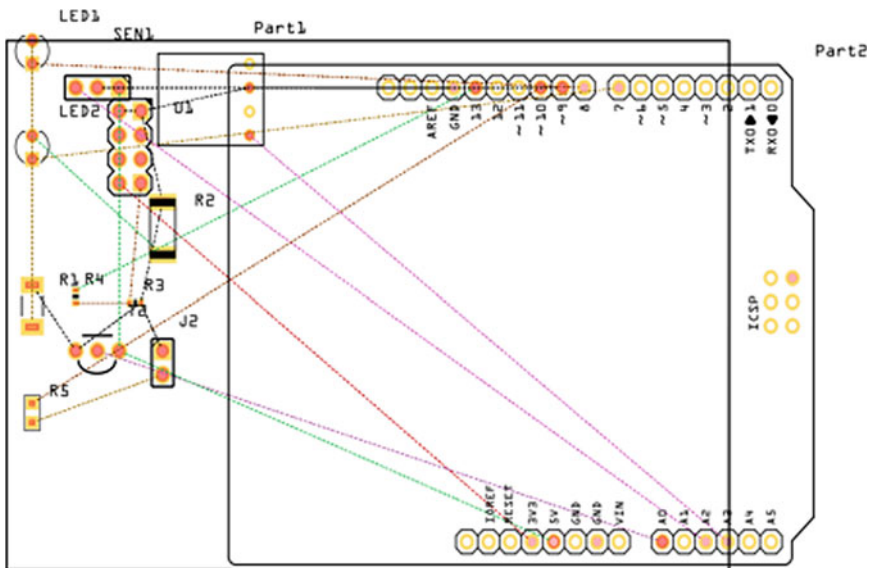


Fig. 2 PCB design of the proposed health monitoring system

3.1 LM-35 Temperature Sensors

The LM-35 output is directly proportional to its temperature and it is a high-precision device. Oxidation and other processes have no effect on an LM 35 as a result of its sealed sensor circuitry. A thermistor is usually used to measure the temperature more accurately. Therefore, we used an LM 35 temperature sensor comprising of three pins such as VCC (Pin 1), Out (Pin 2), and Ground (Pin 3). VCC and ground are linked to the VCC and Ground pins, respectively. The LM-35 is powered by a voltage ranging from 4 to 20 V, while a 5 V supply is utilized to power the Arduino board. The LM-35 is usually powered with voltages between 4 and 20v which is the same voltage range for the Arduino Uno. The signal pin is usually connected to the analog pin on the microcontroller board. The values are converted into Fahrenheit after being read.

3.2 Pulse Sensor

A pulse sensor makes use of noise cancellation and amplification circuitry and a simple heart rate sensor to get fast readings. It usually has to be connected to a 5v power supply and then placed on the fingertip or the earlobe. The pulse sensor module has three terminals—Ground, VCC, and Out. The signal pin of the pulse sensor module is connected to the analog pin A0 of the Arduino. Ground is connected to the common ground and the VCC is connected to 5 V DC output.

3.3 MQ-2 Gas Sensor

The output voltage of an MQ-2 Gas sensor is proportional to the gas per smoke concentration. A built-in potentiometer in the sensor allows one to be able to adjust the sensor sensitivity depending on how sensitive the gas to be detected is. The MQ-2 sensor has 4 pins. Table 1 depicts the pins of the MQ-2 sensor.

Table 1 Pins of the MQ-2 gas sensor

Pin	Connection to Arduino Uno
A0	Analog pins
D0	Digital pins
GND	GND
VCC	5 V

3.4 ESP8266 Wi-Fi Modem

The ESP8266 Wi-Fi Module acts as a gateway between the microcontroller and the cloud platform. It connects to the Wi-Fi network with the help of the built-in TCP/IP protocol stack. The module is available in two variants: ESP-12 and ESP-01. ESP-12 is composed of 16 pins used for interfacing while ESP-01 is composed only of 8 pins for interfacing. Table 2 shows the pin configuration of the ESP-12.

In particular, the ESP-01 model is used in this work. The ESP-01 model has the following pin configurations as presented in Table 3.

The common ground is usually connected to the Ground pin while the RESET and VCC are connected to the 3.3 VDC. Digital pins 12 and 9 are connected to the Tx and Rx pins of the module, respectively.

T-test was used to determine the extent of difference between the data obtained from the proposed system and the standard system using two measures, namely: temperature reading and heart rate reading. Analysis of variance (ANOVA) data analysis method was also used. ANOVA is a statistical analysis tool that separates observed aggregate variability within a data set into two components: systematic components and random factors.

Table 2 Pin configuration of ESP8266 ESP-12 Modem

Pin number	Pin name	Pin function
1	RESET	Active low external reset signal
2	ADC(TOUT)	ADC pin analog input
3	CH_PD	Active high chip enable
4	GPIO16	General purpose Input–Output (GPIO)
5	GPIO14	GPIO
6	GPIO12	GPIO
7	GPIO13	GPIO
8	VCC	Power supply
9	Ground	Ground
10	GPIO15	GPIO, should be connected to ground for booting from internal flash
11	GPIO1	GPIO, serial Tx1
12	GPIO0	GPIO, launch serial programming mode if low while reset or power ON
13	GPIO4	GPIO
14	GPIO5	GPIO
15	GPIO3	GPIO, serial Rx
16	GPIO1	GPIO, serial Tx

Table 3 Pin configuration of ESP8266 ESP-01 modem

Pin number	Pin name	Pin function
1	Ground	
2	GPIO1	GPIO, serial Tx1
3	GPIO2	GPIO
4	CH_PD	Active high chip enable
5	GPIO0	GPIO, launch serial programming mode if low while reset or power ON
6	RESET	Active low external reset signal
7	GPIO3	GPIO, serial Rx
8	VCC	Power supply

4 Results and Discussion

Table 4 shows the results of both the developed device and the standard device from which comparisons have been made to identify the discrepancies between the system using t-test and ANOVA data analysis methods. Figure 3 shows the error bars from the t-test data analysis carried out to compare the results of the developed device and standard device. For the t-test, the P-value is greater than 0.05 or 5%, indicating differences between the means, which is not statistically significant. In comparison, the P-value or t-test is less than 0.05 or 5% which suggests that the differences between some of the standards are statistically significant.

From the t-test analysis carried out, it is observed that the t-test is greater than 5% of the temperature values indicating that the difference between the values for temperature of the developed device and the standard device is not statistically significant, which means that the null hypothesis is to be accepted. It is also observed that the error bar of temperature values from the standard device and the developed system overlaps. The results obtained from the ANOVA experiment are presented in Table 5.

From the ANOVA test of the temperature reading in Table 6, we observe that the P-value is greater than 0.05, and as such, we can say that there are no significant differences between the means and the null hypothesis. Figure 4 shows the error margin from the analysis of the t-test data. “SS” means “the sum of squares due to the source”, “DF” means “the degree of freedom in the source”, “MS” means “the mean sum of squares due to the source”, “F” means “the F-statistic”, and “P” means “the P-value”.

Figure 4 shows the error margin from the t-test data analysis which was carried out to compare the developed device and standard device heart rate readings. It is observed that the error bar of temperature values from the standard device and the developed system overlaps, indicating that the difference between the devices are significantly small. From the t-test analysis carried out, it is observed that the t-test is greater than 5% for the heart rate values indicating that the differences between the values for temperature for the developed device and the standard device are not

Table 4 Performance evaluation of the system

S/N	Name of patient	Temperature reading from device (celcius)	Temperature reading from standard equipment (celcius)	Heart rate reading from device (bpm)	Heart rate reading from standard equipment (bpm)
1	I. Nnamah	33	36.3	67	77.6
2	O. Grace	34	36.8	86	76.2
3	J. Doumu	38	34.6	77	87.1
4	O. Oyebanji	31	33.9	85	74.9
5	M. Akintan	37	35.2	76	75.7
6	E. Dafe	40	37.3	57	66.6
7	O. Timothy	35	36.7	78	88.1
8	P. Ujunwa	39	35.9	68	78.3
9	Pr. Ujunwa	35	37.7	79	88.2
10	N. Nwosu	32	34.8	60	70.1
11	L. Ujunwa	39	35.8	110	99.8
12	C. Nwosu	38	36.9	110	99.6
13	E. Ifeanyi	34	37.2	89	90.5
14	N. Ikedi	40	36.5	67	76.9
15	C. Ujunv	38	35.1	85	94.8
16	O. Mbisike	34	34.6	84	74.5
17	D. Olaoye	35	37.7	71	80.4
18	F. Hassan	37	34.8	69	78
19	P. Ujunwa	34	36.9	69	77.3
20	J. Prince	37	37.2	70	79.1
	Mean	36	36.095	77.85	81.685
	St. Dev	2.65	1.1691089	14.057307	9.2601168
	Variance	7.052631579	1.366815789	197.6078947	85.74976316
	n			20	20
	T-test	88.44%		29.05%	

statistically significant, which means that the null hypothesis is to be accepted. It is also observed that the error bar of heart rate values from the standard device and the developed system overlaps. Table 7 shows the output of the GZAIR SA103 CO Detector in five different locations. The device was tested in five different locations based on their level of carbon monoxide (CO). When the level of CO is less than 50 ppm: no alarm is triggered, when the level of CO is greater than 50 ppm: weak alarm, and when the level of CO is greater than 200 ppm, a strong alarm is triggered.

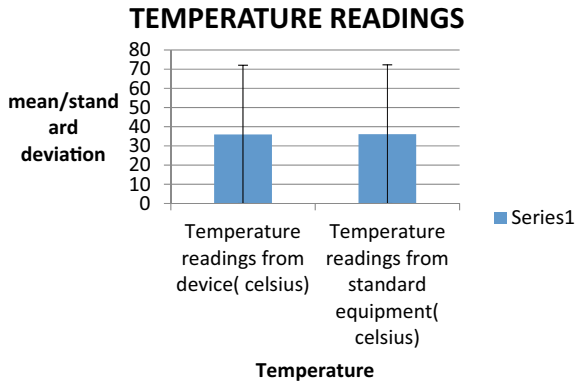


Fig. 3 The error analysis of the t-test data analysis

Table 5 Data derived from the single factor ANOVA is used to carry out comparisons between systems

ANOVA: single factor						
<i>Summary</i>						
Groups	Count	Sum	Average	Variance		
33	19	687	36.15789	6.918129		
36.3	19	685.6	36.08421	1.440292		
<i>ANOVA</i>						
Source of variation	SS	df	MS	F	P-value	F crit
Between groups	0.051579	1	0.051579	0.012342	0.912159	4.113165
Within groups	150.4516	36	4.179211			
Total	150.5032	37				

Table 6 Single factor ANOVA data analysis for the temperature values

Single factor ANOVA data analysis for the temperature values						
<i>Summary</i>						
Groups	Count	Sum	Average	Variance		
67	18	1420	78.88889	209.1634		
77.6	18	1477	82.05556	94.31791		
<i>ANOVA</i>						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	90.25	1	90.25	0.594765	0.445909	4.130017699
Within Groups	5159.182	34	151.7407			

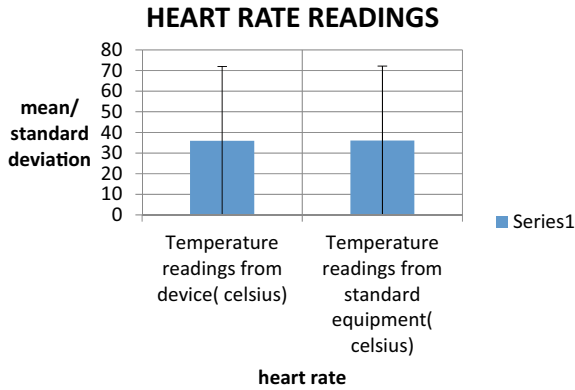


Fig. 4 Error evaluation of the t-test data analysis

Table 7 The outputs of the GZAIR SA103 CO Detector at the five (5) different locations

Location	Level of CO concentration (ppm)	Weak alarm	Strong alarm
Location 1 (normal atmosphere)	0.2	–	–
Location 2 (Home)	3.5	–	–
Location 3 (on the road)	14.4	–	–
Location 4 (beside car exhaust)	512.3	–	YES
Location 5 (Beside burning wood)	3245.1	–	YES

5 Conclusion and Recommendation

This paper presented the development and evaluation of a health monitoring system for hazardous gas detection. Performance evaluation was performed by comparing the values derived from the system to the standard device used in hospital. The t-test and ANOVA data analysis were used to analyze the results. The results show no statistically significant difference between the means for both the temperature and the heart rate sensors indicated by the t-test and P-test values. In the future, we recommend the following for a better system: Provision for analyzing the data being monitored should be made so that the system can automatically to diagnose any illness of the patient, possibly through the aid of machine learning. An application could be added to highlight healthy meals and exercise routines that the patients can use to improve their health.

The system can incorporate a GPS module to detect and send the patient’s location to the medical personnel for easy location in worse situations. There should be a

GSM module to instantly inform the medical personnel when the heart rate and the temperature exceed a certain threshold.

More sensors could be added to the system to monitor more physiological parameters of patients. Such as sensors to monitor the blood pressure and the oxygen levels of a patient can be incorporated into the system. The system can feature a backup battery to increase the reliability of the system.

References

1. Baabood AH, Baomar TA, Valsalan P (2020) IOT based health monitoring system. *J Crit Rev* 739–743
2. Pardeshi V, Sagar S, Murmurwar S, Hage P (2017) Health monitoring systems using IoT and raspberry Pi—a review. In: *International conference on innovative mechanisms for industry applications (ICIMIA 2017)*, pp 134–137
3. Ruman MR, Barua A, Rahman W (2020) IoT based emergency health monitoring system. In: *International conference on industry 4.0 technology (I4Tech)*, pp 159–162
4. Khan T, Chattopadhyay MK (2017) Smart health monitoring system. In: *IEEE international conference on information communication, instrumentation and control (ICICIC-2017)*, pp 1–6
5. Krishnan DR, Gupta SC, Choudhury T (2018) An IoT based patient health monitoring system. In: *International conference on advances in computing and communication engineering (ICACCE-2018)*, pp 1–7
6. Tan ET, Halim ZA (2018) Health care monitoring system and analytics based on internet of things framework. *IETE J Res* 1–8
7. Shetty HB, Ankitha S (2018) A review on health monitoring system using IOT. *Int J Eng Res Technol* 1–3
8. Dirja NI, Hardisal B, Rudi AC (2019) Heart rate monitoring and simulation with the Internet of Thing-based (IOT) alquran recitation. *J Publ Inf Eng Res* pp. 221–225
9. Yadessa AG, Salau AO (2021) Low cost sensor based hand washing solution for COVID-19 prevention. In: *International conference on innovation and intelligence for informatics, computing, and technologies (3ICT)*, pp 93–97. <https://doi.org/10.1109/3ICT53449.2021.9581821>
10. Pratiksha Y, Ashwin IW, Sheena S, Thakare A (2018) A review on patient monitoring system using IOT. *Int J Recent Innov Trends Comput Commun* 152–154

Discovering Novelty via Transfer Learning



Shafkat Islam and Bharat K. Bhargava

Abstract Modern-day intelligent systems, i.e., autonomous vehicles, smart manufacturing, etc., rely on distinct data-driven machine learning (ML) models for performing different safety-critical operations such as pedestrian detection, chemical plant control, among others. Since the performance of ML models depends on their generalizability capability, handling out-of-distribution data during the operational phase is of paramount importance for enhancing the adaptability of artificial intelligence (AI) systems. Hence, finding the root causes of novelty is critical for minimizing its impact on the performance of AI systems. However, detection of novelties is not trivial since each AI system operates in a specific environment or agent settings, whereas small changes cause the detection mechanism to adapt to different environmental constraints. For example, detecting novelties in identical intelligent navigation systems differs if the system is deployed in two different countries, though the operation of a navigation system is indistinguishable. In this paper, to reduce the dependability of novelty detection mechanisms on environment's or agent's attributes, we propose transfer learning-based novelty detection mechanisms for inter-domain applications. In this regard, we analyze the importance of feature transformation to enhance novelty detection systems' transferability. The proposed detection mechanisms aim at augmenting AI systems with rapid responsiveness to novel surroundings, thus, making AI systems responsible and trustworthy. We conduct multiple experiments on state-of-the-art neural networks, i.e., ResNet50, Mobile-Net, and benchmark datasets, i.e., MNIST.

Keywords Novelty detection · Transfer learning · Open-world AI

S. Islam (✉) · B. K. Bhargava
Purdue University, West Lafayette, USA
e-mail: islam59@purdue.edu

B. K. Bhargava
e-mail: bbshail@purdue.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
S. Jain et al. (eds.), *Semantic Intelligence*, Lecture Notes in Electrical Engineering 964,
https://doi.org/10.1007/978-981-19-7126-6_6

1 Introduction

With the emergence of hardware acceleration and high-speed connectivity, the proliferation of compute-intensive distributed AI systems has been observed in different application domains, including transportation, warfare, manufacturing, agriculture, and bio-engineering. In recent years, AI systems have been utilized for safety-critical operations such as autonomous navigation, nuclear plant control, autonomous plant irrigation, and automatic manufacturing process control. Hence, AI systems are required to have high accuracy, responsiveness, trustworthiness, and responsible.

Since the operational environments for the AI systems usually comprise typical open-world challenges, i.e., novel or out-of-distribution (OOD) instances, it is obvious that the AI systems are designed to tackle such unusual environmental behaviors. Responsiveness of the AI systems depends on its handling capacity of novel instances during the testing or operational phase. In this regard, the detection or identification of novel samples comes first, and the responsiveness is intuitively dependent on the functional capability of the AI system in detecting novelties within an environment.

However, delineating novelties in different environments, even within the environment for different agents, contexts, or interactions, is nontrivial as each entity in the environment consists of distinct attributes. Hence, novelty detection in AI systems is cumbersome and requires entity-specific knowledge, making such detection in AI systems both inefficient and impractical in extreme cases. To address this issue and develop a pseudo-generalizable novelty detection technique, we propose a transfer learning (TL) based novelty detection framework for inter-domain applications. In this regard, transfer learning can minimize both the entity dependency (through feature transformation) and a large amount of novelty data requirement for similar but distinct environments for inter-domain applications. The main contributions of this research are stated below.

- We formulate a TL-based supervised novelty detection mechanism for inter-domain AI applications. TL can reduce the resource (i.e., computation and training data) requirement for developing an efficient novelty detection scheme compared to the traditional supervised methods.
- We analyze the impact of feature transformation for minimizing the overhead of novelty detection system. We conduct experiments to analyze the effectiveness of TL for detecting novelties with limited number of novelty samples.

2 Related Works

The outlined novelty detection problem covers three distinct topics in machine learning, including supervised learning, transfer learning, and anomaly and novelty detection. In the following, we present the existing researches in this domain along with the research gap.

Multiple research works have considered novelty detection a classification task

in literature, whereas the objective has been to distinguish OOD instances from the well-known training or test distribution [7, 10]. Different supervised detection techniques, i.e., auto-encoder based semantic approach [11], generated novelty based training data augmentation [9], and deep neural network-based method [6] have been proposed for identifying novelties in AI systems. Novelty detection has been explored in deep reinforcement learning [4] environments as well. In contrast with novelty, anomalies are defined as unnatural behavior observed by AI systems during inference [2]. Suppose the variance of known data distribution (which can contain enough diverse data) is significantly high; in that case, an anomaly during inference can be considered a novelty, whereas the reverse notion is not always true [11]. This observation implies that novelty detection can be more generalizable than anomaly detection.

Machine learning-based network anomaly detection techniques, have been widely studied in literature [3, 5]. However, anomaly detection in other domains, i.e., in video [8] and sensor data [12] have recently received attention from researchers. Since novelty detection is somewhat similar to anomaly detection, such supervised learning techniques can be plug-able for novelty detection with appropriate modification depending on applications. However, these detection mechanisms depend on specific environment settings and are sensitive to minor changes in the predefined environment features. This attribute limits the widespread applicability of such detection mechanisms.

As AI systems usually operate in dynamic environments where the predefined environment/agent features can change abruptly, detecting novelties in such environments has enormous implications since the term 'novelty' for different entities (or agents) in the environment is not identical which can differ by context, and even on interactions. In such a dynamic situational environment utilizing environment-specific novelty detection mechanisms will fail due to the lack of its generalizability. In literature, very few works address the generalizability problem of anomaly detection techniques [1] whereas the existing novelty detection techniques for AI systems are lacking in such studies. This paper proposes a transfer learning-based generalizable novelty detection mechanism for addressing the adaptability issue of such detection techniques in various inter-domain AI applications.

3 Proposed TL-based Novelty Detection

We begin this section by introducing an example of a dilemma in defining novelty with minor changes in the environment.

High-level Problem Description: We consider an autonomous vehicle is equipped with an intelligent navigation system (INS). The INS relies on a vision-based object detection module for efficient and safe navigation. We assume that the object detection module uses a convolutional neural network (CNN) and is trained with a dataset with images of urban settings. Since the INS is enabled with a novelty detection feature, the training dataset of the CNN model consists of a 'novel' class that is

responsible for detecting unusual or unnatural instances during the inference phase. So the INS can make a prompt response in the presence of a novel instance. The training dataset only considers urban settings; the CNN module may recognize a ‘horse’ in the road as a novel instance. In contrast, if an identical CNN detection module is trained with a dataset of rural settings, the INS may not consider a ‘horse’ in the road as a novel. This dilemma holds different agents in the same environment or extreme cases based on contexts, and interaction among agents can also influence. One naive solution to this problem can be developing multiple classes of novelty (i.e., ‘urban novelty’ and ‘rural novelty’) during the training phase; however, this type of universalizing is impractical and inefficient.

Therefore, the existing literature focuses on developing environment or agent-specific novelty detection techniques that remain highly sensitive to environment or agent features, and slight changes in any of these may degrade the detection system’s performance. Hence, we can outline novelty into four distinct classes as described below.

Environmental Novelty: This type of novelty is related to the environmental features, and it differs from one environment to another. For instance, in the example mentioned above, the novelty differs due to the change in environment (i.e., urban and rural).

Agent-centric Novelty: The novelty for different agents differs according to the role of agents in the environment. For example, in warfare (environment), a novelty for soldiers (agents) who are fighting on the battlefield and novelty for the commander (agents) who is leading the troop from battle stations are different.

Contextual Novelty: Novelty also depends on the context of the environment and the agent. For instance, the novelty for the soldier (agent) when in the battle-station and on the battlefield are distinctly different.

Interaction-centric Novelty: Interaction among agents also plays a significant role in defining novelty, such as the novelty when two soldiers are interacting differs from novelty when a soldier and a commander are interacting.

Fig. 1 Overview of transfer learning for inter-domain novelty detection

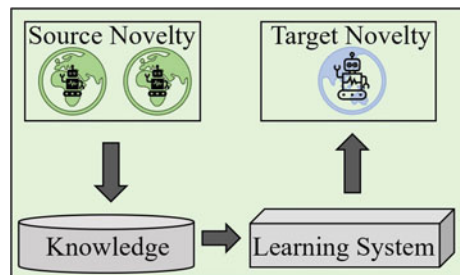


Fig. 2 TL-assisted Novelty detection deployment

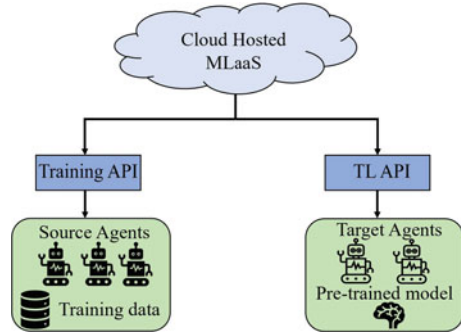


Figure 1 illustrates the proposed TL-based novelty detection method where the knowledge extracted from the source domain is utilized and fine-tuned by the target domain. In the following, we describe different novelty transfer methods (Fig. 2).

3.1 Transforming Novelty Features

Feature transformation is a popular TL technique in which the source classifier’s features are transformed consistently with the target classifier’s feature [13]. In this method, each common latent feature is transformed into new representations according to the data distribution of the target classifier. Such transformation aims to minimize the differences between the marginal and conditional differences and establish a bridge between the features of two distinct classifiers. Multiple difference matrices such as maximum mean discrepancy (MMD), Kullback-Leibler Divergence (KLD), Bregman Divergence (BD), Hilbert-Schmidt Independence Criterion (HSIC), and Jensen-Shannon Divergence (JSD), can be used while gauging the difference between two corresponding feature distributions based on the feature data type (i.e., image, text, or numerical). Feature transformation can be further divided into two distinct types as described below.

Feature Augmentation: In this method, distinct features are augmented by simple replication. For instance, in the case of single-source and single target, the feature set is augmented with three different feature sets, i.e., general, source, and target feature set.

Feature Reduction: In this method, source and target features are extracted, clustered, encoded, and mapped to transfer efficiently between two distinct classifiers. Depending on the application and data distribution, any or combination of reduction techniques can be utilized.

Feature Alignment: This method transforms the implicit features, i.e., statistical and spectral, for better feature representation in addition to the explicit feature transformation. This type of transformation technique helps in extracting latent characteristics of data distributions.

Since the definition of novelty varies even in similar applications, the feature transformation techniques can be adapted to transfer the known knowledge of source applications and adjust the detection technique suitable for target applications. This can reduce training overhead by minimizing training time and making the novelty detection mechanism more generalizable.

3.2 Transferring Novelty at Parametric Level

In this approach, the novelty detection model parameters in the source domain can be directly transferred to the target domain. The target domain classifier can fine-tune the parameters by freezing all the layers except the last few-layer neurons for adjusting the novelty instances in that domain. However, the level of adjustment or fine-tuning depends on the statistical distance between the source and target domain. This type of TL approach can be suitable for the example mentioned above, in which the INS can adjust its novelty detection module when it goes into the rural area from the urban by adjusting weights of last few layers.

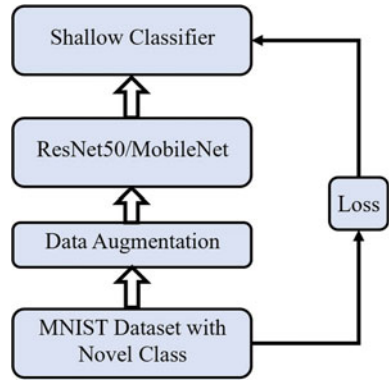
3.3 Transferring Novelty via Model Ensemble

In the model ensemble technique, multiple classifiers are stacked to make the output prediction. In this regard, the target novelty detection module can collect multiple source models from its domain and then perform ensemble techniques to find the appropriate combination of classifiers. The target module can construct a weighting or voting-based approach to select models among the available sources. This technique can achieve significant generalizability in the novelty detection model and fasten the detection training period significantly (Fig. 3).

3.4 Novelty Detection Deployment Architecture

To formalize the TL-assisted novelty detection deployment process, we assume a cloud-hosted machine learning as a service (MLaaS) provider. The cloud service provider hosts a novelty detection training process for the agents. Agents communicate with the service provider through a secure training application programming interface (API). On the contrary, if an agent (or target agent) wants to utilize pre-

Fig. 3 Transfer learning architecture for novelty detection



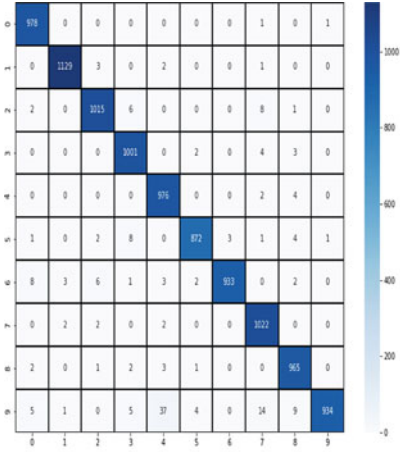
trained detection models by other agents (or source agents), it must communicate with the service provider through the TL API. The target agent can fine-tune the model with that pre-trained model either by itself or through the service provider. Figure 2 illustrates the proposed cloud assisted TL based novelty detection deployment architecture.

4 Experimental Analysis

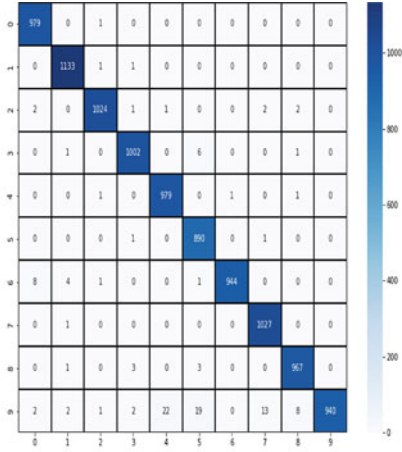
We conduct experiments to analyze the performance of transfer learning in detecting novelty classes where the target domain consists of fewer novelty samples compared to non-novelty samples. Figure 4 illustrates the experimental results where ImageNet is used as source domain and MNIST is used as target domain. We use two different neural network architecture, i.e., ResNet50 and MobileNet, in both the cases we only train the last two layers keeping rest of the feature extraction layers frizzed. We also limit the number of training epochs less than 5 for each of the cases. From the simulation results, we observe that ResNet50 outperforms MobileNet in detecting novelty class with an accuracy of 93%. With the increase in novelty samples, we observe that the detection accuracy for transfer models becomes higher for both cases. Figure 3). illustrates the novelty detection architecture used in this experiment.

5 Conclusions

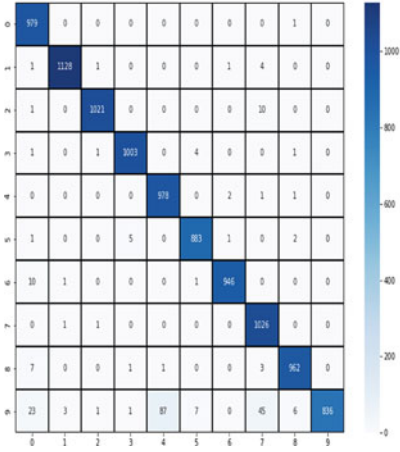
Detecting and accommodating novelties in real-time and taking prompt countermeasures is of paramount importance for any AI system. In this paper, we described a transfer learning-assisted novelty detection architecture for AI systems. Transfer learning can enhance the generalizability capability of novelty detection systems and reduce the training overhead (in terms of computation and data resources) compared to traditional supervised learning techniques for resource-constrained agents, thus, ensuring the adaptability of AI agents in dynamic environments. We have conduct



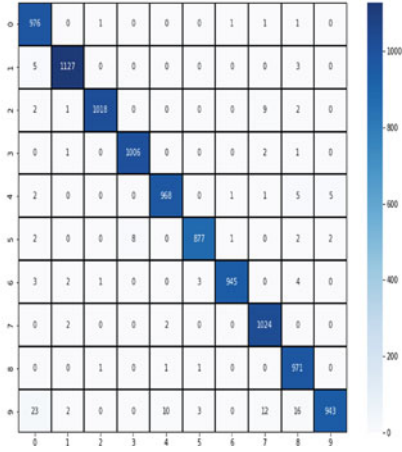
Neural: MobileNet, Source: ImageNet, Target: MNIST



Neural: MobileNet, Source: ImageNet, Target: MNIST



Neural: ResNet50, Source: ImageNet, Target: MNIST



Neural: ResNet50, Source: ImageNet, Target: MNIST

Fig. 4 Transfer Learning based Novelty Class detection with fewer novelty samples (< 0.5%) compared to non-novelty samples. We denote class ‘9’ in MNIST dataset as novelty class and results show on novelty sizes of (a) 100, (b) 200, (c) 100, and (d) 200

several preliminary experiments to gauge the efficacy of TL-based novelty detection and accommodation system. In the future, we will focus on developing robust novelty accommodation techniques and conduct an empirical study to gauge their efficacy.

Acknowledgements This research is supported, in part, by the Defense Advanced Research Projects Agency (DARPA) and 3th Air Force Research Laboratory (AFRL) under the contract number W911NF2020003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, or the U.S. Government.

References

1. Arifuzzaman M, Islam S, Arslan E (2021) Towards generalizable network anomaly detection models. In: 2021 IEEE 46th conference on local computer networks (LCN), pp 375–378. IEEE
2. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. *ACM Comput Surv (CSUR)* 41(3):1–58
3. Cooper S, Bhuiyan M, Arslan E (2020) Machine learning for data transfer anomaly detection. In: *IEEE/ACM supercomputing*
4. Haliem M, Aggarwal V, Bhargava BK (2021) Novelty detection and adaptation: a domain agnostic approach. In: Jain S, Groppe S (eds) *Proceedings of the international semantic intelligence conference 2021 (ISIC 2021)*, New Delhi, India, February 25–27, 2021. *CEUR Workshop Proceedings*, vol 2786, pp 73–77. *CEUR-WS.org*. <http://ceur-ws.org/Vol-2786/Paper11.pdf>
5. Hou B, Hou C, Zhou T, Cai Z, Liu F (2021) Detection and characterization of network anomalies in large-scale rtt time series. *IEEE Trans Netw Serv Manag* 18(1):793–806
6. Kliger M, Fleishman S (2018) Novelty detection with gan. [arXiv:1802.10560](https://arxiv.org/abs/1802.10560)
7. Miljković D (2010) Review of novelty detection methods. In: *The 33rd international convention MIPRO*. IEEE, pp 593–598
8. Nesen A, Bhargava B (2020) Knowledge graphs for semantic-aware anomaly detection in video. In: *2020 IEEE third international conference on artificial intelligence and knowledge engineering (AIKE)*. IEEE, pp 65–70
9. Nesen A, Solaiman K, Bhargava B (2021) Dataset augmentation with generated novelties. In: *2021 third international conference on transdisciplinary AI (TransAI)*. IEEE, pp 41–44
10. Pimentel MA, Clifton DA, Clifton L, Tarassenko L (2014) A review of novelty detection. *Signal Process* 99:215–249
11. Rausch A, Sedeh AM, Zhang M (2021) Autoencoder-based semantic novelty detection: towards dependable ai-based systems. *Appl Sci* 11(21):9881
12. Yadav M, Malhotra P, Vig L, Sriram K, Shroff G (2016) Ode-augmented training improves anomaly detection in sensor data from machines. [arXiv:1605.01534](https://arxiv.org/abs/1605.01534)
13. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2020) A comprehensive survey on transfer learning. *Proc IEEE* 109(1):43–76

An Ontology for Social Media Data Analysis



Sarika Jain, Sumit Dalal, and Mayank Dave

Abstract Social media is one of the valuable information sources which present much data to the researchers. This information is mainly analyzed by machine learning and the deep learning methods, which is focused on feature engineering. We present a taxonomy of the different features. These features are learnt during the training for analyzing the textual information, specifically available on social platforms. This ontological view of the data will represent knowledge in a more understandable form besides interpreting the machine learning results for various tasks related to the social data analysis. We have taken Depression as the use-case purpose. The ontology is designed in the Protégé. The validation of the ontology is carried out with designed competency questions.

Keywords Deep learning · Knowledge graph · Depression · Machine learning · Ontology · Social MediaData · Twitter

1 Introduction

Mental health is an essential aspect of human life to live a productive and energetic life. People tend to ignore their mental health for various reasons like inaccessible health services, limited time for themselves, etc. Nevertheless, technological advancements provide researchers an opportunity for pervasive monitoring to include users' social data for their mental health assessment without interfering with their daily life. People share their feelings, emotion, and daily activities related to work

S. Jain (✉) · S. Dalal

Department of Computer Applications, National Institute of Technology, Kurukshetra 136119, India

e-mail: jasarika@nitkkr.ac.in

S. Dalal

e-mail: sumitdalal9050@gmail.com

M. Dave

Computer Engineering, National Institute of Technology, Kurukshetra 136119, India

e-mail: mdave@nitkkr.ac.in

and family on social media platforms (Facebook, Twitter, Reddit). These posts can be used for extracting features or looking for particular words, phrases that can be used to assess if a user has depression or not.

Various machine learning and deep learning methods have been devised and applied for mental health assessment from users' social data. These techniques mainly consider correlation information and miss contextual information of the domain. Analyzing social data with traditional statistical and machine learning approaches has limitations like poor big data handling capacity, semantics, and contextual/background knowledge inclusion. So a hybrid approach that handles semantics and big data should be considered for better results.

We develop an ontology to represent the domain information related to the social data analysis. Ontology is a formalization of a domain's knowledge [1]. The main principles of ontology are to re-use sharing domain knowledge between agents in a language understandable to them (user or software). Ontology has been developed and used in different application domains. [2, 3] design ontology for analyzing user reviews in social media. [4, 5] employ ontology for text documents' classification while [6] use ontology for email classification. [7, 8] developed an ontology for collecting and analyzing the COVID-19 data. [9] develop an ontology for disease event detection from Twitter. [10] use topic modeling methods to extract essential topics from transportation planning documents for constructing intelligent transportation infrastructure planning ontology. However, ontology-based techniques for depression classification and monitoring from social data have been insufficiently studied.

The machine learning and statistical approaches consider limited contextual information. Moreover, it is not easy to interpret their results. For this reason, deep learning models are considered complete black boxes. Personalization of the system is another issue that needs to be in focus. For the implementation purpose, we chose depression as the domain. We aim to develop an underlying ontology for the personalized and disease-specific knowledge graph, to monitor a depressive user through his publicly available textual social data. The ontology is designed using Ontology Web Language and Resource Description Framework in the Protégé. The validation of the ontology is carried out with designed competency questions. Our contributions in this paper are as follows:

1. To develop an ontology for analyzing social media posts. The features of social posts manipulated by machine learning and deep learning techniques are arranged in a taxonomy. This way, structured data help in the interpretation of output produced.
2. We write competency questions to describe the scope of FeatureOnto to detect and monitor depression through social media posts.

The remaining paper is organized into four sections. Section 2 discusses the related literature. The FeatureOnto development approach and its scope will be discussed in Sect. 3. Section 4 discusses the conceptual design of the FeatureOnto and the evaluation of the same. Conclusion and future work is discussed in the last section.

2 Literature

This section discusses previous research on depression ontology development using various sources or employing the available ontology for depression detection or monitoring.

2.1 *Ontology-Based Sentiment Analysis*

Sentiment analysis is a crucial aspect in detecting depression from social posts. However, there are applications other than mental health assessment where it is functional. Sentiment extraction of user posts/reviews is a popular application that considers the affective features of the posts. The authors consider eight emotion categories to develop an emotion ontology for the sentiment classification [11]. [12] employ entity extraction tools for extracting entities and mapping semantic concepts from user reviews. They use the extracted semantic features with unigrams for Twitter sentiment analysis. [13] apply the ontology-based approach for sentiment analysis.

2.2 *Ontology in Healthcare Domain*

In the healthcare domain, ontologies have been employed for quite a long time. [14] employ Unified Medical Language System (UMLS) ontology to extract health-related named entities from user tweets. [15] implement DBpedia, Freebase, and YAGO2 ontologies for determining behavior addiction category of social users. [16] develop ontology as a bridge between the device and space-specific ontologies for ubiquitous and personalized healthcare service environment. [17] build ontology as a catalog of drug abuse, use, and addiction concepts for the social data investigation. [18] extract concepts and their relations from clinical practice guidelines, literature, and social posts to build an ontology for social media sentiment analysis on childhood vaccination. [19] build ontology with personality traits and their facets as classes and sub-classes, respectively, for personality measurements from Twitter posts. [20] design a monitoring framework for diabetes and blood pressure patients that consider various available ontologies such as medical domain and patient's medical records, wearable sensor, and social data.

2.3 *Depression Monitoring and Ontology*

Authors employ ontologies in depression diagnosis and monitoring. Either they build ontology or use available ones. [21] propose to add temporal dimension in ontology

Table 1 Comparison of the FeatureOnto and other depression ontologies used in literature

	Main classes	Dimensions covered	Entities source considered	Availability
O1	Depression, symptom, activity	D1, D2, D3, D4, D5, D6	Literature	–
O2	Symptoms, treatments, life, feelings	D1, D6	SNSs	–
O3	Diagnostics, sub-types, risk factors, sign and symptoms, intervention	D6	CPG, literature, SNSs, FAQs	–
O4	Patient, disease, symptom	D2, D3	Literature	–
O5	Patient, doctor, activity, diagnosis, treatment diary	D1, D2, D3, D4	General scenario	–
FeatureOnto	Patient, symptom, posts, user profile, feature	D2, D3, D6, D7	Literature	Yes

for analyzing depressed user’s linguistic patterns and ontology evolution over time in his social data. [22] represent explicitly defined patient data, self-questionnaire, and diagnosis result in the semantic network for preventing and detecting depression among cancer patients. [23] construct a vocabulary of suicide-related themes and divide them into sub-classes concerning the degree of threat. WordNet is further used for semantic analysis of machine learning predictions for suicide sentiments on Twitter. Some works that build an ontology for depression diagnosis are discussed below. We assign ontologies unique IDs such as O1, O2, etc. These IDs are used in Table 1 to mention the particular ontology.

O1. [24] provide a ubiquitous framework based on ontology to assist the treatment of people suffering from depression. The ontology consists of concepts related to the user’s depression, person, activity, and depression symptoms. Activity has sub-classes related to the social network, email, and geographical activities. The person has a sub-class Person type which further defines a person into User, Medical, and Auxiliary. We are not sure if a patient history is considered or not.

O2. [25] extract concepts and their relationships from posts on the daily-Strength, to develop the Onto-Depression ontology for depression detection. They use tweets of family caregivers of Alzheimer’s. Onto-Depression has four main classes: Symptoms, Treatments, Feelings, and Life. The symptom is categorized into general, medical, physical, and mental. Feelings represent positive and negative aspects. Life

Table 2 Description of different dimensions considered

Dimension	Dimension ID	Description
Activity	D1	This facet covers physical movements, social platforms, daily life activities, etc.
Clinical record	D2	It is related to patient profile; provides historical context; and covers clinical tests, physician observations, treatment diary, schedules, etc.
Patient profile	D3	The dimensions cover disease symptoms, education, work condition, economical, relationship status, family background, etc.
Physician profile	D4	This aspect describes a physician in terms of his expertise, experience, etc.
Sensor data	D5	This element is related to the smartphone, body, and background sensors
Social posts	D6	It is affiliated with the content of posts by a user on SNS
User’s social profile	D7	Social media profile provides an essential aspect of user personality

class captures what the family caregivers are talking about. Treatments represent concepts of medical treatment.

O3. [26, 27] develop ontology from clinical practice guidelines and related literature to detect depression in adolescents from their social data. The ontology consists of five main classes: measurement, diagnostic result and management care, risk factors, and sign and symptoms.

O4. [28] build an ontology for depression diagnosis using Bayesian networks. The ontology consists of three main classes: Patient, Disease, and Depression_Symptom. Depression symptoms are categorized into 36 symptoms.

Table 3 Schema-based competency questions

Competency questions
1. Retrieve the labels for every sub-class of the class Content?
2. “Topics” is the sub-class of?
3. What type of feature is “Anger”?

Table 4 Knowledge graph-based competency questions

Competency questions
1. What is the sleeping pattern of a user/patient (user can be normal patient)?
2. In which hour user messages frequently?
3. How many posts has low valence in a week?
4. Emotional behavior pattern considering week as a unit?
5. Daily/weekly average frequency of negative emotions?
6. Compare daily/weekly/overall average number of first person pronoun and second/third person pronouns?
7. What are the topics of interest for a depressed user?
8. Anger-related words used frequently or not?
9. Find the pattern of psycho-linguistic features?

O5. [29] developed ontology based on Cognitive Behavioral Theory (CBT) to diagnose depression among online users at the current stage. Their focus is to lower the threshold access of online CBT. The ontology consists of the patient, doctor, patient record, and treatment diary concepts.

Work in [30] created ontology for social media users to detect suicidal ideation from their knowledge graph. Their work is similar to our work but they considered limited features taxonomy, moreover we focus depression detection from personalized knowledge graph.

Table 1 compares distinct ontologies built in different research papers to detect and monitor depression which are compared on four parameters (Main Classes, Dimensions Covered, Entities Source Considered, and Availability and Re-usability). We extracted seven dimensions (Activity, Clinical Record, Patient Profile, Physician Profile, Sensor Data, Social Posts, and Social Profile) from the related literature. A description of each dimension, along with dimension ID, is given in Table 2. We cannot find the ontologies built by other authors online. We are not sure if these are available for re-use or not, so the Availability and Re-usability column is blank. O1 ontology has scope over almost all the dimensions we have considered.

3 Designing FeatureOnto Ontology

The focus of ontology development is to analyze the social textual data and interpret the results produced by the machine learning or deep learning models. Mainly, authors focus on n-gram features of social media posts, but FeatureOnto also considers other features. We follow the ‘‘Ontology Development 101’’ methodology for FeatureOnto development [31]. An iterative process is followed while designing the ontology life cycle.

Step 1. Determining Domain and Scope of the Ontology

We create a list of competency questions to determine the ontology's domain and scope [32]. FeatureOnto ontology should be able to answer these questions, e.g., what are the textual features of social media posts? The ontology will be evaluated with these questions. Tables 3 and 4 provide the sample of competency questions where Table 3 is derived to check the ontology schema, i.e., ontology without any instance. In comparison, questions in Table 4 are derived keeping in mind the use case of depression monitoring of a social user. Queries of Table 4 are out of scope for this paper as here we are only presenting the schema.

Step 2. Re-using the Existing Ontologies

We search for available conceptual frameworks and ontologies on social data analysis at BioPortal [33], OBOFoundry, and LOD cloud. Ontologies representing sentiment analysis or depression classification, or other social media analysis tasks on the web (Google Scholar, Pubmed) and the kinds of literature are searched for the required concepts and relationships. We have done a comprehensive search but could not find a suitable ontology that could be re-used fully. We find some ontology and can inherit one or more classes from them. Most of the inherited classes are given attributes as per our requirements. Table 5 shows our efforts toward implementing the re-usability principle of the semantic web. Figure 3 is presented in the next section and gives a diagrammatical representation of the inherited entities. Different colors define each schema. The solid and the dotted line shows immediate and remote child-parent relations between classes. Most inherited entities belong to Schema, MFOEM, and HORD, while APAONTO, Obo are the least inherited ontologies. We did not find suitable classes for UniGrams, BiGrams, Emoticon, and POSTags. So we use our schema to represent these classes. The solid and the dotted lines represent the property and the sub-class relationship between two entities.

Step 3. Extracting Terms and Concepts

Keeping in mind our use case, we read literature on depression and mental disorders detection from social data using machine learning or lexicon-based approaches and extract terms related to features considered for classification. We found that different textual features are extracted and learned in machine learning or deep learning training phase [34, 35], e.g., bigrams, unigrams, positive, or negative sentiment words. Table 5 shows different entities and sub-entities present in the FeatureOnto ontology. It also provides information about the various available schemas for an entity and the schema used for the inheritance. We also search social networking data to extract additional terms. The extracted terms are used for describing the class concepts.

Step 4. Developing and Ontology and Terminology

We have defined the classes and the class hierarchy using the top-down approach. The ontology is developed using Protégé [36]. The ontology is uploaded on BioPortal.

Table 5 Entities and namespaces considered in the FeatureOnto

Entity	Sub-entities	Schema selected	Available schemas
Content	UniGrams, BiGrams, POSTags	–	–
Emoticon	–	–	–
Emotion	Arousal, positive valence, negative valence	MFOEM	MFOEM, SIO, VEO
	Dominance	APAONTO	APAONTO, FB-CV
GenderType	Schema	Schema, GND	–
Person	Patient	Schema	FOAF, Schema, Wiki-data, DUL
	Social media user	HORD	NCIT ² , SIO, HORD
Post	–	HORD	HORD
Psycho linguistic	Anger, anxiety, sad	MFOEM	MFOEM, SIO, VEO, NCIT
	Pronoun	–	–
Symptoms	–	Obo	NCIT, SYMP, RADLEX, Obo
Topic	–	–	EDAM, ITO

Step 5. Evaluating the Scope of the Ontology

A set of competency questions is given in Tables 3 and 4. For scope evaluation of the FeatureOnto, answers to the SPARQL queries built on the questions from Table 3 are considered. Results of the queries are discussed in the coming sections.

4 FeatureOnto Ontology Model

Following the steps discussed in the previous section, we design FeatureOnto ontology. A high-level view of the FeatureOnto ontology is presented in Fig. 1. The complete FeatureOnto structure (at the current stage) has five dimensions (Patient, Symptom, Posts, User Profile, and Feature) covered by various classes in the figure. Most of the entities in our ontology belong to the Social Post dimension. The solid and the dotted lines represent the property and the sub-class relationship between two entities. Figure 1 gives a conceptual schema of the proposed model. FeatureOnto uses existing ontologies to pursue the basic principle of ontology implementation.

Scope Evaluation of FeatureOnto: Tables 3 and 4 present the competency questions related to the schema and instances. This work is related to the building of the schema only, and hence we executed queries on schema only. Below, queries are built on questions from Table 3.

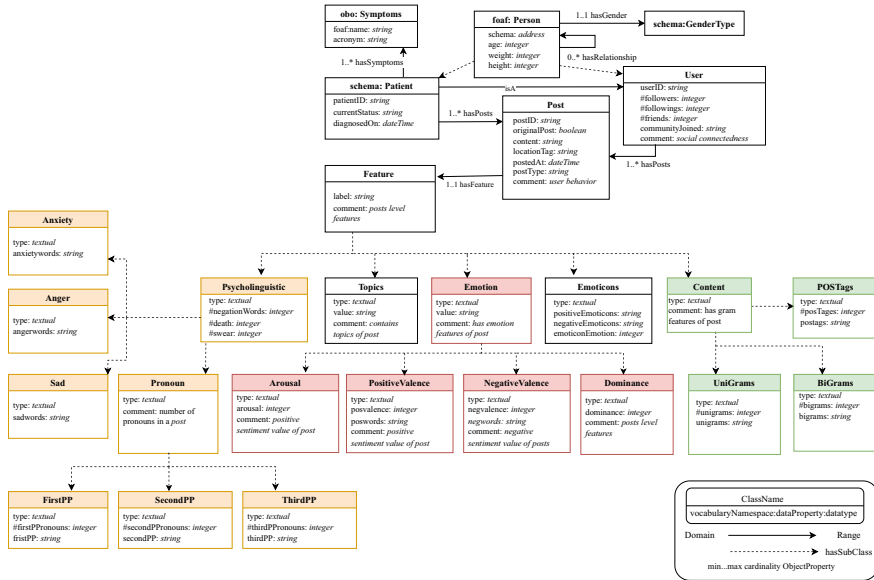


Fig. 1 Conceptual design of the FeatureOnto

5 Conclusion

We developed a first version of the social data analysis ontology to provide a taxonomy of social media posts’ features (use mental health assessment or depression classification/monitoring is taken). Posts carry huge information regarding many aspects. This information can be placed into different feature categories. These features are widely used in sentiment analysis, mental health assessment, event detection, user profiling, document classification, and other natural language and image processing tasks. The ontology will be used to create a personalized depression knowledge graph in the future. For this reason, it does not focus on the concepts from clinical practice guidelines and depression literature at the current stage. We will also extend the ontology to include other concepts related to depression in the future.

References

1. Gruber TR (1995) Toward principles for the design of ontologies used for knowledge sharing? Int J Hum-Comput Stud 43(5–6):907–928
2. Konjengbam A, Dewangan N, Kumar N, Singh M (2018) Aspect ontology based review exploration. Electron Commer Res Appl 30:62–71
3. Wang D, Xu L, Younas A (2018) Social media sentiment analysis based on domain ontology and semantic mining. In: International conference on machine learning and data mining in pattern recognition. Springer, pp 28–39

4. Malik S, Jain S (2021) Semantic ontology-based approach to enhance text classification. In: International semantic intelligence conference, Delhi, India, 25–27 Feb 2021. CEUR Workshop Proceedings, vol 2786, pp 85–98 (2021)
5. Allahyari M, Kochut KJ, Janik M (2014) Ontology-based text classification into dynamically defined topics. In: IEEE international conference on semantic computing. IEEE, pp 273–278
6. Taghva K, Borsack J, Coombs J, Condit A, Lumos S, Nartker T (2003) Ontology-based classification of email. In: Proceedings ITCC 2003. International conference on information technology: coding and computing. IEEE, pp 194–198
7. Dutta B, DeBellis M (2020) CODO: an ontology for collection and analysis of COVID-19 data. arXiv preprint. [arXiv:2009.01210](https://arxiv.org/abs/2009.01210)
8. Patel A, Debnath NC, Mishra AK, Jain S (2021) Covid19-IBO: a Covid-19 impact on Indian banking ontology along with an efficient schema matching approach. *New Gener Comput* 39(3):647–676
9. Magumba MA, Nabende P (2016) Ontology driven disease incidence detection on Twitter. arXiv preprint. [arXiv:1611.06671](https://arxiv.org/abs/1611.06671)
10. Chowdhury S, Zhu J (2019) Towards the ontology development for smart transportation infrastructure planning via topic modeling. In: ISARC. Proceedings of the international symposium on automation and robotics in construction, vol 36. IAARC Publications, pp 507–514
11. Sykora M, Jackson T, O'Brien A, Elayan S (2013) Emotive ontology: extracting fine-grained emotions from terse, informal messages
12. Saif H, He Y, Alani H (2012) Semantic sentiment analysis of twitter. In: International semantic web conference. Springer, pp 508–524
13. Kardinata EA, Rakhmawati NA, Zuhroh NA (2021) Ontology-based sentiment analysis on news title. In: 2021 3rd East Indonesia conference on computer and information technology (EIConCIT). IEEE, pp 360–364
14. Batbaatar E, Ryu KH (2019) Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach. *Int J Environ Res Public Health* 16(19):3628
15. Krishnamurthy M, Mahmood K, Marcinek P (2016) A hybrid statistical and semantic model for identification of mental health and behavioral disorders using social network analysis. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 1019–1026
16. Kim J, Chung K-Y (2014) Ontology-based healthcare context information model to implement ubiquitous environment. *Multimed Tools Appl* 71(2):873–888
17. Lokala U, Daniulaityte R, Lamy F, Gaur M, Thirunarayan K, Kursuncu U, Sheth AP (2020) DAO: an ontology for substance use epidemiology on social media and dark web. *JMIR Public Health Surveill*
18. On J, Park H-A, Song T-M (2019) Sentiment analysis of social media on childhood vaccination: development of an ontology. *J Med Internet Res* 21(6):e13456
19. Alamsyah A, Putra MRD, Fadhilah DD, Nurwianti F, Ningsih E (2018) Ontology modelling approach for personality measurement based on social media activity. In: 2018 6th international conference on information and communication technology (ICoICT). IEEE, pp 507–513
20. Ali F, El-Sappagh S, Islam SR, Ali A, Attique M, Imran M, Kwak K-S (2021) An intelligent healthcare monitoring framework using wearable sensors and social networking data. *Futur Gener Comput Syst* 114:23–43
21. Martín-Rodilla P (2020) Adding temporal dimension to ontology learning models for depression signs detection from social media texts. In: Proceedings of ENASE, pp 323–330
22. Benfares C, Idrissi YEBE, Hamid K (2018) Personalized healthcare system based on ontologies. In: International conference on advanced intelligent systems for sustainable development. Springer, pp 185–196
23. Birjali M, Beni-Hssane A, Erritali M (2017) Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Comput Sci* 113:65–72

24. Petry MM, Barbosa JLV, Rigo SJ, Dias LPS, Büttendörfer PC (2020) Toward a ubiquitous model to assist the treatment of people with depression. *Univ Access Inf Soc* 19(4):841–854
25. Kim HH, Jeong S, Kim A, Shin D (2018) Analyzing Twitter data of family caregivers of Alzheimer's disease patients based on the depression ontology. *Advances in computer science and ubiquitous computing*. Springer, pp 30–35
26. Jung H, Park H, Song T-M (2016) Development and evaluation of an adolescents' depression ontology for analyzing social data. *Nursing informatics 2016*. IOS Press, pp 442–446
27. Jung H, Park H-A, Song T-M (2017) Ontology-based approach to social data sentiment analysis: detection of adolescent depression signals. *J Med Internet Res* 19(7):e7452
28. Chang Y-S, Fan C-T, Lo W-T, Hung W-C, Yuan S-M (2015) Mobile cloud-based depression diagnosis using an ontology and a Bayesian network. *Futur Gener Comput Syst* 43:87–98
29. Hu B, Hu B, Wan J, Dennis M, Chen H-H, Li L, Zhou Q (2010) Ontology-based ubiquitous monitoring and treatment against depression. *Wirel Commun Mob Comput* 10(10):1303–1319
30. Cao L, Zhang H, Feng L (2020) Building and using personal knowledge graph to improve suicidal ideation detection on social media. *IEEE Trans Multimed*
31. Noy NF, McGuinness DL (2001) Ontology development 101: a guide to creating your first ontology. Stanford knowledge systems laboratory technical report KSL-01-05 and ..
32. Grüninger M, Fox MS (1995) Methodology for the design and evaluation of ontologies
33. Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story M-A, Smith B, team N (2012) The national center for biomedical ontology. *J Am Med Inform Assoc* 19(2):190–195
34. Dalal S, Jain S, Dave M (2019) A systematic review of smart mental healthcare. In: *Proceedings of the 5th international conference on cyber security & privacy in communication networks (ICCS)*
35. Dalal S, Jain S (2021) Smart mental healthcare systems. *Web semantics*. Elsevier, pp 153–163
36. Musen MA (2015) The protégé project: a look back and a look forward. *AI Matters* 1(4):4–12

The Coronavirus Disease Ontology (CovidO)



Sumit Sharma and Sarika Jain

Abstract This paper presents the Coronavirus Disease Ontology (CovidO), a superset of the available Coronavirus (COVID-19) ontologies, including all the possible dimensions. CovidO consists of an ontological network of thriving distinct dimensions for storing coronavirus information. CovidO has 175 classes, 169 properties, 4141 triples, 645 individuals with 264 nodes and 308 edges. CovidO is based on standard input of coronavirus disease data sources, activities, and related sources, which collects and validates records for decision-making used to set guidelines and recommend resources. We present CovidO to a growing community of artificial intelligence project developers as pure metadata and illustrate its importance, quality, and impact. The ontology developed in this work addresses grouping the existing ontologies to build a global data model.

Keywords Coronavirus ontology · Ontology learning · Semantic metadata · Semantic web · COVID-19

1 Introduction and Motivation

The World Health Organization declared a Public Health Emergency of International Concern on 30 January 2020 and a pandemic on 11 March 2020 [1]. At the same time, novel coronavirus (COVID-19) pandemic data has been collected by various data sources like the World Health Organization (WHO) and DXY.CN, BNO News, National Health Commission (NHC) of the People's Republic of China, China Cen-

S. Sharma (✉) · S. Jain
National Institute of Technology Kurukshetra, Kurukshetra 136119, India
e-mail: sharma24h@gmail.com

S. Jain
e-mail: jasarika@nitkkr.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
S. Jain et al. (eds.), *Semantic Intelligence*, Lecture Notes in Electrical Engineering 964,
https://doi.org/10.1007/978-981-19-7126-6_8

ters for Disease Control and Prevention (CCDC), Hong Kong Health Department, Macau Government, Taiwan CDC, US Centers for Disease Control and Prevention (CDC), Government of Canada, Australian Government Department of Health, European Centers for Disease Prevention and Control (ECDC), Ministry of Health Singapore (MOH), and others. The COVID-19 data sources and access links are shown below.

- WHO. <https://www.who.in>.
- DXY.cn. <https://www.dxy.cn>.
- BNO News. <https://bnonews.com>.
- National Health Commission (NHC). <http://en.nhc.gov.cn>.
- China CDC (CCDC). <https://www.chinacdc.cn/en/>.
- Hong Kong Health Department. <https://www.dh.gov.hk>.
- Macau Government. <https://www.gov.mo/en/>.
- Taiwan CDC. <https://www.cdc.gov.tw/En>.
- US CDC. <https://www.cdc.gov/>.
- Government of Canada. <https://www.canada.ca/en.html>.
- Australian Government. Department of Health <https://www.health.gov.au/>.
- ECDC. <https://www.ecdc.europa.eu/en>.
- Ministry of Health Singapore (MOH). <https://www.moh.gov.sg/>.
- Ministry of Health and Family Welfare. <https://www.mohfw.gov.in/>.

A significant issue is that the data sources related to COVID-19 are heterogeneous, static, and broad in scope. So many heterogeneous and stationary data sources create situations where data is sometimes under-utilized or, in more extreme cases, not used for the decision-making process [2]. Another vital issue of COVID-19 is to provide semantic (machine understandable) representation of data from various exciting fields such as research, health, resources, drugs, and treatment. Ontology is emerging to solve these issues. Ontology solves the problem of changing user expectations and data integration demands driven by its volatility in a rapidly growing digital market and societal challenges related to resource efficiency [3]. Ontologies have proven practical tools for representing domain knowledge, integrating data from disparate sources, and supporting many semantic applications [4].

This paper presents an OWL-based Coronavirus disease ontology (CovidO) that defines all the possible aspects and relations to describe COVID-19. To develop CovidO, we have defined seven dimensions that cover all essential aspects of COVID-19: (1) About the COVID-19 infectious disease, symptoms, drugs, and treatment; (2) Information statistics of COVID-19 cases in a geographical region; (3) COVID-19 patients information with the cause of infection and exposure of pandemic; (4) COVID-19 related resources and their availability in a location; (5) Impact of COVID-19 in different verticals like education, finance, business, research and social; (6) Various guidelines and prevention and vaccine mandates by public authority; (7) Global and biomedical research on COVID-19. Several ontologies (e.g., Infectious Disease Ontology (IDO), Virus Infectious Disease Ontology (VIDO), Coronavirus Infectious Disease Ontology (CIDO), etc.) have been created for coronavirus one after another. They comprehensively and thoroughly describe coronavirus disease.

These ontologies individually fail to cater to all the dimensions of coronavirus. The ontology developed in this paper aims to group existing ontologies to construct a common global data model with the unified purpose. The main contributions of this paper are:

- to provide a list of dimensions to cover every aspect of coronavirus-related information.
- to develop standard metadata (Providing a schema) called CovidO as a global data model to annotate the Covid-19 information.

This paper presents our work to be independent of external ontologies, we defined a new namespace <https://w3id.org/CovidO> with the prefix CovidO (registered entry at <http://prefix.cc>) for all classes used in the ontology. As a permanent URL service, we use w3id.org.

The content of this paper is organized as follows. The related work describes literature on the various existing COVID-19 associated ontologies and their scope boundary. The dimension section describes the incidence of the COVID-19 information dimensions in the primary analysis and provides a summary analysis. The conceptual design and scope of CovidO are described in the CovidO section to build the ontology. The ontology design section defines the competency questions to determine the scope of CovidO with abstract design. The method section outlines the model to develop CovidO. In the evaluation section, we present the evaluation for the COVID-19 schema, in addition to simple predictions for the future incidence of COVID-19. Some concluding remarks are given in the conclusion.

2 Related Work

Several ontologies represent the COVID-19 pandemic in different contexts. We have found some ontologies related to COVID-19, each representing a completely different scope of COVID-19. They are briefed here:

- O1: COVID-19 Infectious Disease Ontology (IDO-COVID-19):** IDO-COVID-19 [5] is the most particular version of CIDO (Coronavirus Infectious Disease Ontology) [6], containing information about COVID-19 and its cause SARS-CoV-2. IDO-COVID-19 adheres to the OBO Foundry design philosophy by extending the CIDO in the same way as the CIDO extends VIDO (Virus Infectious Disease Ontology) [5] and the VIDO extends the IDO (Infectious Disease Ontology) [7]. IDO-COVID-19 also pulls concepts from other ontologies, such as SARS-CoV-2, imported from NCBITaxon. IDO-COVID-19, CIDO, and VIDO ontologies consist of concepts related to disease dimension, other dimensions are not covered.
- O2: COviD-19 Ontology for the case and patient information (called CODO):** CODO [8] is an ontology that contains COVID-19 case data in a format that can be used by other ontologies and software systems and is based only on OWL

and different W3C standards. CODO tracks specific pandemic cases, including information such as how the patient is considered to have been infected and potential further contacts who may be at risk owing to their association with the infected individual. CODO also tracks clinical tests, travel history, available resources, actual demand (e.g., ICU bed, invasive ventilators), trend analyses, and forecast increases. The CODO ontology covered the cases, patients, and resources dimensions.

- O3: **COVID-19 surveillance Ontology (COVID-19):** COVID-19 surveillance Ontology [9] is a COVID-19 application ontology that intends to provide COVID-19 cases and respiratory information by obtaining data from multiple medical records systems. This ontology is constructed as a taxonomy with only 32 classes. COVID-19 verified by a lab test, SARS-CoV-2 identified, Probable COVID-19, Clinical codes, Possible COVID-19, Suspected COVID-19, Under investigation, Exposure, COVID-19 excluded are the ten core ideas of COVID-19 ontology. The COVID-19 ontology was created using the protégé tool, and its format is based on the OWL language.
- O4: **COVID19-IBO:** The COVID-19 Impact on Banking Ontology (COVID-19-IBO) [10] is a knowledge graph which covers semantically the COVID-19's impact on Indian banking industry under the "Impact on business vertical" dimension. In addition, the authors have provided a schema matching technique with satisfactory results for mapping the COVID-19 ontologies.
- O5: **Kg-COVID-19:** The KG-COVID-19 [11] framework is used to create customized COVID-19 knowledge graphs. The FAIR (Findable, Accessible, Interoperable, and Reusable) approach is followed by KG-COVID-19, which combines various COVID-19 heterogeneous biomedical data and covers the COVID-19 disease, symptoms, and treatment dimensions.
- O6: **COVID-19 Ontology:** The COVID-19 ontology includes the function of molecular and cellular entities in viral-host interactions throughout the virus life cycle and a wide range of medical and epidemiological concepts associated with COVID-19. A scalable new coronavirus (SARS-CoV-2) entity is represented as an ontology. As a prominent target of ongoing COVID-19 medicinal research, the ontology contains a broad scope on chemical entities ideal for drug repurposing. The ontology's performance was evaluated using Medline and the Allen Institute's COVID-19 corpus.
- O7: **DRUGS4COVID19:** DRUGS4COVID19 [12] identifies drugs and their associations with COVID-19. The ontology's core concepts include drug, effect, disease, symptoms, disorder, chemical substance, and so on.
- O8: **COVIDCRFRAPID:** The World Health Organization's (WHO) COVIDCR-FRAPID [13] ontology is a semantic data model for the COVID-19 RAPID case record form. COVIDCRFRAPID ontology provides semantic references to the form filled by patients during the treatment as questions and responses. It shows a variety of application scenarios, including graph-based machine learning.

O9: **ROC**: Ontology (Country Responses toward COVID-19) ROC [14] enables data integration from heterogeneous data sources and answers interesting questions. ROC was designed to assist statistical analysis in exploring and analyzing the efficacy and side effects of government responses to COVID-19 in various nations.

We investigate these existing ontologies, focusing on a group of individuals to discuss a specific topic like drug, protein interaction databases, protein function annotations, COVID-19 patients, and cases. So they have a limited scope that does not cover all aspects of COVID-19. We have found that these ontologies refer to the COVID-19 pandemic but represent different aspects and scopes. We fill this gap, and our work follows the same approach to ontology design and has a common motivation. We have conglomerated all these ontologies into a comprehensive design covering all the required distinct dimensions (COVID-19 cases information, patient information, disease-symptom-treatment, resources, COVID-19 impact, research, and event or news related to COVID-19) discussed in the next section. By adopting established models, we aim to facilitate integration, linking, and reuse across the data sources and make data accessible to a wide range of applications. In addition, new entities have been introduced required.

3 COVID-19 Information Dimensions

With a view to allowing stakeholders in the research community and application developers to reach out and benefit, CovidO has been created as a platform through specific dimensions. There are many domains and sub domains; we group these domains into seven significant divisions that cover all the aspects of COVID-19 related knowledge. These seven dimensions are shown in detail in Table 1. The CovidO ontology is developed based on these dimensions representing COVID-19 information in OWL format and other W3C standards utilized by other ontologies and software systems. The last column of Table 1 describes the core CovidO classes associated with a particular dimension. CovidO allows detailed tracking of specific pandemic dimensions. For example, the diseases and treatment dimension includes clinical test tracking, test report history, illness, symptoms, medication, and clinical measurement and diagnosis. Similarly, CovidO traces other dimensions as well. In brief and with overall dimensions, CovidO monitors the COVID-19 patient's travel history, symptoms, medication, available healthcare facilities, resources, actual need (e.g., ambulance, invasive ICU bed with ventilators), trend study, impact on business verticals, research publications finding, guidelines for public health Safety, news, and growth projections. To the best of our knowledge, we have not found any ontology that describes all the seven dimensions D1 to D7. Nor was any such ontology found which could provide an interlink between them. All ontologies have different scopes and common goals to provide the schema for COVID-19 data.

Table 1 Seven dimensions covering the COVID-19 information

S. No.	Dimension	Description	Core classes in CovidO
1	D1: COVID-19 Cases Information	Attribute a COVID-19 case such as active, recovered, deceased, migrated cases daily across the Geo-location (district, state, and country)	Statistics
2	D2: COVID-19 Patient Information	Represents the COVID-19 patients. Patient symptoms, suspected COVID-19 cases, COVID-19 treatment facility, patient travel history, patient nationality, interpersonal relationships between patients, supposed transmission reason, tracking of the patient test, etc.	Patient
3	D3: COVID-19 Disease, Symptom and Treatment	Describe the various diseases, different variants of disease, viruses and discover their symptom, treatment for disease	Symptom, Disease, Treatment, Vaccine
4	D4: COVID-19 Resources	COVID-19 clinical facility (COVID center, hospital, ambulance, available test), Availability of resources (Doctor, Nurse, Medical Equipment, Medicines, etc.). Availability of bed, ICU bed with oxygen	Resources, COVID-19 Clinical Facility
5	D5: COVID-19 Impact on business vertical	Impact of COVID-19 on various sectors like education sector, banking sector, economic, etc.	COVID-19Impact
6	D6: COVID-19 Related Event and Decision	Lockdown, guideline issued by government, requirement of mask, sanitizer spray, awareness program, event hosting, various exposure of COVID-19 etc.	Lockdown, Prevention, Exposure
7	D7: COVID-19 Research Domain	Provide the information regarding the research on Coronavirus diseases and findings	COVID-19 Research, COVID-19 News

4 The Coronavirus Disease Ontology (CovidO)

The CovidO knowledge representation model encodes knowledge in the form of classes, relationships, properties, instances, and axioms [12]. Our work is inspired by COviD-19 Ontology (codo), and we have taken their work forward with some new features and new dimensions that cover all aspects of COVID-19. We have significantly expanded the capabilities of the codo ontology model made for COVID-19 cases and patients information, i.e., changes to classes, properties, relations to be extensible. We are giving annotations for ontology concepts that have already been produced but neither annotated nor labeled.

4.1 Design Methodology

From the survey of the literature one can find that there are several ontology development methodology (ODM) that carried set of activities to create ontologies. In ODM, Knowledge acquisition, integration, and alignment drive the speed of building ontology and these come with the risk of redundancy, consistency, and conflicts. Invoking the existing ODM, we opted for a mixed approach of Diligent [15] and Methontology [16] to develop the CovidO. The general procedure, and functions to obtain CovidO describe in four phases as DataToMetadata (D2MD): **1. Ontology Requirement Specification (ORs):** In the ORs process, set the goal for ontology development and study the feasibility of the scope of ontology. We examine heterogeneous data sources given in the introduction to building ORs for CovidO and state-of-the-art requirement analysis according to the dimension of CovidO given in Sect. 3. ORs document help to design competency questions (given in Table 4) to define CovidO's scope. CovidO domain knowledge should be organized as a meaningful ORs model at the knowledge level. After gathering sufficient information and ORs, we create a CovidO conceptual model that describes the problem and its solution. **2. Ontology Development Phase (ODP):** ODP phase is the 2nd phase in the methodology which is responsible for deciding the Ontology Architecture, Designing Conceptual Map, Encoding in the OWL format. We have assumed that initial ontology is already constructed in the form of codo. It saves initial ontology development time and will help expand ontology in a distributed manner with different stakeholders' objectives as a dimensions. According to extendable objectives make the changes in the ontology locally and then revise updates to satisfy consistency and verify scope. The analysis attempts to find the similarities in changing requests and users' ontology. The study looks for commonalities between the ontologies of changing requests and users. Instead, decide what changes should be made to construct conceptual design. Once have a conceptual design, then implementation can be done in two intermediate ways: formalization and ontology design. In formalization, to transform the conceptual model into a formal or semi-compatible model. In the ontology design, to make ontology use some ontology editor like protégé with sup-

ported formal language. **3. Validation & Evaluation:** Validation step is performed to validate the ontology based on different criteria like content validity, application-based analysis, etc. **4. Deployment:** In this phase makes the ontology to reuse and available for further use. Where the publication of the ontology is on the cloud or public portal with take care of maintenance and proper updates.

4.2 *New Features*

CovidO is formed by reusing existing ontologies adding new classes and properties to cover all the dimensions listed in Sect. 3. Table 2 represents the core classes of CovidO with referral namespace. We have divided the Statistics class of codo into two parts, codo: Statistics class and CovidO: Resource. Statistics class represents the actual cases information, and Resource class represents the resources utilized for covid-19. Similarly, other classes of codo like Status, Symptom, Disease, and CovidTestingFacility have been added according to dimensions. Table 3 describes some new relationships between the classes of CovidO that were not present in other Covid ontologies. Currently, CovidO contains 175 classes, 169 relationship types, and goes to evaluation. We applied the Pellet Reasoner to verify that CovidO is consistent.

The top-level class structure diagram and relationships between core concepts are shown in Fig. 1. There are many types of relationships between concepts. The content ensures the semantic consistency of relationships between concepts and facilitates logical axioms and reasoning definitions.

4.3 *CovidO Scope*

Competency Questions (CQs) play an essential role in the ontology development lifecycle, representing the ontology requirements. Some CQs have been formalized through COVID-19 data sources described in the introduction section. We could structure and expand the conceptual modeling design of codo through CQs to obtain the CovidO. The CQs have broad coverage on COVID-19 data, So we are trying to map it into seven dimensions of COVID to mitigate it and cover all possible aspects of the pandemic so far. Table 4 represents some competency questions with the respective dimension that CovidO is expected to answer.

4.4 *Reusing the Ontology Concept*

According to [17], there are two perspectives to reuse ontology: (1) assembling, extending, specializing, and adapting other ontologies that are components of the

Table 2 Core entities and namespaces considered in CovidO and their description

S. No.	Core entities	Namespace	Description
1.	dbo:Continent	dbpedia.org	A continent is any of several large landmasses
2.	dbo:Town	dbpedia.org	A town is a human settlement
3.	covid0:TownWiseStatistics	w3id.org/CovidO	Statistical information of covid case in a town
4.	codo:Statistics	w3id.org/codo	update, information about the infected cases
5.	covid0:Status	w3id.org/CovidO	provide the status as recovered, hospitalized, deceased
6.	covid0:TestResult	w3id.org/CovidO	Provide test result positive or negative
7.	covid0:Resources	w3id.org/CovidO	Details of Covid-19 related resources
8.	covid0:ClinicalMeasurement	w3id.org/CovidO	Describe clinical measurement required for covid testing, i.e., blood test, oxygen saturation, temperature etc.
9.	covid19IBO:Impact	semanticweb.org/archana/ontologies/2021/5/untitled-ontology-6	Covid-19 impact on business verticals
10.	covid0:Covid-19	w3id.org/CovidO	Information about the Covid-19
11.	covid0:Covid-19Study	w3id.org/covid0	Covid-19 study on application, news, article, global research, origin of covid etc.
12.	covid0:Covid19FungalInfection	w3id.org/CovidO	Fungal infections in covid positive patients and different variants
13.	covid0:LaboratoryTest	w3id.org/CovidO	Laboratory Test for covid specific
14.	covid0:Treatment	w3id.org/CovidO	A planned operation based on data for the delivery of health care
15.	covid0:Vaccine	w3id.org/CovidO	Antigenic chemicals are used in preparations to stimulate the immune system and elicit an immunological response

final ontology, or (2) integrating other ontologies on the single concept that integrate all concept. We have used second approach, integration of other ontologies. The core concepts for CovidO are determined as prevention, vaccine, hospital facility, disease, infection, disorder, virus, SARS-Cov2, Coronavirus, agent, patient, disease, symptom, drug, treatment, organization, impact, host, diagnosis, statistics, place, etc. Which is already in use in other ontologies O1 to O12. CovidO integrates the concept with existing ontologies to pursue the basic principle of ontology implementation. Figure 2 represents the terms inherited by CovidO from available schema. Different

Table 3 Some new relations (core object properties) and related domains and range, which have been considered in CovidO

S. No.	Object properties	Namespace	Domain	Range
1.	isPartOf	w3id.org/covido	dbo:Continent	dbo:Continent
2.	hasResources	w3id.org/covido	codu:Covid-19DedicatedFacility	covido:Resources
3.	hasClinicalFinding	w3id.org/covido	schema:MedicalClinic	covido:ClinicalFinding
4.	isVaccinated	w3id.org/covido	foaf:Person	dbo:Vaccine
5.	hasImpactOn	w3id.org/covido	covido:Covid-19	xmlns:Organization
6.	hasPrevention	w3id.org/covido	covido:Mndate	covido:Prevention
7.	hasDecision	w3id.org/covido	foaf:Person	covido:Decison
8.	hasImpactOn	w3id.org/covido	xmlns:Organization	xmlns:Organization
9.	mandatissued	w3id.org/covido	covido:PublicAuthority	covido:Mndate
10.	hasRiskFactor	w3id.org/covido	xmlns:Organization	covido:RiskFacotr
11.	hasGovern	w3id.org/covido	covido:PublicAuthority	xmlns:Organization

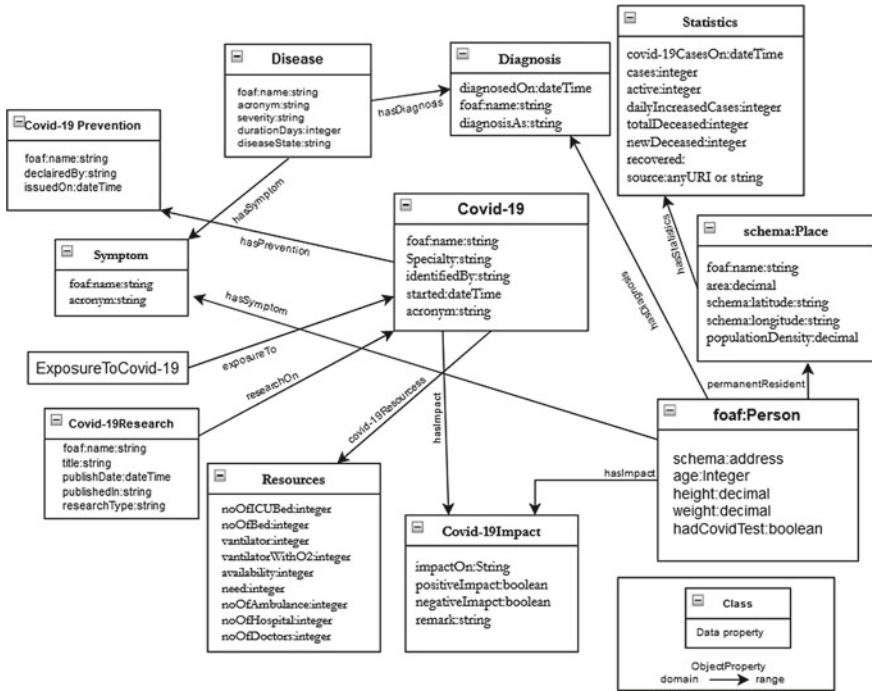


Fig. 1 Class structure diagram of the top levels of the core class hierarchy in CovidO

Table 4 CovidO competency questions excerpt

Dimension	Q. No.	Competency questions
D1	CQ1.	When did country c have the first COVID-19 case?
	CQ2.	How many COVID-19 cases have been found at any (location) place l on date t?
	CQ3.	How many patients died in Continent r on date t?
D2	CQ4.	What is the travel history of a patient p?
	CQ5.	How many patients traveled from / to Continent r on date t?
	CQ6.	List all the patients between age 18 and 30
D3	CQ7.	What are the different variants of COVID-19?
	CQ8.	What are the prevention vaccines for COVID-19?
	CQ9.	What are the drugs for COVID-19 treatment?
	CQ10.	What are the clinical measurements of a COVID-19 patient p?
D4	CQ11.	What are the health care facilities for COVID-19 in a place l?
	CQ12.	How many ICU beds are in a hospital h at a place p?
	CQ13.	How many ambulances are available in a hospital h on date t?
D5	CQ14.	What are the business verticals on which COVID-19 has a positive impact?
	CQ15.	Provide a list of all organizations at high risk?
D6	CQ16.	When was the first lock down announced in a country c?
	CQ17.	What is the exposure of COVID-19 spread?
	CQ18.	What is prevention advice issued by state public authority for covid-19?
D7	CQ19.	What are the research articles published for COVID-19 in a month?
	CQ20.	What are the news headlines and their sources of COVID-19 on date t?

colors define each schema. The solid and the dotted line show immediate and remote child-parent relations between classes. Most inherited entities belong to Schema, BFO, and CODO, while Kg4Grug, SYMP, and VO are the least inherited ontologies.

We use a permanent URL service, w3id.org, that makes it independent of external ontologies. We have defined a new namespace convention <https://w3id.org/CovidO> for all classes used in CovidO, with the prefix covido (the entry registered at <http://prefix.cc>). A common concept is integrated into a single concept that unifies them all. The concept selection is based on the best fit for the scope and dimension of CovidO. For determination of scope, competency questions were created and are publicly available via GitHub <https://github.com/sumitsnit/CovidO>.

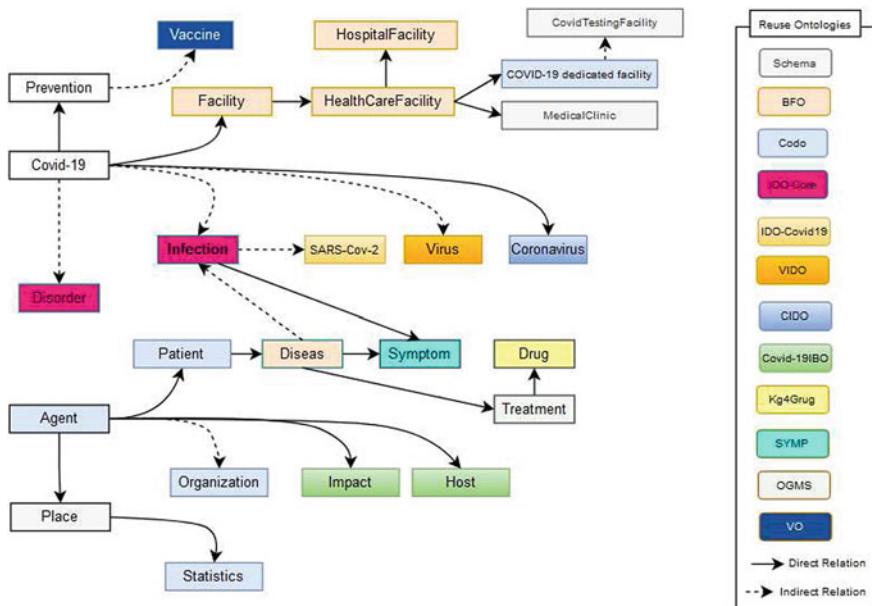


Fig. 2 Reuse core concepts of CovidO ontology extracted from the related ontologies

5 Evaluation

We evaluated CovidO in two ways: using SPARQL query and OOPS! Pitfall Scanner. SPARQL query describes the accessibility of elements and OOPS! Pitfall represents the RDF quality of CovidO.

5.1 SPARQL Query Evaluation

A set of competency questions over CovidO is given in Table 4. For scope evaluation of the CovidO, answers to the SPARQL [18] queries built on the questions from Table 4 are considered. CovidO is schema-based ontology, so schema-level SPARQL query is allowed till now. Once the ordered representation is done as user requirement from the coronavirus distributed and heterogeneous data sources, CovidO instance-level questions are permitted to answer. We evaluate CovidO on schema-level competency questions based on the D3 dimension. The SPARQL queries and their results are shown in Table 5. The prefix and their namespaces used for SPARQL queries are shown in Table 6.

Table 5 CovidO schema-level queries and their SPARQL queries and results obtained

S. No.	Question	SPARQL query	Result
CQ8.	What are the prevention vaccines for Covid-19?	SELECT ?CovidVaccine WHERE { ?Vaccine rdfs:subClassOf covidov:Vaccine. ?CovidVaccine rdfs:subClassOf ?Vaccine. }	Novavax, Covaxin, Moderna, SputnikLite, ZyduScadila, Covishield, Corbevax, SputnikV,
CQ10.	What are the clinical measurements of a COVID-19 patient?	SELECT * WHERE ?Vaccine rdfs:subClassOf covidov:ClinicalMeasurement.	Temperature, OxygenSaturation, BloodMeasurement, CardioVascular, BloodChemMeasurement, Diabetes

Table 6 CovidO prefixes table

Prefix	Namespace
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns
owl:	http://www.w3.org/2002/07/owl
rdfs:	http://www.w3.org/2000/01/rdf-schema
xsd:	http://www.w3.org/2001/XMLSchema
foaf:	http://xmlns.com/foaf/0.1/
ndf:	https://www.semintelligence.org/ns/whonto
xmlns:	http://xmlns.com/foaf/0.1/
dbo:	http://dbpedia.org/ontology/
schema:	https://schema.org/
covidov:	http://www.w3id.org/covidov/

5.2 OOPS! Pitfall Evaluation

We have used OOPS! Pitfall scanner [19] to examine CovidO. OOPS! Pitfall ontology diagnosis online tool detects 40 different types of pitfalls in OWL ontologies, including semantic and structural checks and best practices verification. OOPS! Pitfall scanner has two components, Pitfall Scanner, and Suggestion Scanner. Pitfall Scanner checked the ontology syntax, and Pellet reasoner has analyzed the logical consistency of ontology. Whereas the Suggestion scanner has thrown some suggestions for possible errors of ontology elements. We resolved the problems reported by OOPS! in the CovidO and updated CovidO.

6 Conclusion

This work presents Coronavirus Disease Ontology (CovidO): (1) Describes complete knowledge of COVID-19; (2) Maps existing ontologies and create standard metadata to understand and share COVID-19 knowledge. (3) Acts as a vocabulary for researchers, engineers, and developers to find the commonly used COVID-19 gesticulation for particular accordance's, and to use the scope and dynamics of a specific gesture. The intention and scope of the CovidO can be summarized in seven dimensions as discussed above. As future work, some possible use case will be added to make it more integrating linguistics. In addition, we intend to release and deploy the CovidO RESTful service in the Cloud and API clients to the leading coronavirus annotation [20]

References

1. World health organization. <https://covid19.who.int/>. Accessed 10 Feb 2022
2. Dividino R, Soares A, Matwin S, Isenor AW, Webb S, Brousseau M (2018) Semantic integration of real-time heterogeneous data streams for ocean-related decision making. Defense Research and Development Canada = Recherche et développement pour la ..
3. Osborne F, Motta E (2015) Klink-2: integrating multiple web sources to generate semantic topic networks. In: International semantic web conference. Springer, pp 408–424
4. Ding L, Kolari P, Ding Z, Avancha S (2007) Using ontologies in the semantic web: a survey. In: Ontologies. Springer, pp 79–113
5. Beverley J, Babcock S, Cowell L, Smith B (2021) The covid 19 infectious disease ontology
6. He Y, Yu H, Ong E, Wang Y, Liu Y, Huffman A, Huang HH, John B, Lin AY, Arabandi S et al (2021) Cido: the community-based coronavirus infectious disease ontology
7. Babcock S, Beverley J, Cowell LG, Smithd B (2020) The infectious disease ontology
8. Dutta B, DeBellis M (2020) Codo: an ontology for collection and analysis of covid-19 data. [arXiv:2009.01210](https://arxiv.org/abs/2009.01210)
9. de Lusignan S, Liyanage H, McGagh D, Jani BD, Bauwens J, Byford R, Evans D, Fahey T, Greenhalgh T, Jones N et al (2020) In-pandemic development of an application ontology for covid-19 surveillance in a primary care sentinel network. *JMIR Publ Health Surveill*
10. Patel A, Debnath NC, Mishra AK, Jain S (2021) Covid19-ibo: a covid-19 impact on Indian banking ontology along with an efficient schema matching approach. *New Gener Comput* 39(3):647–676
11. Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, Shefchek KA, Good BM, Balhoff JP, Fontana T et al (2021) Kg-covid-19: a framework to produce customized knowledge graphs for covid-19 response. *Patterns* 2(1):100155
12. Badenes-Olmedo C, Chaves-Fraga D, Poveda-Villalón M, Iglesias-Molina A, Calleja P, Bernardos S, Martín-Chozas P, Fernández-Izquierdo A, Amador-Domínguez E, Espinoza-Arias P et al (2020) Drugs4covid: Drug-driven knowledge exploitation based on scientific publications. [arXiv:2012.01953](https://arxiv.org/abs/2012.01953)
13. Bonino L (2020) Who covid-19 rapid version crf semantic data model. *BioPortal*
14. Qundus JA, Schäfermeier R, Karam N, Peikert S, Paschke A (2021) Roc: an ontology for country responses towards covid-19. [arXiv:2104.07345](https://arxiv.org/abs/2104.07345)
15. Pinto HS, Staab S, Tempich C (2004) Diligent: towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies. In: *ECAI*, vol 16. Citeseer, p 393

16. Fernández-López M, Gómez-Pérez A, Juristo N (1997) Methontology: from ontological art towards ontological engineering
17. Pinto HS, Martins J (2000) Reusing ontologies. In: AAAI 2000 spring symposium on bringing knowledge to business processes, vol 2. AAAI, Karlsruhe, Germany, p 7
18. Seaborne A, Manjunath G, Bizer C, Breslin J, Das S, Davis I, Harris S, Idehen K, Corby O, Kjernsmo K et al (2008) Sparql/update: a language for updating rdf graphs. W3c Member Submission 15
19. Gómez-Pérez A, Oops!(ontology pitfall scanner!): supporting ontology evaluation on-line
20. Sharma S, Jain S (2021) Comprehensive study of semantic annotation: variant and praxis

An Ethnolinguistic Research Agenda for Intelligent Autonomous Systems



Bharat K. Bhargava, Sarika Jain, and Abhisek Sharma

Abstract Depicting intelligence by autonomous systems by extracting patterns from heterogeneous data streams poses significant computational and analytical challenges in real-time. We need technologies that can provide multilingual services while respecting each culture and its values and ethics. In many cases, language is not the only reason for misunderstandings and unsatisfactory services, but sometimes cultural properties also play an equivalent role. Current technologies can prominently understand and work with multilingual information; what is required is to embed the cultural constructs and properties during automation. Ethnolinguistics is precisely talking about that, i.e., the relationship between language and culture. The semantic representations of language and cultural differences can prove expressive enough in making judgements when decisions cross geographic boundaries. The vision is to utilize a hybrid of both the syntactic (deep learning) and the symbolic (knowledge graphs) approaches to modeling behaviour. This paper tries to lay forward the current status and future research directions for the vision of intertwining ethnolinguistics and knowledge bases for the intelligent autonomous systems (IAS).

1 Introduction

One of the dreams of AI has been to achieve Artificial General Intelligence, i.e., rather than being best at something, the machines become autonomous agents. This is impossible without developing cognitive skills. Consider use case scenarios where we wish to emulate a human, determine suspicious behaviour, do cross-country

B. K. Bhargava
Purdue University, West Lafayette, IN, USA
e-mail: bbshail@purdue.edu

S. Jain · A. Sharma (✉)
National Institute of Technology Kurukshetra, Kurukshetra, India
e-mail: abhisek_61900048@nitkr.ac.in

S. Jain
e-mail: jasarika@nitkr.ac.in

business, or move between geographic boundaries; the most important requirement is to represent and communicate context. The machines must represent the relationships and the meaning of data for sensible prediction, decision-making, conflict resolution, and all sorts of reasoning tasks in the real world.

Current AI systems lack explainability and hence trust. We propose a semantic framework to make computers understand cultural constructs and utilize the same in depicting intelligence. As knowledge graphs give AI the required context, we take knowledge graphs as the default data model for contextual understanding in cross-country intelligent applications. The semantic representations of language and cultural differences can prove expressive enough in making judgements when decisions cross geographic boundaries. The vision is to utilize a hybrid of both the syntactic (deep learning) and the symbolic (knowledge graphs) approaches to modeling behaviour. The symbols act as lingua-franca between humans and the statistical AI models. A culturally rich knowledge graph will be built by utilizing the natural language processing techniques and machine learning. Expert rules will be codified and services in the form of APIs will be provided for proper accessibility of the knowledge graph. The developed semantic framework is planned to be available on the linked open data cloud so that AI applications can embrace the developed technology for social benefit.

This paper outlines the context and research questions behind ethnolinguistics, a study of language and culture, and how they relate to each other. We start from “Intelligent Knowledge Bases” [1] and go all the way to “Intelligent Ethnolinguistic Knowledge Bases” which hold all the details needed to provide effective services and build applications in a culturally rich manner. The scientists are already working upon intelligent knowledge bases involving complex underlying semantics and pragmatics that are capable enough to put data in context. These knowledge bases can represent multilingual information also [2, 3]. After exploring the available literature on ethnolinguistics, we found that culture speaks the language, i.e., both are intertwined. The culture teaches us how to behave and interact with others in society, and language facilitates this social interaction [4].

The preservation of the cultural diversities/similarities between different communities and the exchange/transmission of information/knowledge dealing with the cultural aspects of a region or a community demands a technology that allows enough flexibility and specializes in cataloging metadata semantically. While a lot of effort has been put into developing such tools for English content, relatively little effort has been put into localizing them. Localization is a more comprehensive process than only translating text from one language to another; it addresses cultural and non-textual components and linguistic issues when adapting a product or service for another country or locale. (Localization = Translation + Culturalization). Localization differs from translation in that it focuses not only on linguistic adaptation but on a regional/cultural adaptation of contents.

In the rest of this paper, we outline these aspects in much more detail. Section 2 defines intelligent knowledge bases and Intelligent Ethnolinguistic Knowledge Bases along with their properties and some use cases. The first part of Sect. 3 then digs a bit deeper into the collective literature of ethnolinguistics and knowledge base. In

part 2 of Sect. 3, we have devised some questions that need to be answered to realize ethnolinguistics capabilities by computers.

2 Overview

When can interaction be categorized as effective? To answer this question, we should be aware of the fact that there are various ways to express oneself, and there are multiple factors/details that need to be content with while communicating. To make communication effective, we should be acquainted with the culture and language of each other. The language part is being covered in many ways, but the cultural part (the representation and accessibility of it) is still in its starting phase.

As technology progresses, we need ways to make computers understand the culture. We as a community have to an extent covered the ways and methodologies to make language understandable by computers. And the reason we should work on cultural constructs is because culture and language are an integral part of each other. To do so, we will be using knowledge bases. The reason behind using knowledge bases along with some details of ethnolinguistics are stated in the subsections.

2.1 *Intelligent Knowledge Bases*

The ability to perform inference on stored information is possible through “Intelligent Knowledge Bases.” In an intelligent knowledge base, the information is stored in the form of a graph that allows hidden (or previously unknown) connections to be recognized (inferencing) and enrichment of the knowledge base.

Representation and Storage: the available options are Triple Store (Ontology), RDBMS and NoSQL databases (especially the graph databases like Allegro Graph¹). In our case, we are using ontology. The advantage of ontology over the RDBMS database is that ontology uses open-world assumption (OWA). As there are various domains where a clear result can’t be provided like medical, in medical we can’t certainly say anything without testing (or getting concrete proofs) if the information is not present. In the case of RDBMS, if the information is not present, it will return that the patient doesn’t have a particular disease, but that can’t be certainly said without proper testing, so in cases like these, ontologies are the way to go as it goes with OWA and allows reasoning to infer new information. Ontologies have formalized semantics which allows reasoners to infer new information without the need of manually writing it.

¹ <https://allegrograph.com/>.

Applications of Intelligent Knowledge base

The applications of intelligent knowledge bases can be anything that requires capabilities from storing and accessing information to depict details (which involves semantics). Applications such as:

- (a) Semantic web
- (b) AI-based content recommender
- (c) Wearable devices
- (d) Autonomous vehicles
- (e) Search engines

Use Case: Business Environment

1. The framework will utilize the semantic tools and applications to fetch information from subject matter experts, surveys, and descriptive textual data/documentation to populate the domain-specific semantic knowledge graph. Human’s personal knowledge graph will be developed (Human’s cultural background, local, country, etc.). Sources for these personal details will be like social networking profiles of the person, organizational profiles, etc.
2. Knowledge graph of business environment domain will be made.

Take the organizational hierarchy below where A is the type manager, B is three types of team leaders, and C is three types of team members (Fig. 1).

Say the organization has three people of different cultures at the position of manager. The B1 type of team leader also has three instances, and so on.

A—a1, a2, a3 (3 instances of A of different cultures).

B1—b11, b12, b13, etc.

B2—b21, b22, b23, etc.

Now say in different scenarios, a1 interacts with b21, a1 interacts with b22, a1 interacts with b11, a2 interacts with b12, a2 interacts with b21, and so on. All of these above-stated scenarios involve a pair of people communicating with each other, and there are multiple possibilities of a combination of cultural differences that can occur. Combinations like.

- (a) **Huge cultural distance:** First-person in the above interaction is from Scandinavian culture, and the second person is from Asia.

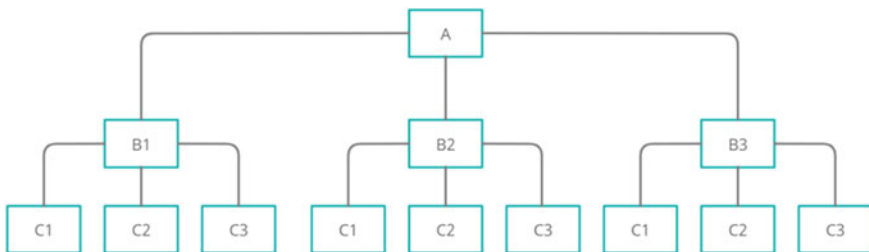


Fig. 1 Organizational Member’s Hierarchical Representation

- (b) **Moderate cultural distance:** In this category, the conversations between people from countries like Germany and UK.
- (c) **Low cultural distance:** In this, the two persons have little cultural distance. That means that one of them is from a country like India and the other is from a country like China. Where one person is from India and the other is from China.

The above categories are differentiated based on the number of cultural traits that the cultures have in common. A high number of common traits means that the cultures are similar to each other and vice versa for the cultures with a smaller number of common traits.

2.2 *Intelligent Ethnolinguistic Knowledge Bases*

Ethnolinguistics means the relationship between language and culture of people who uses that language. As the name suggests, intelligent ethnolinguistic knowledge bases means the knowledge base containing ethnolinguistic information and core concepts of any domain. To represent ethnolinguistics in a way that is understandable by the computers, knowledge bases are preferred as they allow keeping the semantic and context of data intact.

Culture and Language

There are six linguistic units (phonetics, phonology, morphology, syntax, semantics, and pragmatics). All these components are collectively termed as the ethnolinguistic factors and heavily influence the human's perception and behaviour. Ethnolinguistics is a subfield on linguistics that deals with the relationship between culture and language and how it changes the ways people from different cultures perceive the world around them [5]. Semantics has been explored in recent years by the Linked Data community. The notion of context (pragmatics) is still a research subject; moreover, most of these activities are primarily tailored to the English language. Pragmatics is about using language in social contexts and provides the ability to understand the speaker's intention. The general semantics and pragmatics are highly influenced by the cultural background and the non-textual factors.

Ethnolinguistics is a multidisciplinary area of research that explores the relationship between languages, culture and conceptualizations. The ultimate aim of ethnolinguistics is to examine the relationship between language, culture, and conceptualizations.

Intercultural Communication

When it comes to communication, then every factor that we discussed so far in ethnolinguistics. Below are some characteristics of intercultural communication and how it gets affected.

- (a) Gumperz defines contextualization cues as “verbal and non-verbal meta-linguistic signs that serve to retrieve the context-bound presuppositions in terms of which component messages are interpreted.”
- (b) Cultural conceptualizations provide a basis for constructing, interpreting, and negotiating intercultural meanings.
- (c) “Cross-cultural variation at the conceptual level calls for a strongly meaning-oriented and interpretive approach to the study of intercultural communication,” and that is what Cultural Linguistics has to offer.
- (d) Through various examples, intercultural communication needs to have a common ground where people from different cultures are sensitive to and aware of cultural differences.

3 Background and Agenda

3.1 Literature

If people use cultural knowledge/understanding to interpret human speech, then the computer should too. To increase its efficiency in interpreting humans and their languages. We have language-specific information in a semantically rich format. These representations of language can be constructed through the use of models like lemon [6]. Many knowledge bases like Yago, Wikidata, DBPedia, etc., supports/contain multilingual information. But cultural information is still missing in all of them. We will perform tasks like interpreting human speech, texts (even texts that are written years ago), etc., more prominently through information-related cultural constructs.

Language is a part of the culture. Researchers have stated that without language, it is challenging for culture to develop and prosper. It is also to be noted that culture is responsible for a variety of changes in the language.

Both of them are bound in a vicious cycle, as also stated by [7]:

- Language reflects the culture and is shaped and influenced by the culture.
- In other words, “A language is a part of a culture, and a culture is a part of a language.”

“It is likely that all native knowledge of language and culture belongs to cultural schemas and the living of culture and the speaking of language consist of schemas in action” [8].

The set of characteristics that makes someone multicultural are context (like cultural heritage, history, interpersonal relations, etc.), skills and abilities, acculturation process (how a person has acquired multicultural properties), identification, and cognition.

Currently available technologies are capable enough to do tasks from building an ontology [9] to creating a knowledge graph with every information related to concepts in it. We already have knowledge bases that contain language-specific information.

Some ontologies are available that are developed with cultural constructs as part of them (such as Jean Petit et al. 2017), but these ontologies contain problems and errors.

Pheto et al. [10] has developed an ontology formalizing all the required aspects of culture contributing to the representation of cultural knowledge. The ontology has been written in the context of Botswana, which is a small African country and is likely to be generic enough for the global context. This cultural ontology will aid in better machine understanding and will benefit the sharing, reuse, and portability of knowledge across heterogeneous platforms.

Degl’Innocenti [11] states that the Referential Space Model (Vector Spaces) tends to provide more culturally sensitive and satisfactory results in calculating the semantic relationships. Gromann [12] mentions that the Neural Language Models (NLMs) (embeddings) learn implicit semantic representations of sequences on their hidden layer(s), resulting in a dense real-valued vector for each entity. The author proved that the extensions of ontology-lexicon and ontology-terminology models injected into NLMs are promising for context-sensitive tasks.

Robinson [13], has talked about how Nordic society citizens are skeptical about trusting AI and related technologies. To build people’s trust, authors have suggested making national policies on AI by upholding cultural values and personal rights. Along with trust, people should know how the technology works and see what is being collected about them and how it is being used (i.e., transparency and openness). There has been a lot of work for preserving cultural and historical knowledge and monuments. The work by Eckert et al. [14] is one of them providing a knowledge base for Jewish culture and history. It is developed based on multilingual information from general-purpose knowledge bases and encyclopedias. All in all, they are developing a single hub for contextualized knowledge by compiling a subset of data from various sources. Yang et al. [15] have discussed how online services generate data quickly, and there are cultural correlations in them. They have proposed architecture for handling and processing of cultural knowledge from these services. Currently, the architecture and their work are centered on working with Chinese sources and developing systems based on their insights.

Scope: Many popular and widely used knowledge bases such as DBpedia,² Wikidata,³ YAGO,⁴ etc., also contain some part of cultural information along with vast knowledge about almost everything in the world. But their main focus is knowledge in general, and they don’t put much effort into keeping cultural information intact with everything else and need a lot of work to embed cultural knowledge with previously added concepts when cultural constructs were not defined. In most systems, the focus is rather narrow as they want to build a system or knowledge base for some specific country or group. A broader consideration of culture is necessary. The knowledge bases that are already available are rather general and unfathomable in size, which makes them harder to process and use. These systems are working with various NLP

² www.dbpedia.org.

³ www.wikidata.org.

⁴ yago-knowledge.org.

techniques and using deep learning techniques like recurrent neural networks. But the approaches are contextually unaware as they lack contextual knowledge, which will be available through contextualized culturally rich knowledge graphs.

3.2 Agenda

This review of the ethnolinguistic adaptation of intelligent knowledge bases provides the background for language with cultural constructs. We have discussed and examined the current state of cultural contextualization and use. We have an ambitious agenda informed by the urgent need to develop understanding in an area of research that has been under-developed hitherto. Our research questions are:

(1) **How culture and language connects**

In the literature, we found that culture and language are intertwined, but what exact change a specific attribute makes and how it can be measured in quantifiable terms still needs to be answered. In this sense, our first research question is (RQ1) **“What advantages cultural knowledge will provide if referred at the time of interpreting language?”**

(2) **Culture and understanding**

In continuation of the first question that talks about a culture or a group in a collective sense, our second question is focused more on studying the effects culture has on an individual. Our second research question is (RQ2) **“Does culture play a role in understanding each other?/How does culture affect the person’s perception of the world?”**

(3) **Culture and language for computers**

Several NLP tools already exist that work with the basic concepts of linguistics except for semantics and pragmatics. The question then arises is that how computers can use culture and its information to increase the efficiency of tasks that a computer can already perform, and how a computer can become capable of performing some tasks because of cultural information. In this light, the third research question is (RQ3) **“How computers can make use of cultural constructs to better understand human language/speech?/How systems will be able to understand better if we add culture?”**

(4) **Available Technologies**

After talking about what cultural constructs can offer to computer systems, now we need to investigate the capabilities of current technologies in realizing those benefits that we as a community have visualized.

(5) **Determining the set of tools that can support**

Keeping RQ4 (a & b) in mind, we have to look into how and what combination of available tools can be used to provide culturally rich knowledge interpretation and services associated with it. Our fifth question is (RQ5) **“What combination**

of tools can be used to support culturally rich services?"/How to leverage all developed in 4.

(6) The current state of technological capabilities

We have some possible ways and results from some attempts at making culturally rich ontologies. Still, they are incomplete and contain errors because of which it is unclear that (RQ6a) **How much current services provide a proper depiction of human sentences?** And (RQ6b) **How to improve the effectiveness of existing services?**

4 Discussion and Solution

Devising an intelligent autonomous system (IAS) to achieve artificial general intelligence moves focus to the type of representations that are meaningful, structured, and not resource-hungry. Figure 2 depicts the framework for the contextualization of culture for robust and intelligent applications.

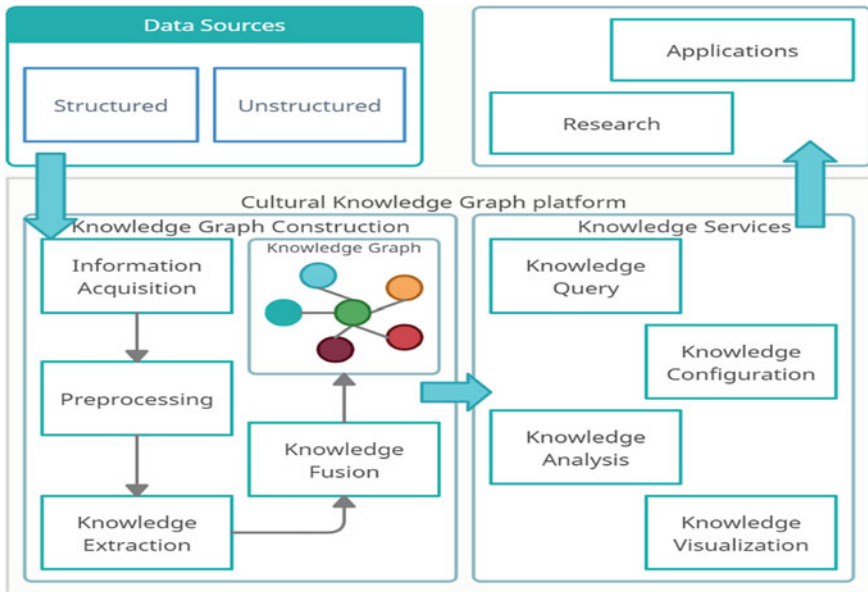


Fig. 2 Contextualization of Cultural Knowledge

4.1 Knowledge Graph Construction

First, we need all the information we can assemble about culture and its dimensions in a structured form (see Fig. 3). The cultural dimensions given by Hofstede [16] and scales given by Meyer [17] are termed as cultural properties here without loss of generality. They include Power Distance, Uncertainty Avoidance, Individualism/Collectivism, Masculinity/Femininity, Long/Short Term Orientation, and Indulgence/Restraint. A contextualized culture knowledge graph will be built from multiple sources.

- a. **Structured Sources:** These are the ones that represent the information they have in some structured format such as triple stores, RDF, etc. Some of these sources are DBpedia, YAGO, and Wikidata. These information extraction tools are APIs for accessing information from relational databases and SPARQL endpoints for knowledge extraction from various standard knowledge repositories. The tasks involved include entity recognition, relationship extraction, and disambiguation.
- b. **Unstructured Sources:** We move towards unstructured sources like web pages, articles, and books for the concepts that are not available from structured sources. We will be using a collection of tools such as Open Information Extraction (OIE), regular expression-based methods, etc., for information extraction and construct structured representation of the unstructured data on these sources. The process of extracting information from unstructured sources and putting them into RDF or n-triples format have 2 steps:
 1. **Data Collection:** For this, web crawlers will be developed and used, to crawl the web to find data explaining cultural properties, its connection with humans and their behaviour, and how it has evolved to the version it is now.
 2. **Data Preprocessing:** This involves filtering and cleaning the data by removing all the stop words, links, etc. Secondly, named entities will be recognized in the text and the relationship between these entities will be determined. Then we will disambiguate the entities and relations to decide

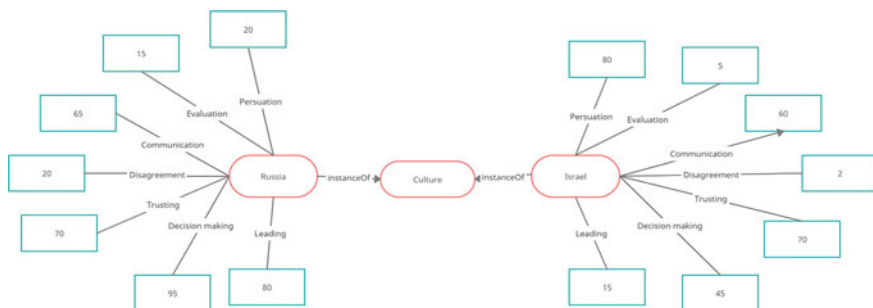


Fig. 3 Sample instance from the knowledge graph

the right position where that entity belongs. And along the above set of tasks, we will keep finding and removing redundancies.

After the construction of knowledge graph, we proceed to knowledge graph enrichment. As knowledge graph enrichment is an iterative process we will be using entity and instance matching tools to keep redundancy in check.

4.2 Rules

These are the set of rules that states that a set of properties devise particular behaviour in a specific culture. These are the general rules that are useful for every use case there is which works with cultural constructs or is affected by them.

Sample Rule:

```
R1 (LowPowerDistance(?country)^Collectivism(?country)
^Indulgence(?country)... ) → preferRightToEquality(?country)
```

4.3 Use Cases

1. Detecting Suspicious Behaviour

There are various autonomous systems available and are at work that are being used to detect suspicious behaviour (like [18, 19, 20]). Through using cultural knowledge, these autonomous systems can get more insights into the task at hand.

- a. **Detect Person's Locale:** We first need to identify the culture that a person belongs. For this, a combination of techniques can be used, such as using facial features or using a unique identification database of the government (after getting permission from the government) to get details on where the person is from and what culture s/he belongs. Facial recognition feature, which is a part of most of the autonomous systems in addition with the cultural information of facial features is the novelty involved.
- b. **Behaviour Detection Specific Rules:** In addition to the rules defined in the generalized cultural knowledge graph we will need specific rules for every application. These rules will be something that will define a combination of cultural constructs that can be used in the justification for the behaviours that are being categorized as suspicious.

2. Moving to Different Country

There are differences between cultures that one needs to be aware of when moving to a different country to keep him/herself out of trouble. Refer to Fig. 4, the rules constructed such as.

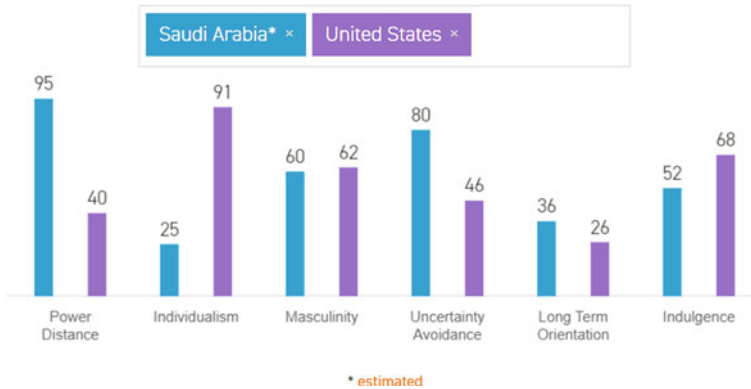


Fig. 4 Comparison of cultural dimensions of Saudi Arabia and the USA⁵

```
R2 (HighPowerDistance(?country) ^ HighUncertaintyAvoidance
(?country) ^ ...) → beCautious(?user)
```

will mean that if a person moves from the USA to Saudi Arabia s/he has to keep in mind that on many levels they have been cautious as there is less equality in the society (High power distance) and if some mishappening occurs they may not get much help from local people (as people in Saudi Arabia prefer avoiding confrontations).

5 Conclusion

Culture along with language plays an important role in effectively understanding the true meaning of presented information. Current systems are able to work with multilingual information but lacks in handling cultural variance. While answering the deviced quetions, we as a community can enhance the capabilities of the systems to handle cultural constructs and understand the semantics behind similar textual representation of information with difference in meaning dependent upon culture.

Declaration of Interest Statement The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

References

1. Jain S, Jain NK (2014) A globalized intelligent system. In: 2014 international conference on computing for sustainable global development (INDIACom). IEEE, pp 425–431

⁵ <https://www.hofstede-insights.com/country-comparison/saudi-arabia,the-usa/>.

2. Jain S, Kysliak A (2022) Language-agnostic knowledge representation for a truly multilingual semantic web. *Int J Inform Syst Model Design (IJISMD)* 13(1). IGI Global
3. Jain S, Chaudhary D, Jain NK (2011) Localization of EHCPRs system in the multilingual domain: an implementation. In: International conference on information systems for Indian languages. Springer, Berlin, Heidelberg, pp 314–316
4. Cronk L (2019) *That complex whole: culture and the evolution of human behavior*. Routledge
5. Wierzbicka A (1992) *Semantics, culture, and cognition: universal human concepts in culture-specific configurations*. Oxford University Press on Demand
6. McCrae J, Aguado-de-Cea G, Buitelaar P, Cimiano P, Declerck T, Gómez-Pérez A, Gracia J, Hollink L, Montiel-Ponsoda E, Spohr D, Wunner T (2012) Interchanging lexical resources on the semantic web. *Lang Resour Eval* 46(4):701–719
7. Jiang W (2000) The relationship between culture and language. *ELT J* 54(4):328–334
8. Sharifian F (2015) Language and culture: overview. *Routledge Handbook Lang Cult* 1:3–17
9. Noy NF, McGuinness DL (2001) Ontology development 101: a guide to creating your first ontology
10. Phefo OS, Kefitile N, Hlomani H (2015) Towards the cultural knowledge ontology. In: 2015 IEEE international conference on information reuse and integration. IEEE, pp 526–533
11. Degl'Innocenti D (2017) Multilingual keyphrase extraction and advanced localisation strategies
12. Gromann D (2020) Neural language models for the multilingual, transcultural, and multimodal Semantic Web. *Semantic Web*, (Preprint), pp 1–11
13. Robinson SC (2020) Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (A.I.). *Technol Soc* 63:101421
14. Eckert K, Dadvar M (2019) JudaicaLink: a knowledge base for Jewish culture and history. *Umanistica Digitale* 3(4)
15. Yang Y, Zhang G, Wang J, Ye S, Hu J (2017) Public cultural knowledge graph platform. In: 2017 IEEE 11th international conference on semantic computing (ICSC). IEEE, pp 322–327
16. Hofstede G (2011) Dimensionalizing cultures: The Hofstede model in context. *Online Read Psychol Cult* 2(1):2307–2919
17. Meyer E (2014) *The culture map: breaking through the invisible boundaries of global business*. Public Affairs
18. Renckens IR (2014) Automatic detection of suspicious behaviour. Master's thesis
19. Lee WK, Leong CF, Lai WK, Leow LK, Yap TH (2018) ArchCam: Real time expert system for suspicious behaviour detection in ATM site. *Expert Syst Appl* 109:12–24
20. Ouivirach K, Gharti S, Dailey MN (2013) Incremental behaviour modeling and suspicious activity detection. *Pattern Recogn* 46(3):671–680

QuantumRNG, A Random Number Generator Using One Qubit



Dara Ekanth, Bheemanathy Saketh Chandra, and Meena Belwal

Abstract This paper deals with the work done in generating truly random numbers using Quantum Computing. Since Quantum Computing is in its early stage of developing, we have Quantum Computers with only a few Qubits, so efficient use of those Qubits is necessary. We have developed an algorithm that generates n bit truly random numbers using only one Qubit and minimal resources. There exist many algorithms to generate random numbers, but they use more resources compared to our algorithm. In this paper, we are going to discuss about the algorithms to generate single random number and multiple random number. In this work, we have tested our algorithm with few other algorithms and our algorithms outperform other algorithms most of the time. The challenges we faced is, the Quantum Computers now available are prone to noise and they are not giving the proper results as expected, so with the advancement of technology, if the noise is reduced, we might get the expected results with no error.

Keywords Qubit · Quantum computing · Truly random numbers · Quantum particles · Qiskit · QuantumRNG

The original version of this chapter was revised: Author name “Bheemanathy Saketh Chandra” has been updated in the chapter. The correction to this chapter is available at https://doi.org/10.1007/978-981-19-7126-6_23.

D. Ekanth (✉) · B. S. Chandra · M. Belwal
Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India
e-mail: daraekanth3@gmail.com

B. S. Chandra
e-mail: b.sakethchandra9@gmail.com

M. Belwal
e-mail: b_meena@blr.amrita.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023, corrected publication 2023

S. Jain et al. (eds.), *Semantic Intelligence*, Lecture Notes in Electrical Engineering 964, https://doi.org/10.1007/978-981-19-7126-6_10

1 Introduction

Quantum Computing works by following the laws of Quantum Physics and Quantum Mechanics. By taking the advantage of the interference, superposition and entanglement the computation is made by Quantum particles [1]. The Quantum particles such as electrons and protons do not obey the laws of classical physics, since Quantum particles exhibit dual nature of particles a separate branch is evolved in physics i.e., Quantum Physics. Quantum computers can solve complex problems that cannot be solved by any classical supercomputers. With the evolution of technology, many are developing Quantum computers by taking the advantage of Quantum particles but the ultimate base for it is Quantum Physics. By definition, Quantum Mechanics is random so with the help of that randomness we can generate random numbers [2]. Some developed Quantum Computers using photons, Semi-Conductors, trapped ion Quantum computers etc. Truly random numbers are those which cannot be predicted by any means, and they won't have any patterns in their series [3]. Most of the classical random number libraries are not truly random they are called pseudo-random numbers because they consist of some patterns. The random number generating libraries such as NumPy and Random initially take a seed to generate some random numbers, the series is depending on the seed they are taking initiative. The truly random numbers won't depend on any initial seed value to generate the random numbers [4]. The main advantage of the truly random numbers is we can easily randomize data to generate OTPs, passwords and many more. With the help of these we can decrease the reuse of past OTPs or passwords, this in turn helps in less data breach [5, 6]. In recent applications, these random number generators are used in cellular communication for safe key hierarchy [7]. And these random numbers have a huge scope in many cryptographic applications like Diffie Helman key exchange and securing communications via sequence in a secure channel. We implemented our algorithm using Qiskit framework developed by IBM, and used IBM's Quantum Computers to implement and test our algorithm. The main idea of our algorithm is the superposition in Quantum Computing, this means a Quantum bit can exist in two different states at a given point of time and no one can able to predict in which state the Quantum bit existed until the Quantum bit collapse [8]. Using the Hadamard Gate we achieved the superposition state and we collapsed the Quantum bit using the measure gate. We used Borel Normality Test for testing our algorithm.

The remaining of this article is organized as follows:

Section 2 speaks about the work that has been done in the field of random numbers, Sect. 3 speaks about the design of the Quantum circuit to generate random numbers, Sect. 4 will demonstrate about the implementation of the algorithm, in Sect. 5 the algorithm will be tested with few other algorithms to check the quality of randomness, Sect. 6 will talk about results and discussions.

The key contribution of this work is:

- (1) Algorithm for generating truly random numbers.
- (2) Library-anyone can install and use it on the go.

2 Literature Survey

Many classical algorithms that we currently have to generate random numbers are not truly random numbers in fact they are called pseudo-random numbers [9], means there exists a sequence in generated numbers, because they initially depend on a seed value [10] to generate random numbers with a mathematical function generally. Truly random numbers are those which doesn't follow any sequence in generating random numbers. Random numbers have greater use in many cryptographic applications. We can test the numbers generated by a quantum computer is random or not by some algorithms [11]. Like any other Pseudo-random numbers, the random numbers generated by quantum computers cannot be replicated using any technique [12]. Nowadays random numbers are very helpful in generating OTPs, RECAPTCHA and passwords [13]. Incryptography, public and private keys play a major role in security. The key exchange algorithms are prone to get hacked by adversaries. By using Quantum key exchange, we can overcome those adversary attacks [14]. Security is the highest priority for finance, defence companies and user data protection, with the help of quantum communication data can be transferred securely through the quantum channels [15]. With the help of quantum physics, it is proven that by unlocking the potential of quantum properties many problems can be solved that cannot be solved classically [16]. Nowadays quantum computers are developed by following the principles of quantum physics [17]. By definition quantum physics is random [18, 19] so by taking the advantage of those randomnesses we can easily generate random numbers. The laws of quantum physics are far different from the laws of classical physics, quantum particles have some special properties like entanglement, super position and interference. With the help of super position of a quantum particle, we can achieve the randomness, because when a quantum particle is in super position anyone won't have any idea in which state it would be until the state collapse [20, 21]. With the help of Hadamard gate we can achieve super position state, so that we can make a quantum particle to be both states at a time, measure gate will be used to collapse the super position of the quantum particle [22, 23].

3 Design

For designing our circuit, we have used the Hadamard gate and measure gate. Our designed circuit can be seen in Figs. 1 and 2.

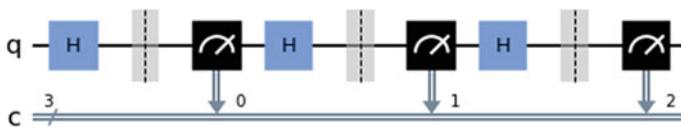


Fig. 1 Designed quantum circuit for generating the random number using one qubit

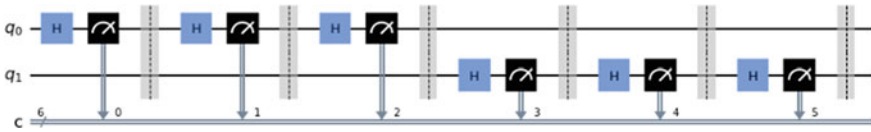


Fig. 2 Designed quantum circuit for generating the random numbers using multiple qubit for parallel numbers generation

Here in Fig. 1, we can observe that initially, we input a “q” value which is then passed to the Hadamard gate, since Hadamard gate forms the superposition of two quantum states we will get the probability of $|0\rangle$ and $|1\rangle$ as 50 per cent, which means both will have equal chances of getting as output, we will get the output by collapsing the superposition states, so to collapse the states we used measure gate. And we have developed an efficient Quantum circuit for minimizing the work overload on the qubits by distributing the work to a few more number of other available qubits see in Fig. 2.

4 Implementation

We implemented the design discussed in the previous section. We developed algorithms using the Qiskit framework. The implemented algorithms are mentioned in the below code snippets.

4.1 Algorithm for Single Qubit Number Generator

Input nbit: Number of digits in binary to generate random numbers.

qubit: Number of qubits that want to generate random numbers.

shots: provide a number of shots to run to generate random numbers.

backend: Provide the appropriate Quantum simulator/device.

1: **Initialization:** a Quantum circuit with an appropriate number of Quantum Registers and Classical Registers.

2: **begin**

3: for $j = 1, \dots, \text{nbit}$ do

4: Add Hadamard gate to the circuit

5: Measure the circuit with the measure gate

6: Increment the count by 1

7: **end for**

8: **end**

Execute the circuit on the backend with the appropriate number of shots.

Output: The output will be a bit string that has to be converted into an integer. The converted numbers will be the random numbers generated.

4.2 Algorithm for Multiple Qubit Parallel Numbers Generator

Input nbit: Number of digits in binary to generate random numbers;
qubit: Number of qubits want to generate random numbers;
shots: provide a number of shots to run to generate random numbers;
backend: Provide the appropriate Quantum simulator/device.

```

1: Initialization: a Quantum circuit with an appropriate number of Quantum Registers and Classical Registers.
2: begin
3: for i = 1,...,qubit do
4:   for j = 1,...,nbit do
5:     Add Hadamard gate to the circuit
6:     Measure the circuit with the Measure Gate
7:     Increment the count by 1
8:   end for
9: end for
10: end

```

Execute the circuit on the backend with the appropriate number of shots.

Output: The output will be a bit string that has to be converted into an integer. The converted numbers will be the random numbers generated.

5 Testing Our Algorithm

5.1 Borel Normality Test

The Borel normality test [24] is really helpful whenever we want to examine a given sequence is truly random or not, Borel normality test provides the best results with minimum effort. For example, let's take a sequence of bits that are assumed to be random sequence, 110,011,101,101,110... by seeing the sequence many may answer that the next number might be 1. Since by seeing the pattern, we can able to predict the next coming bit. So, here in this example, we can't guarantee that this sequence is purely random. For this type of small sequence, we may easily conclude but for a sequence with large numbers having periodicity with large intervals it will be difficult to detect that sequence is random or not. So, to find out a sequence is random or not

we used Borel normality test for our testing. To discuss more about Borel normality test, let's take a sequence of bits $S = \{101, 011, 010, 010, 110\dots\}$ the main idea is to divide the sequence with equal size substrings, and then finding the frequencies or probabilities of each generated substring. By using the mathematical formula, we can form 2^i substrings with i characters. So, when the value of $i = 1$ we have to look for the frequencies or probabilities of only $\{0, 1\}$ and if the value of $i = 2$ we have to look for the frequencies or probabilities of $\{00, 01, 10, 11\}$ since they are only the possible values to be formed when we divide the sequence into respective length substrings. By Borel Normality test the substrings with truly random are bound to be

$$\left| \frac{N_i^j(l)}{|l|_i} - \frac{1}{2^i} \right| < \sqrt{\frac{\log_2(n)}{n}}, j = 0, \dots, 2^i - 1$$

The substrings which satisfy to be within this bound proves to be truly random numbers according to the Borel normality test. Or according to our test if we want to compare two random number sequences then the sequence having highest probabilities or frequencies which is having the minimum probability or frequency and the sequences whose is the lowest probability or frequency having the greatest value is said to be the best random sequence. Or in other words, the sequence having lowest difference between the highest and lowest probabilities or frequencies is said to be the best random sequence.

5.2 Testing with Python Random Function

For testing efficiency of our algorithm, we compared our algorithm with different algorithms that are already existed.

Here we plotted the probabilities of the results generated by python random function against random numbers generated. We got the highest probability value of 0.080 for 0 and the lowest value of 0.046 for 9 (Fig. 3).

5.3 Testing with QuantumRNG

Here we plotted the probabilities of the results generated by QuantumRNG against random numbers generated. We got the highest probability value 0.101 for 0 and the lowest value 0.049 for 14 and 15 (Fig. 4).

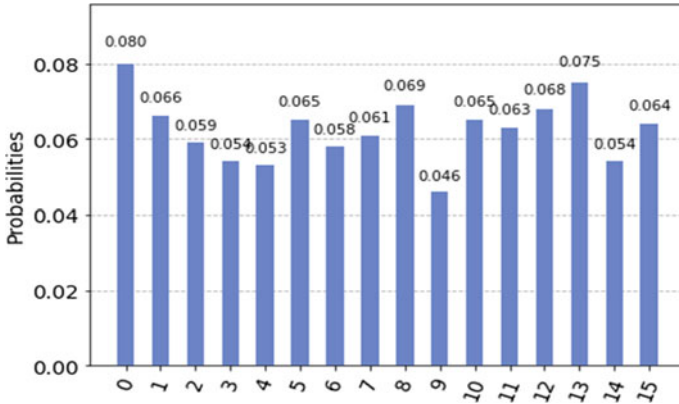


Fig. 3 Output (probabilities of the numbers generated by python random function)

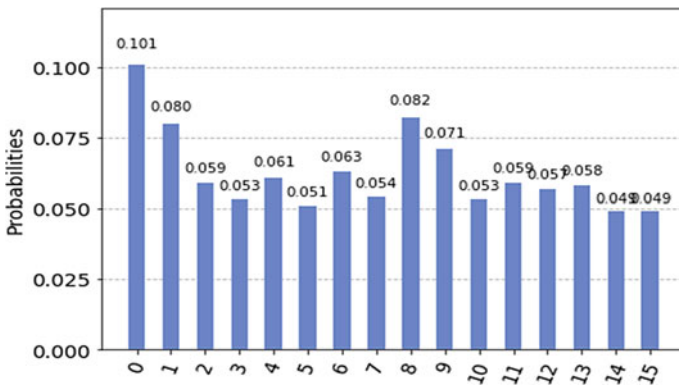


Fig. 4 Output (probabilities of the numbers generated by QuantumRNG)

5.4 Testing with QiskitRNG

Here we plotted the probabilities of the results generated by QiskitRNG against random numbers generated. We got the highest probability value of 0.113 for 0000 and the lowest value of 0.024 for 0100 (Fig. 5).

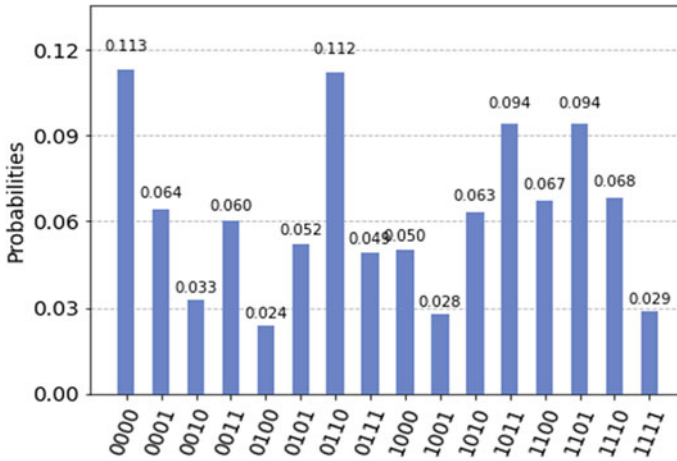


Fig. 5 Output (probabilities of the numbers generated by QiskitRNG)

6 Result and Discussions

6.1 Results from *Ibmq_belem*

Here we discuss about comparisons and behavior of different algorithms when executed in System- “*ibmq_belem*”.

In Figs. 6 and 7 the graphs are showing the results of our QuantumRNG as well as QiskitRNG when executed in the system “*ibmq_belem*” available in IBM Quantum Systems. As per Borel Normality test, we observe occurrence of all digits in sequence and compute their probability. Generator having high occurrences is less random. In terms of probability, algorithm which has least probability has less number of occurrences i.e., algorithm is more random. Comparing Figs. 6 and 7, QiskitRNG has the highest probability which is 0.105 whereas, QuantumRNG has highest probability of 0.085. Therefore, it can be concluded from this result, that QuantumRNG is more random than QiskitRNG.

By following the Borel Normality test, by comparing Figs. 8 and 9, it shows that QuantumRNG has the highest probability which is 0.086 whereas, Python_random has highest probability of 0.078. But we cannot conclude that Python_random is more random than QuantumRNG. This is because, Python_random function has no noise influence on its performance, as it can be executed on our system which is why noise is none. Whereas, QuantumRNG is run on IBM Quantum systems that are still under development stage that adds noise to the result.

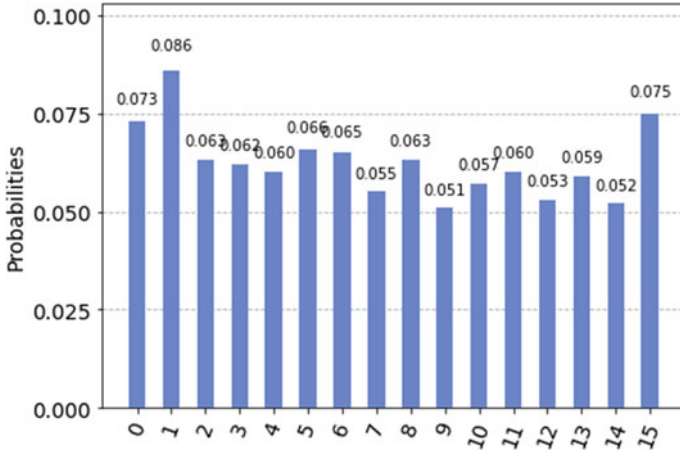


Fig. 6 Probabilities of the numbers generated by QuantumRNG using Ibmq_belem comparing with QiskitRNG

6.2 Results from Ibmq_manila

Here we discuss about comparisons and behavior of different algorithms when executed in System-“ibmq_manila”.

As per Borel Normality test, we observe occurrence of all digits in sequence and compute their probability. Generator having high occurrences is less random. In terms of probability, algorithm which has least probability has less number of occurrences

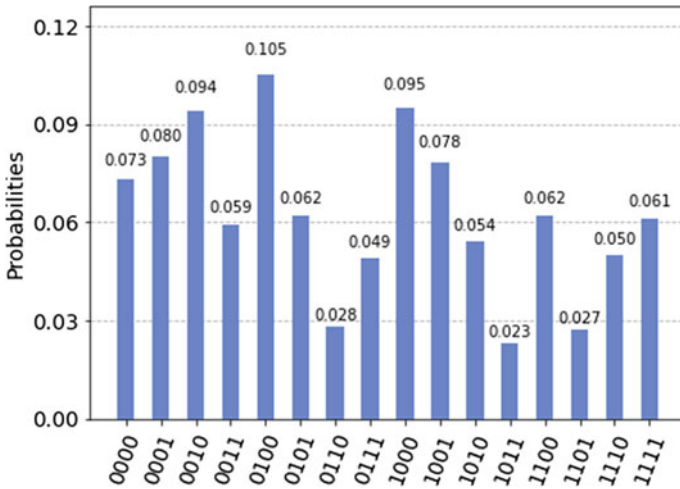


Fig. 7 Probabilities of the numbers generated by QiskitRNG using Ibmq_belem comparing with QuantumRNG

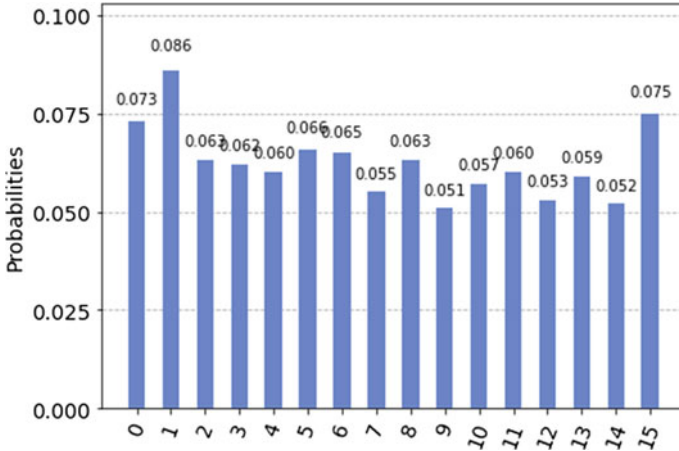


Fig. 8 Probabilities of the numbers generated by QuantumRNG using Ibmq_belem comparing with Python Random Function

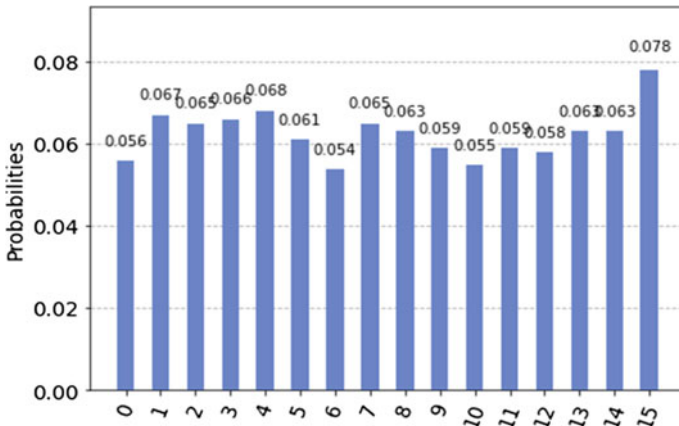


Fig. 9 Probabilities of the numbers generated by Python Random Function comparing with QuantumRNG

i.e., algorithm is more random. By comparing Figs. 10 and 11, QiskitRNG has the highest probability which is 0.113 whereas, QuantumRNG has highest probability of 0.084. Therefore, it can be concluded from this result, that QuantumRNG is more random than QiskitRNG.

As per the Borel Normality test, QuantumRNG has the highest probability which is 0.084 whereas, Python_random has highest probability of 0.073. But we cannot conclude that QuantumRNG is more random than QiskitRNG. This is because, Python_random function has no noise influence on its performance, as it can be executed on our system which is why noise is none. Whereas, QuantumRNG is run

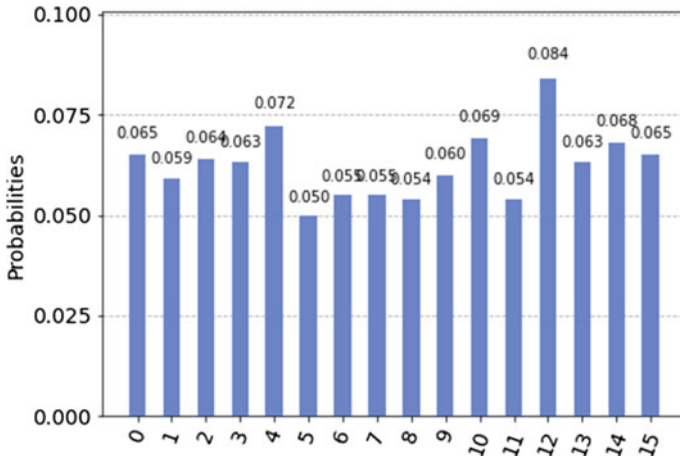


Fig. 10 Probabilities of the numbers generated by QuantumRNG using Ibmq_manila comparing with QiskitRNG

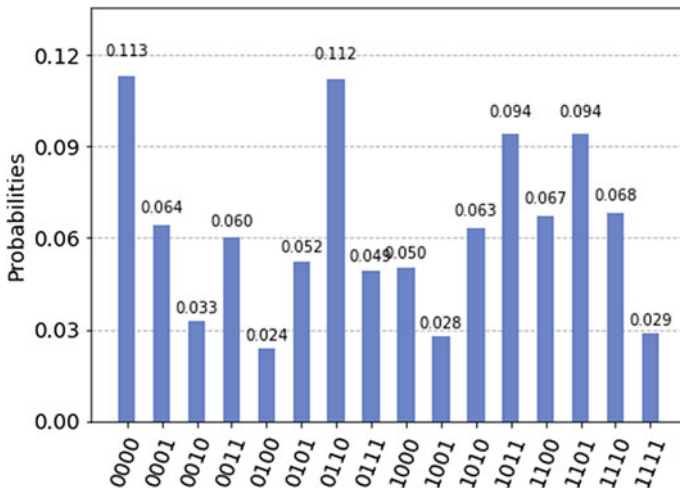


Fig. 11 Probabilities of the numbers generated by QiskitRNG using Ibmq_manila comparing with QuantumRNG

on IBM Quantum Computers which are under development stage so they are subject to noise (Figs. 12 and 13).

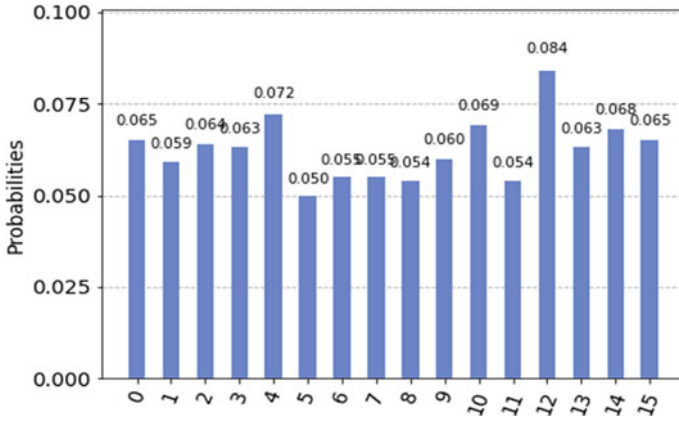


Fig. 12 Probabilities of the numbers generated by QuantumRNG using Ibmq_manila comparing with Python Random Function

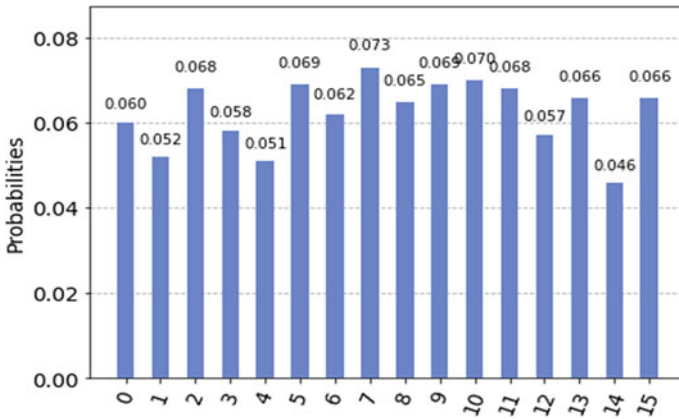


Fig. 13 Probabilities of the numbers generated by Python Random Function comparing with QuantumRNG

6.3 Results from Ibmq_qasm_simulator

Here we discuss about comparisons and behavior of different algorithms when executed in System- “ibmq_qasm_simulator”.

As per the Borel Normality test, by comparing Figs. 14 and 15, QiskitRNG has the highest probability which is 0.112 whereas, QuantumRNG has highest probability of 0.070. Therefore, it can be concluded from this result, that QuantumRNG is more random than QiskitRNG.

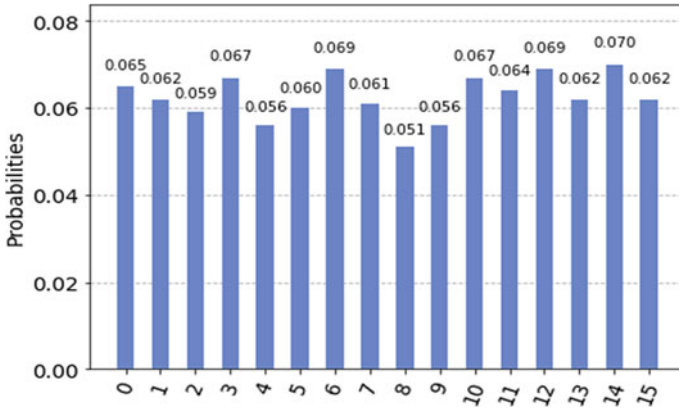


Fig. 14 Probabilities of the numbers generated by QuantumRNG using Ibmq_qasm_simulator comparing with QiskitRNG

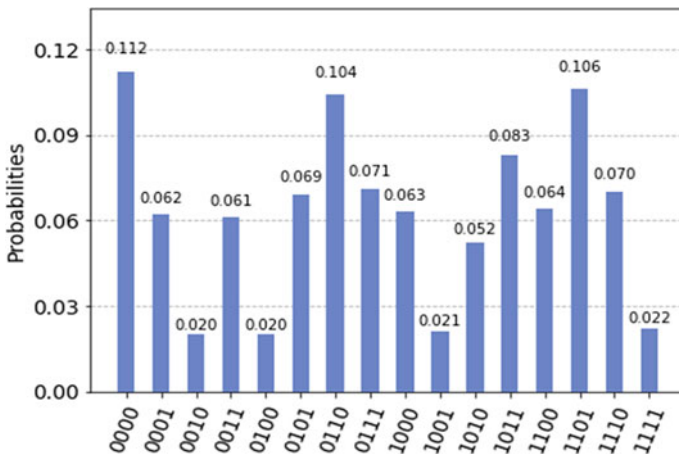


Fig. 15 Probabilities of the numbers generated by QiskitRNG using Ibmq_qasm_simulator comparing with QuantumRNG

As per the Borel Normality test, by comparing Figs. 16 and 17, QuantumRNG has the highest probability which is 0.070 whereas, Python_random has highest probability of 0.074. Since the current job is executed in a simulator which doesn't have any noise, so QuantumRNG performed better compared to python algorithm.

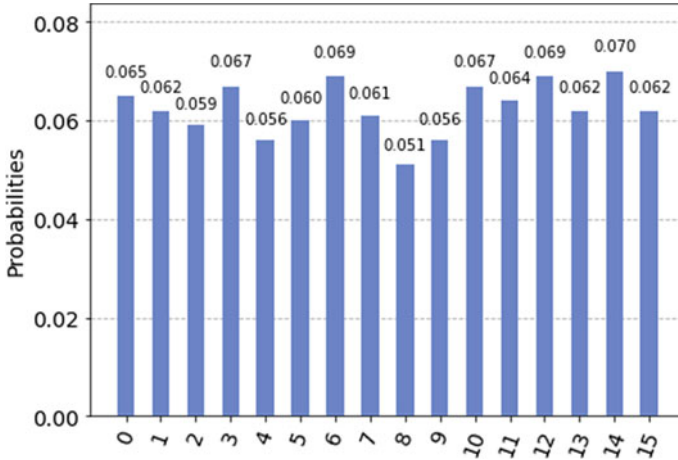


Fig. 16 Probabilities of the numbers generated by QuantumRNG using Ibmq_qasm_simulator comparing with Python Random Function

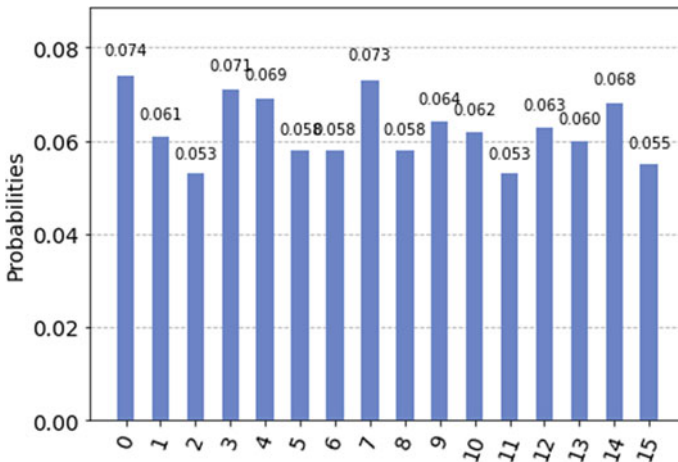


Fig. 17 Probabilities of the numbers generated by Python Random Function comparing with QuantumRNG

7 Conclusion

In this work, we have developed algorithms for single random number generator and multiple random number generator. We have tested our algorithms using Borel Normality test. The results show that our algorithms outperform the other standard algorithms most of the time. Finally, we can conclude that QuantumRNG is low cost and efficient quantum algorithm to generate truly random numbers using only one Qubit. As QuantumRNG is written in Python programming language, it is compatible with most of the operating systems, and we can easily use them in many microcontrollers and microprocessors.

References

1. Herrero-Collantes M, Garcia-Escartin JC (2017) Quantum random number generators. *Rev Modern Phys* 89(1):015004
2. Indhumathi Devi D, Chithra S, Sethumadhavan M (2019) Hardware random number generator using FPGA. *J Cyber Secur Mob* (2019):409–418
3. Calude CS, Dinneen MJ, Dumitrescu M, Svozil K (2010) Experimental evidence of quantum randomness incomputability. *Phys Rev A* 82(2):022102–0221028. <https://doi.org/10.1103/PhysRevA.82.022102>
4. Jofre M, Curtly M, Steinlechner F, Anzolin G, Torres JP, Mitchell MW, Pruneri V (2011) True random numbers from amplified quantum vacuum. *Opt Express* 19:20665–20672
5. Poornachandran P, Nithun M, Pal S, Ashok A, Ajayan A (2016) Password reuse behavior: how massive online data breaches impacts personal data in web. *Innovations in computer science and engineering*. Springer, Singapore, pp 199–210
6. Hari S, Kavinkumar C, Niketh GK, Harini N (2019) Enhancing security of one time passwords in online banking systems. *Int J Recent Technol Eng* 7:319–324
7. Arul R, Raja G, Almagrabi AO, Alkathheiri MS, Chauhdary SH, Bashir AK (2019) A quantum-safe key hierarchy and dynamic security association for LTE/SAE in 5G scenario. *IEEE Trans Ind Inform* 16(1):681–690
8. Vallone G, Marangon DG, Tomasin M, Villoresi P (2014) Quantum randomness certified by the uncertainty principle. *Phys Rev A* 2014. APS
9. Shastry MC, Nagaraj N, Vaidya PG (2006) The B-exponential map: a generalization of the logistic map, and its applications in generating pseudo-random numbers
10. Shiva Prasad R, Siripagada A, Selvaraj S, Mohankumar N (2019) Random seeding LFSR-based TRNG for hardware security applications. *Integrated intelligent computing, communication and security*. Springer, Singapore, pp 427–434
11. Kavulich JT, Van Deren BP, Schlosshauer M (2021) Searching for evidence of algorithmic randomness and incomputability in the output of quantum random number generation. *Phys Lett A* 388:127032
12. Calude CS, Dinneen MJ, Dumitrescu M, Svozil K (2010) Experimental evidence of quantum randomness incomputability. *Phys Rev A* 82(2):022102
13. Sadhu A, Das K, De D, Kanjilal MR, Bhattacharjee P (2022) A QCA-based improvised TRNG design for the implementation of secured nano communication protocol in ATM services. *Computational advancement in communication, circuits and systems*. Springer, Singapore, pp 281–290
14. Aji A, Jain K, Krishnan P (2021) A survey of quantum key distribution (QKD) network simulation platforms. In: 2021 2nd global conference for advancement in technology (GCAT). IEEE, pp 1–8

15. Shrivastava S, Ramesh TK (2019) Integration of SDN controller, time-sliding window, and quantum key distribution with resource allocation strategy in optical networks for high security. In: 2019 global conference for advancement in technology (GCAT)
16. Williams CP, Clearwater SH (1998) Explorations in quantum computing. Telos
17. Steane A (1998) Quantum computing. *Rep Prog Phys* 61(2):117
18. Acín A, Masanes L (2016) Certified randomness in quantum physics. *Nature* 540(7632):213–219
19. Calude CS (2004) Algorithmic randomness, quantum physics, and incompleteness. In: International conference on machines, computations, and universality. Springer, Berlin, Heidelberg, pp 1–17
20. Aerts D, de Bianchi MS (2016) The extended Bloch representation of quantum mechanics: explaining superposition, interference, and entanglement. *J Math Phys* 57(12):122110
21. Goff A (2006) Quantum tic-tac-toe: a teaching metaphor for superposition in quantum mechanics. *Am J Phys* 74(11):962–973
22. Tipsmark A, Dong R, Laghaout A, Marek P, Ježek M, Andersen UL (2011) Experimental demonstration of a Hadamard gate for coherent state qubits. *Phys Rev A* 84(5):050301
23. Aharonov D (2003) A simple proof that Toffoli and Hadamard are quantum universal. arXiv preprint quant-ph/0301040
24. Calude C (1993) Borel normality and algorithmic randomness. *Developments in language theory*, p 113

Sign Language Detection Using Machine Learning



P. Ilanchezhian, I. Amit Kumar Singh, M. Balaji, A. Manoj Kumar,
and S. Muhamad Yaseen

Abstract Sign language detection project is to detect the sign language hand gestures, which really helps the common people like is to understand what a deaf or mute people are trying to converse with us. The sign language detection translates the sign language, in which user forms a hand shape that is structured signs or gestures. In sign language, the configuration of the fingers, the orientation of the hand, and the relative position of fingers and hands to the body are the expressions of a deaf and mute person. Based on this application, the user must be able to capture images of the hand signs or gestures using web camera and they shall predict the hand signs or meaning of the sign and display the name of sign language on screen. At first, we will be taking sample images of different signs, for example, hello, eat, thankyou, etc. Then we are going to label the images with the LabelImg python application file, which is very helpful for object detection. The LabelImg application file develops an XML document for the corresponding image for the training process. In the training process, we have used TensorFlow object detection API to train our model. After training the model, we have detected the sign language or hand gestures in real time; with the help of OpenCV-python, we access the webcam and load the configs and trained model, so that we have detected the sign languages in real time.

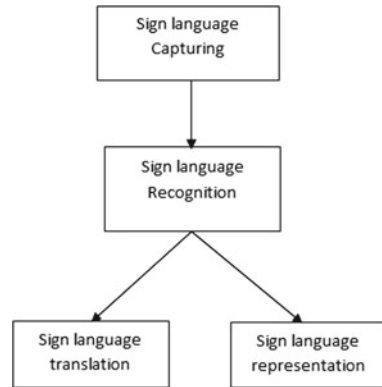
Keywords Sign language · Hand gestures · LabelImg · TensorFlow object detection API

1 Introduction

Deaf–mute people rely heavily on sign language to communicate with people or to communicate with each other [1]. The importance the sign language is being ignored, normal people tend not to give much importance to sign language unless and otherwise their loved ones are unable to speak or hear [2]. Sign language is one of the possible ways to communicate with people who are deaf and mute. In

P. Ilanchezhian (✉) · I. A. K. Singh · M. Balaji · A. M. Kumar · S. M. Yaseen
IT Department, Sona College of Technology, Salem, Tamil Nadu, India
e-mail: ilanchezhianp@sonatech.ac.in

Fig. 1 Application diagram of sign language



sign language, the person uses various hand signs or gestures to communicate with another person, each sign refers to a unique name or has a unique meaning [3]. As the deaf–mute people when try to communicate with normal people their communication becomes very difficult, unable to express their information or feelings to them [4, 5].

Sign language detection application can detect the hand sign or gestures and label them with specific meaning tags towards the hand sign [6]. At first, the user needs to have a webcam, the user needs to perform the hand signs of the sign language, and the application predicts the real-time sign language [7]. The Sign language detection application detects the real-time sign language and user instantly gets to know about the meaning of the sign, which is very much useful and becomes very much possible for deaf–mute people to communicate with people [8] (Fig. 1).

2 Literature Survey

In our literature review, we looked at other similar studies that have been applied in the sign language detection field. It shows that there have been many explorations are done to tackle the hand signs or gestures recognition using several methods and algorithms [9].

Recognition of gestures or expressions systems, in order to recognize sign gestures or expressions you'll need this. The research is based on a variety of input sensors, gesture segmentation, feature extraction, and classification approaches [10]. The goal of this work is to examine and compare the methods utilized in SLR systems, as well as classification approaches and techniques that have been applied, and to recommend the most optimistic way in order to conduct future studies [11, 12]. Many of the recently presented studies, such as hybrid approaches and deep learning, contribute to classification methods due to recent advancements in classification methods [13]. The categorization approaches employed in previous sign language recognition systems are the topic of this paper. In this paper [14] presented, HMM-based techniques, as well as their adaptations, have been extensively researched in the past.

In this study, the paper presents various approaches for making it easier for people to recognize signs while communicating and text will be generated as a result of those symbols [15, 16]. Takes a picture with your webcam and converts it to grayscale. It checks for hand motion. In grayscale segmentation, the Otsu thresholding algorithm is used to create an image of a hand gesture of total picture, then it is separated into two classes: hand and background [15]. The ratio between class variance and total variance is used to establish the optimal threshold value. To locate the hand gesture's boundaries in an image Canny edge detection technique is employed. When it comes to canny edge detection, edge-based and threshold-based segmentation were utilized. Otsu's algorithm is used in the system for its simplicity and stability.

In this paper [17], the major goal was to develop an application that would transform sign gesture or language into text in English and voice [17–21], hence facilitating sign language communication. The application uses a webcam to collect visual data, which is then preprocessed using a combinational method before being detected via template matching. The text-based translation is subsequently transformed into audio [21]. In this system's database, there are 6000 photos of English alphabets in this collection. A total of 4800 photos were utilized in the training and, for testing purpose, 1200 images have utilized [17]. The system has an accuracy of 88 percent.

This research proposes a method for determining the number of available fingers in an action that is both efficient and fast which represents the alphabet of binary sign language. The hand and camera do not have to be exactly aligned for this technique to work [18]. The program will make use of an image processing system to recognize the English letter hand signs or sign gesture used by hearing-impaired people for interaction [19]. The project's main objective is to create a digitalized efficient system that empowers people to use natural hand movements to communicate with everyone else [17, 18]. The aim is to build and create a smart system that captures the visual photos as data input of sign language hand motions and produces easily recognized results using image processing, machine learning, and artificial intelligence principles [20].

3 Methodology

The proposed system is sign language detection using object detection and machine learning which can identify various hand signs or gestures in real time using webcam; it also labels the hand signs with the relevant text. There are various modules used in the system—NumPy, OpenCV, TensorFlow, OS, LabelImg, and Object detection API. These are some of the important APIs and Modules used in sign language detection system (Fig. 2).

There are various steps involved in this project as follows:

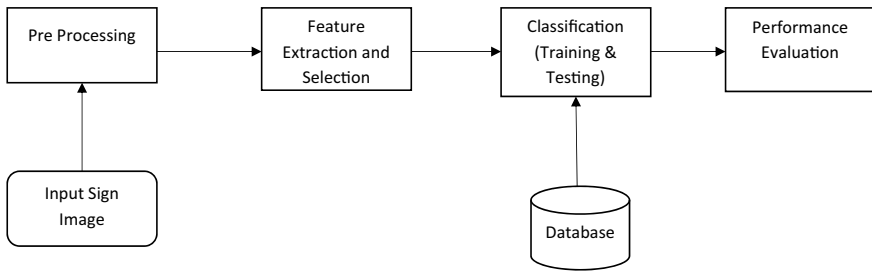


Fig. 2 Sign language detection block diagram

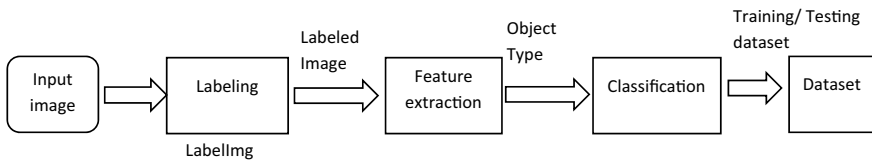


Fig. 3 Image labeling block diagram

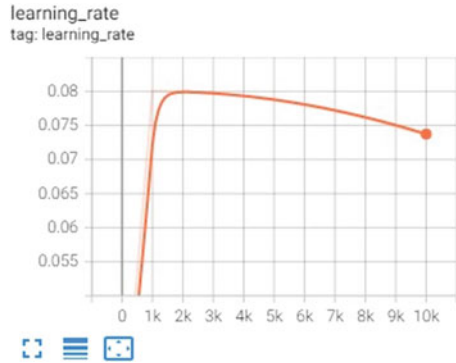
3.1 Image Annotation

The first step is collecting sample images with various hand signs in our project. Here, in this project, we have taken the real images of ourselves not depending on any other sources. Image annotation is the process of labeling image of datasets to train the model. Therefore, it is used to label features that system needs to recognize the gesture. For this, we have used LabelImg tool to label the image. It is free and open-source tool for graphically labeling image. We are labeling each image corresponding to its signs. After labeling the image, this tool creates an XML file for each image, which was really useful for further training (Fig. 3).

3.2 Classification

In this step, we will be classifying or splitting the dataset into two parts: Training dataset and Testing dataset. We have created tf records for both training and testing datasets. These files consist of unique id numbers and label names which were used to classify the sign language or sign gestures.

Fig. 4 Learning rate graph



3.3 Training the Model

After the classification part is done, next is training the model. For this, we have used an application programming interface—TensorFlow Object detection API. It is the foundation for building a deep learning network, so that problems with object detection will be easily solved. It consists of various pretrained models which are referred to as Model Zoo. MobileNet—SSD `ssd_mobilenet_v2_fpnlite_320 × 320_coco17_tpu-8` this is the version we have used in our project. The SSD architecture consists of a single convolution network that learns to predict and classify bounding box locations. As a result, SSD can be trained from beginning to end. Using this object detection API and pretrained model, we have trained our model for 10000 steps and, with these many number of steps, it has a 0.715 learning rate (see Fig. 4).

3.4 Real-Time Detection

Since we have trained our model, we have tested our trained model. So, for the real-time sign language detection, we have to import some python modules, namely OpenCV to access the webcam of the laptop and NumPy which converts the incoming to array format. The modules are loaded and trained. So, we started to test for detecting real-time sign language as the camera began it started video capturing and, with the help of NumPy module, we converted the photos into arrays and system was comparing with model. Then it started detecting the sign languages by drawing the boxes around the hand signs and labeling with corresponding sign language names (Figs. 5 and 6).

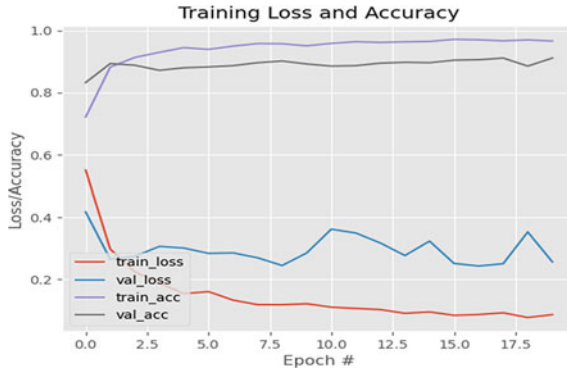


Fig. 5 Training loss and accuracy graph

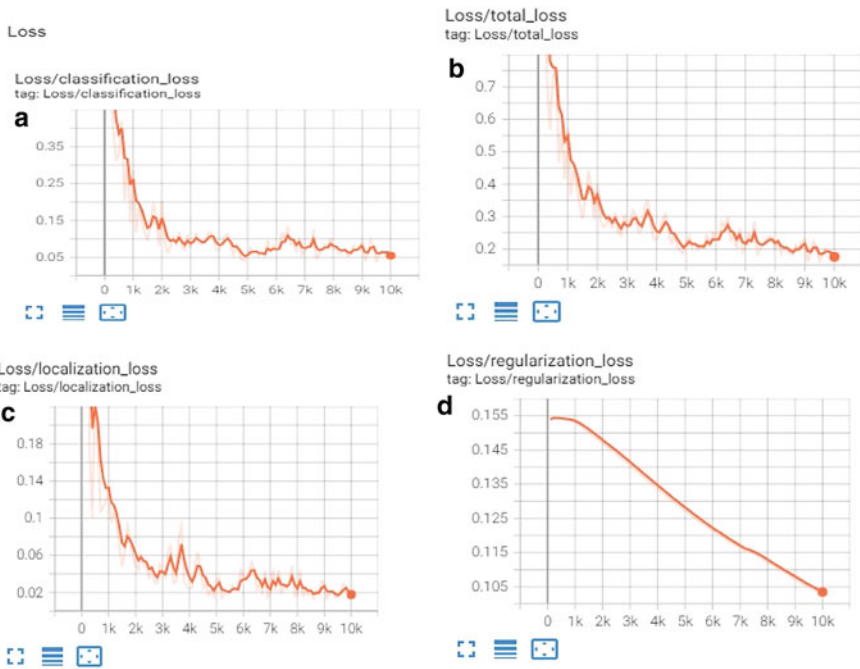


Fig. 6 a Classification loss graph. b Total loss graph. c Localization loss graph. d Regularization loss graph

4 Result

From the comparison table (see Table 1), a comparison between object detection mobnet SSD and convolution neural network. We would like to conclude that, from the above two methods, object detection has a higher efficiency and accuracy

Table 1 Comparison between object detection Mobnet SSD and CNN

	Object detection Mobnet SSD	Convolution neural network
Dynamic detection	94	71
Static detection	97	94
Single-hand sign	99	95
Double-hand sign	95	73
Hand sign switches	91	84

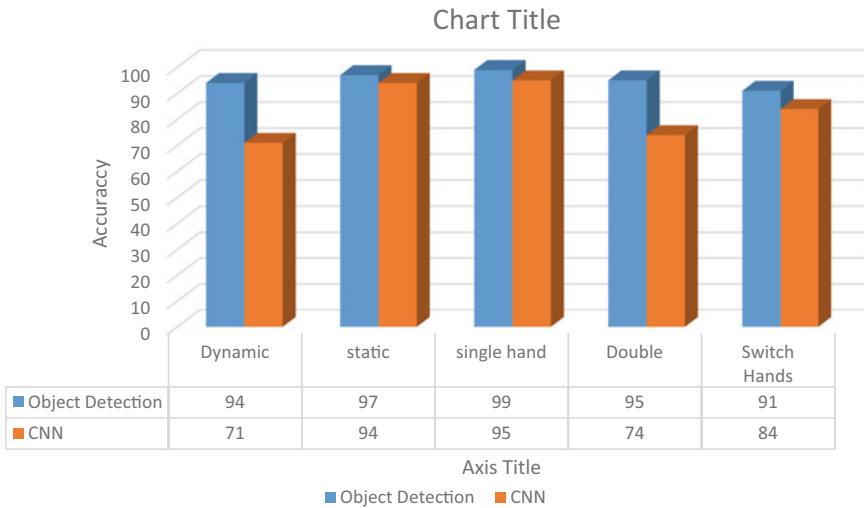


Fig. 7 Comparison graph

compared to CNN. We have obtained the values by testing the object detection-trained model. It is efficient in both dynamic and static detections and also in single and double-hand sign detections and also detects the signs effectively when switch between the signs. In CNN, it is efficient in static detection and single-hand sign detection. Object detection method has better performance in dynamic detection and double-hand signs. It detects hand signs in real-time and has better performance (see Fig. 7).

5 Conclusion

The sign language detection system has evolved from being able to classify simply stationary or fixed symbols and alphabets to being able to identify dynamic symbols and movements in consecutive series of images. Researchers today are giving more thought on creating large-scale vocabularies for sign language recognition systems.

Many developers use a limited vocabulary and homebrew databases to develop sign language detection system. In some countries involved in the implementation of programs for recognizing sign languages, building large-scale databases on behalf of sign language gratitude systems is not yet obtainable. Using Kinect-based data, which gives color and depth stream video, is the most notable. The method of classification used to identify sign gestures or language also depends on the researcher. It is still subjective to compare one method to another, using its own constraints and perspectives on sign language detection system. The direct comparison between approaches is limited because of variation and differences in the sign language. Most countries' variations in sign gesture or language are dependent on their grammar and how they express each word, such as how they present the language by term or by phrase.

References

1. Chowdhury A, Cho SJ, Chong UP (2011) A Background subtraction method using colour information in the frame averaging process. In: Proceedings of 6th international forum on strategic technology
2. Geetha M, Manjusha (2012) A vision based recognition of Indian sign language alphabets using B-spline Approximation. *Int J Comput Sci Eng*
3. Huang Z, Li H (2015) Sign language recognition using convolution neural networks. In: Institute of electrical and electronics engineers international conference on multimedia and expo
4. Bantupali, Xie (2018) American sign language recognition using computer vision and deep learning. In: Institute of electrical and electronics engineers conference on Bigdata
5. Nimisha K, Jacob A (2021) A brief review of the recent trends in sign language. In: Proceedings of the Institute of Electrical and Electronics Engineers 2020 international conference of communication and signal processing
6. Arman, Shashidhar R, Shashank K, Sukumar T, Safeel (2021) A review on Sign language recognition techniques. In: Proceedings of international conference of innovation in technology of IEEE
7. Elham M, Hurroo M (2020) Sign language recognition system using CNN and computer vision. *Int J Eng Res Technol*
8. Al Hammadi M, Muhamed G, Alsulaiman M, Amine M (2020) Sign language recognition with efficient hand gesture representation using deep learning. *Inst Electr Electron Eng*. <https://doi.org/10.1109/ACCESS.2020.3032140>
9. Anderson R, Wiryana F, Ariesta MC, Kusuma GP (2017) Sign language recognition application for deaf mute people. In: 2nd international conference on computer science and computational intelligence
10. Goyal S, Sharma I (2013) Sign language recognition system for mute and deaf people. *IJERT—Int J Eng Res Technol*
11. Raheja JL, Mishra A, Chaudary A (2016) Indian sign language recognition using SVM. *Pattern Recognit Image Anal*
12. Singh A, Kanika, Goyal (2014) A system of Indian sign language recognition system for deaf people. *J Today's Ideas—Tomorrow's Technol*
13. Kudrinko K, Flavin, Zhu X, Li Q (2020) A comprehensive review on wearable sensor based sign language recognition. *Inst Electr Electron Eng Rev Biomed Eng*
14. Davari A, Fanl J, Mekala P, Gao Y (2014) Real time Sign language recognition based on neural network architecture. In: Institute of Electricals and Electronics Engineers 43rd symposium on system theory

15. Raut M, Machhale K, Dhok P, Hora J (2015) Indian sign language recognition system for Deaf people using Otsu's algorithm. IRJET—Int Res J Eng Technol
16. Tewari D, Srivastava S (2012) A visual static Hand gesture recognition in Indian sign language using self-organizing map algorithm. IKEAT—Int J Eng Adv Technol
17. Todkar A, Patil M, Vedak O, Zavre P (2019) Sign language interpreter using ML and image processing. IRJET—Int Res J Eng Technol 6(4)
18. Pramada S, Pranita N, Samiksha N, Saylee D, Archana S (2013) Intelligent sign language recognition using image processing. Int Organ Sci Res J Eng 3(2):45
19. Jeyapal A, Ganesan J, Sabeenian RS, Subramanian L, Anbalagan N (2020) A comparative study of feature detection techniques for navigation of visually impaired people in an indoor environment. J Comput Theor Nanosci 17(1):21–26
20. Saraswathi K, Mohanraj V, Suresh Y, Kumar JS (2021) A hybrid multi feature semantic similarity based online social recommendation system using CNN. Int J Uncertainty Fuzziness Knowl-Based Syst (2021)
21. Akilandeswari J, Jothi G, Naveenkumar A, Sabeenian RS, Iyyanar P, Paramasiyam (2022) Design & development of an indoor navigation system using denoising auto encoder based on CNN for the visually impaired people. Multimedia Tools Appl 81(41)

Scrutinize and Discover of Image of Freshwater Taken by Faraway Realizing Using FFNN and ConvNet Mechanisms



D. Komalavalli, P. Ilanchezhian, A. Diwakar, K. Gayathri, T. S. Indhuja, and R. V. Devadharshini

Abstract Water is the most momentous for all types of species, this need is notably more predominant for anthropoids, and this is since blood in the anthropoid body requires about 90% of water. The quantity of fresh water is on globe remnants constant, but the inhabitants are just too sweeping, this is why there is a more scarcity of freshwater and it is extensively spoken among the people. It is therefore salient to unambiguously gauge the amount of freshwater is on the globe. To estimate it, first take the image utilizing remote sensing and then to discover the information about the water in the image, some mechanism utilizing in image processing is utilized. First, it segregates the attribute highlights of the image to precisely assess the information in the image about water, and then it is trouble-free to discover the objects or scrutinize the objects or formation it meaningful in the image through the image section with those attributes. Therefore, two arrangements are utilized in this paper to recognize these operations more precisely. This paper utilizes the Feed Forward Neural Network (FFNN) system for segregating attributes and then the CNN mechanism for segmentation.

Keywords Freshwater information · Remote sensing image · FFNN · Convolutional Neural Network (CNN) · Accurate analysis

1 Introduction

In general, remote sensing is the ability to perceive information about an object from a distance. Researchers use remote sensing to track atmospheric events, in land applications, earth-based studies and for military observations. But here it is used to monitor the amount of freshwater on Earth. This paper is proposed to analyze the image taken by remote sensing.

Computer vision has been used extensively to analyze an image in the past, but nowadays, because of the development of deep learning, everyone has shifted to

D. Komalavalli · P. Ilanchezhian (✉) · A. Diwakar · K. Gayathri · T. S. Indhuja · R. V. Devadharshini
Department of IT, Sona College of Technology, Salem, India
e-mail: ilanchezhianp@sonatech.ac.in

deep learning. It is very difficult to solve those kinds of problems from a computer point of view but it has become much easier with this deep learning method. In this deep learning mode, we can analyze an image and learn its meanings very easily and without problems.

In these deep learning structures, there are many features for image analysis such as image segmentation, image colorization, object detection, image classification, and so on. So in this paper, the quantity of freshwater can be detected by applying the deep learning methods to the image. For image analysis, methods such as partitioning features and dividing as parts of the image are used.

In [1], by remote sensing, wetlands are identified. For this, the paper editor has used digital image processing. So they have been trying to find the wetlands with the satellite image data. So in order to do this, they have done the neural network in terms of machine learning. This concept has been neglected because machine learning is less efficient than deep learning. In [2], submarine research was carried out to investigate or detect undersea water events. Before that, research was done by diving into the deep sea. Although this has been done from time to time, it has given many a hard time. Moreover, underwater, dark, dim and dusty, the status of the image apprehended by the optic camera is substandard [3]. To ameliorate the quality of the image, the talent adopted a methodology based on BEMD. But this method does not give a good result [4].

In [5], we report the effects of water pollution on a daily basis, That is, water is contaminated with agricultural waste, industrial waste, hospital waste and sewage discharge. In addition, aquatic fishes are classified as waste fish and good fish. The author of this paper has attempted to distinguish between the waste fishes by leaving the day mummies in them. They have therefore used statistical analysis. But this is old time and it is very poor quality. In [6], the perception of images from a distance is studied using CNN methods. In this they have taken the picture of extracting the ships through a distant sensation. They have used a CNN Mechanism of detecting ships by extracting sea water. But, using only the CNN, the quality is not great [7].

In [8], it is very difficult to separate water and land images by artificial aperture radar. This is because the quality is greatly reduced by irregular shape, excessive noise, and dust [9]. The author therefore uses the matrix method to segregate the features of the image; author then used the DoG mechanism to integrate all of the main features. Although the author of this paper has tried, it has given little result. In [10], in this community, all the places with water bodies are the most important, therefore, safeguard the water bodies and to beware of disasters such as tsunami, it is very important to research the image of water bodies [11]. Since this requires a more accurate result, the author of this paper has used methods such as LREP and MTM. But these methods are not working according to their needs.

In [12], due to agriculture and socioeconomics, there is a need for a more accurate separation of rivers. Natural disasters such as flooding require precautionary measures. This can lead to fatalities and property losses. So this method is easier because of the development of technologies like Remote Sensing [13]. The author of this paper has therefore used the method of PSO-SVM. But this process did not give them the results they expected. In [14], in this paper, they have used the deep

learning method for image classification to be very accurate. Of these, there are some problems due to lack of data. So they have used the feature extraction method. For feature extraction, the CNN method is used. Their attitude was largely unanswered [15].

In [16], image processing is now widely used in deep learning areas, traffic issues are getting bigger due to Smart City, because, the problem comes due to the violation of traffic rules, congestion and festivals. This author classifies the problem of traffic with respect to physical features to detect traffic [17]. They have used the classification in deep learning mode to clearly identify the traffic problem from the video. Their accuracy was very low. In [18], the CNN algorithm is used to classify the histopathology image. But, using only the CNN method, the quality of the image does not give much quality. In [19], the aim is to make the deep learning method more useful for analyzing the medical image, so, this paper was proposed by this author. So they used a neural network called SRNet. It requires more money. In [20], the CNN procedure, which is the system of the deep learning method, was used to automatically classify the images, but money was more needed to automate this process.

2 Proposed Rule

The instances of this proposed method are given in the following steps:

Step 1: Firstly, the image taken by remote sensing should be given in this proposed manner as input.

Step 2: This means that have to give this input to the feature separation method. This feature separation method will only extract the features required by MLF.

Step 3: Then the MLF method will give as output what is needed.

Step 4: These outputs then go to the CNN mode as input.

Step 5: This CNN system divides the image into parts and gives clear information (Fig. 1).

(1) Feature extraction

Here the MLF method is utilized to extract the sfeatures of the images. That is, Multi-layer feed-forward (MLF) neural networks. There will be a lot of features in the film, but the separation of the image with all of them is the reason for reducing the image quality. Other than that, a few important features are difficult to obtain, As such; image separation is another reason for reducing image quality when those important features are not available. Therefore, this paper uses this MLF method to reduce these problems. It only extracts the essential and important features of the image and releases it.

The following is a MLP with m-layer. It calculates the one-dimensional output in the n-dimensional input.

1. The activation function of the release perceptron is g_o , and the activation function of the perceptron of the hidden layer is g .
2. With each perceptron in the l_{i-1} layer, the perceptron in the l_i layer is attached. Each of the layers is fully enclosed. Each perceptron therefore depends on the outputs of all the perceptrons in the preceding layer. And this is the weight that connects the two perceptrons is still zero, so it does not lose its common separation. That is, for those who have no contact, it will be equal.
3. Even in the same layer, there is no connection between the perceptrons.

Figure 2 is called as Image of MLP, fully enclosed in three entries with two hidden layers and each contains four perceptrons. Below is a definition for each of the codes:

- w_{ij}^k The weight for the perceptron j in the layer l_k for the inside edge i .
- b_i^k The pro for perceptron i is in the l_k layer.
- h_i^k The product dependence and sum for the perceptron i is in the l_k cascade.
- o_i^k The output for the node i is in the l_k layer.
- r_k The no. of nodes in layer l_k .
- \vec{w}_i^k The weight vector for the perceptron i in the l_k layer.
- \vec{o}^k The o/p vector for layer l_k .

The calculation of the output of MLP is continued in the following steps:

Step 1: Take the input layer i_0 :

The values of the o_i^0 outputs for nodes in the input layer i_0 , Set in the vector $\vec{x} = \{x_1, \dots, x_n\}$ to their corresponding inputs.

Step 2: Calculate the outputs and product sums of each hidden layer from i_1 to i_{m-1} :

For $k \rightarrow$ from 1 to $m - 1$,

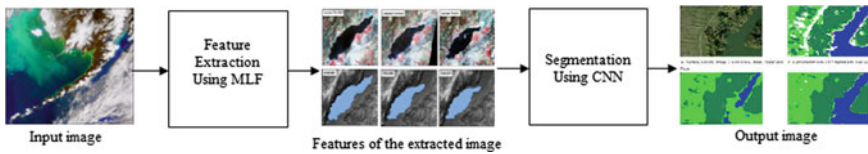
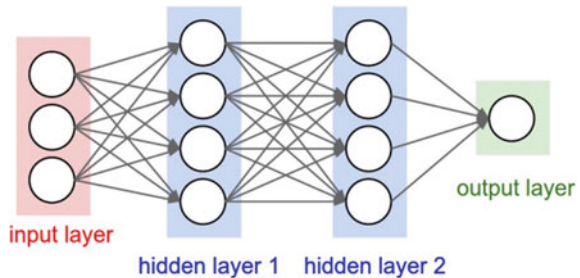


Fig. 1 Flow of analyzing the image taken by the sensation of distance

Fig. 2 Structure of MLP



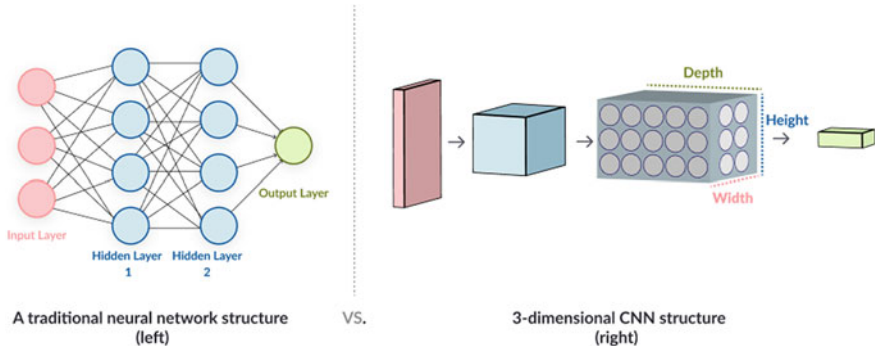


Fig. 3 Structure of 3-D CNN

(1) Compute

$$h_i^k = \vec{w}_i^k \cdot \vec{o}^{k-1} + b_i^k = b_i^k + \sum_{j=1}^{r_{k-1}} w_{ji}^k o_j^{k-1} \text{ for } i = 1, \dots, r_k;$$

(2) Compute

$$o_i^k = g(h_i^k) \text{ for } i = 1, \dots, r_k$$

Step 3: Find the output lm for the Release layer y:

(1) Compute

$$h_1^m = \vec{w}_1^m \cdot \vec{o}^{m-1} + b_1^m = b_1^m + \sum_{j=1}^{r_{m-1}} w_{j1}^m o_j^{m-1}$$

(2) Compute

$$o = o_1^m = g_o(h_1^m)$$

(2) Training MLPs

This refers to the set of input–output pairs $X = \{(\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N)\}$ of as single-layer perceptron, to reduce the mean square error; the learning method consists of refreshing the values of \vec{w}_i^k and b_i^k .

$$E(X) = \frac{1}{2N} \sum_{i=1}^N (o_i - y_i)^2$$

where, o_i indicates the output of the MLP on input \vec{x}_i .

This reduces $E(X)$ to all w_{ij}^k and b_i^k , so it gives it a good extraction, the slope descent should be used to adjust the parameters of w_{ij}^k and b_i^k with the alpha learning rate. The following delta equations are given for each iteration.

Expansion of the right-hand side of the delta rule is taken using back propagation, because it flows backwards through the slope information network. This slope flow develops in the final layer l_m , this is proportional to the difference between the actual output o and the target output y .

$$\Delta w_{ij}^k = -\alpha \frac{\partial E(X)}{\partial w_{ij}^k}$$

$$\Delta b_i^k = -\alpha \frac{\partial E(X)}{\partial b_i^k}$$

(3) Segmentation

CNN methods are great for processing closely interconnected images. It uses a 3 dimensional structure; here are three special neural networks studies that examine the green, blue and red layers of the image. CNN scans only part of the image first, then identifies and extracts only the most important features, and then uses those features to classify the image. CNN Uses 2 or 3 dimensional neural layers to analyze images with conventional, 2 or 3 colored channels. CNN with one-dimensional ones are also very useful. When the features of the section are not required one-dimensional CNN extract important features and enhances the image quality.

A plain vanilla neural network is one in which all neurons in one layer interact with all neurons in the next layer. But this is very limited in its effectiveness in analyzing video and large images. For an average size image with hundreds of pixels and three types of color channels, the no. of parameters used by a traditional neural network can be in the millions; this can be supportive for overindulgent fit.

Use the neural network in important areas of the image, and to manage the no. of parameters uses the 3 dimensional layouts. In this, each set of neurons analyzes the feature of an image or part. All neurons, instead of sending their results to the next neural layer, each group of neurons are focused on identifying a part of the image. This indicates how much each aspect is likely to be part of a segment (Fig. 3).

Usually, the symbol system works in three steps:

1. The first step is a change, in this, the image is scanned into a few pixels at a time, then, each feature belongs to the required section and a segmentation graph is created with probabilities.
2. The second step is pooling. This reduces the dimension of each segment while maintaining its critical data. The pooling standard summarizes the most important image information in the image.

In Fig. 4, This CNN system has the dog's part alone, boat's part alone, and bird's part alone and the cat's part alone, separately divided, The following explanations

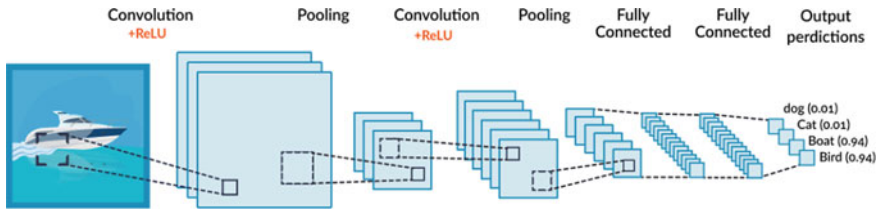
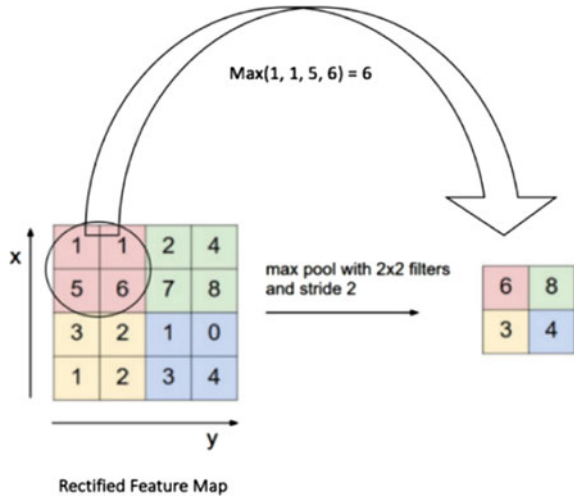


Fig. 4 Working of CNN

Fig. 5 Pixel selection process



show how it works. Many CNN mechanisms maximum use Max Pooling. Here the CNN algorithms extract the highest value from each of the Pixel parts scanned. This is stated Interpretational in Fig. 5,

In this pixel selection process, first, it divides the features in the image into four parts, then, just takes the important part of each area and segment. Another example is given in Fig. 6, in it; the forest, the building, the water, the land and the road are separated separately.

3. Finally, when the parts are properly divided, the CNN algorithm goes to the third step. This is a fully connected neural network, It analyzes the last probabilities, then decides which segment of the given input image will depend on, Then it divides separately into this is water, this is land, and this is forest (Fig. 7).

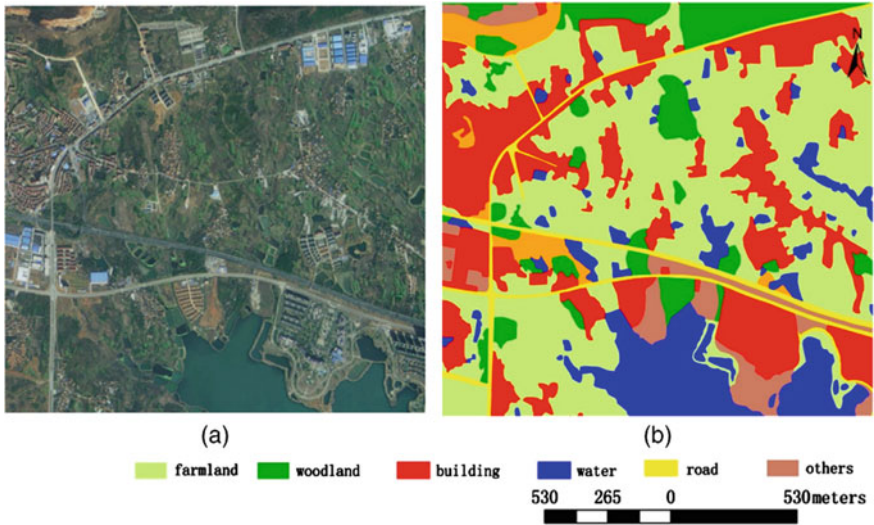
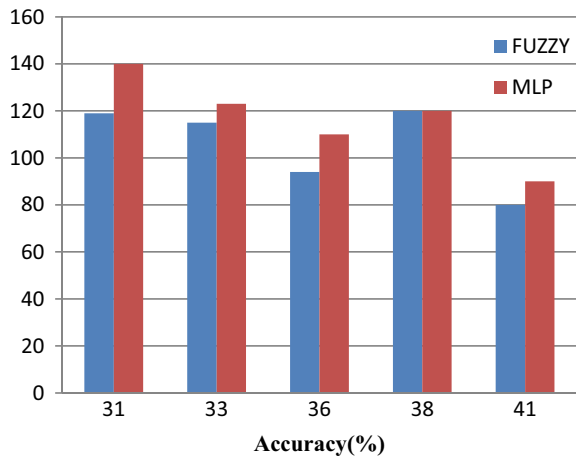


Fig. 6 Segment the remote sensing image

Fig. 7 Comparative diagram of accuracy



3 Results and Discussion

When this accuracy compares the performance, the MLP/FFNN algorithm is much higher than the Fuzzy Logic algorithm. The FFNN algorithm has more values in some important areas, though it has similar values in a few places. This FFNN algorithm is therefore the best in these processes.

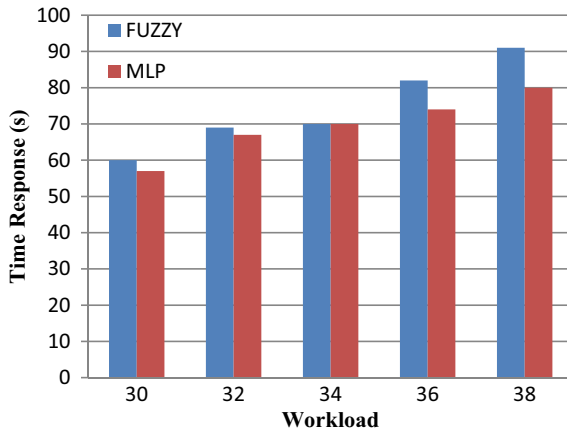


Fig. 8 Comparative chart of time taking

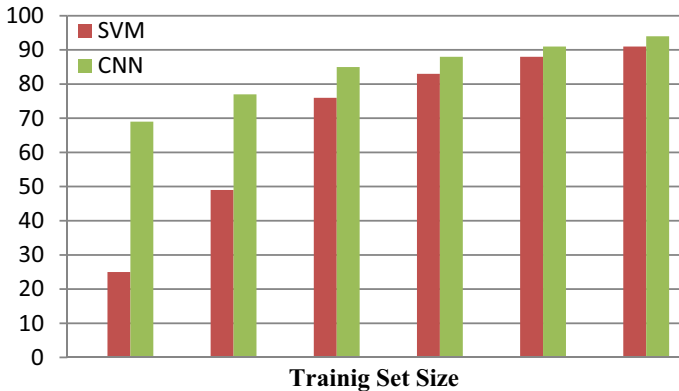


Fig. 9 Performance in accuracy

This MLP process is taking very low compared to the time taking, No matter how much work is put into this process, it takes up very little time. Even in Fig. 8, the FFNN algorithm proves to be better than the Fuzzy Logic Control algorithm (Fig. 9).

In accuracy presentation, the CNN procedure is higher than the SVM algorithm. This is because there is a lot more development performance in CNN than SVM.

4 Conclusion

In this proposed method, the process was developed with the central concept of freshwater detection. This paper takes a lot of hard work for that. For that, it has two

steps, one is FNN/MLP/MLF and another one is CNN. The first method is to analyze a given distance sensing image and then extract the important features. Then all of those features go into the CNN algorithm and some of it actively engages. Then, this method shows and divides only the areas where freshwater is present. In this paper, all the activities that have taken place so far, with these two methods this process was found to be more accurate, quicker, and cheaper. Finally in this paper, these two methods are described in terms of performance better than the other methods.

References

1. Sharapov RV, Varlamov AD (2019) Using neural networks to estimate the degree of wetlands with remote sensing data. In: International youth conference on Radio Electronics, Electrical and Power Engineering (REEPE), 978-1-5386-9334-6/19/\$31.00. IEEE
2. Lakshmi J, Prasanti K, Kalpana S (2019) Visual enhancement techniques for underwater images. In: 5th international conference on advanced computing & communication systems (ICACCS), 978-1-5386-9533-3/19/\$31.00. IEEE
3. Bai Y, Adriano B, Mas E, Koshimura S (2017) Machine learning based building damage mapping from the ALOS-2/PALSAR-2 SAR imagery: case study of 2016 Kumamoto earthquake. *J Disaster Res* 12:646–655
4. Krizhevsky A, Sutskever I, Hinton GH (2012) ImageNet classification with deep convolutional neural networks. In: Proceedings of advances in neural information processing system, pp 1106–1114
5. Mauya R, Dutta MK, Riha K, Kritz P (2019) An image processing based identification of fish exposed to polluted water. In: 42nd international conference on telecommunications and signal processing (TSP), 978-1-7281-1864-2/19/\$31.00. IEEE
6. Lei F, Wang W, Zhang W (2019) Ship extraction using post CNN from high resolution optical remotely sensed images. In: IEEE 3rd information technology, networking, electronic and automation control conference (ITNEC 2019), 978-1-5386-6243-4/19/\$31.00. IEEE
7. Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. *IEEE Trans Pattern Anal Mach Intell* 35(1):221–231
8. Meng Q, Wen X, Yuan L, Xu H (2019) Factorization-based active contour for water-land SAR image segmentation via the fusion of features. *Digit Obj Ident* <https://doi.org/10.1109/ACCESS.2019.2905847>. IEEE Access
9. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011) Multimodal deep learning. In: Proceedings of 28th international conference on machine learning, pp 689–696
10. Meng L, Zhang Z, Zhang W, Ye J, Wu C, Chen D, Song C (2019) An automatic extraction method for lakes and reservoirs using satellite images. *IEEE Access* 7. <https://doi.org/10.1109/ACCESS.2019.2916148>
11. Huang J, Ling XCX (2005) Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans Knowl Data Eng* 17(3):299–310
12. Li X, Lyu X (2019) YaoTong, Shengyang Li and Daofang Liu, An object-based river extraction method via optimized transductive support vector machine for multi-spectral remote-sensing images. *Citation Inf*. <https://doi.org/10.1109/ACCESS.2019.2908232>, [IEEE Access](https://doi.org/10.1109/ACCESS.2019.2908232)
13. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444; LeCun Y et al (1984) Backpropagation applied to handwritten zip code recognition. *Neural Comput* 1(4):541–551
14. Ahn E, Kumar A, Feng D, Fulham M, Kim J (2019) Unsupervised deep transfer feature learning for medical image classification. In: IEEE 16th international symposium on biomedical imaging (ISBI 2019) Venice, Italy
15. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of IEEE international conference computer vision, pp 4489–4497

16. Altundogan TG, Karakose M (2019) Image processing and deep neural image classification based physical feature determiner for traffic stakeholders. In: 7th international Istanbul smart grids and cities congress and fair (ICSG)
17. Architectural Institute of Japan (2018) Report on the damage investigation of the 2016 Kumamoto earthquakes (in Japanese)
18. Wu P, Qu H, Yi J, Huang Q, Chen C, Metaxas D (2019) Deep attentive feature learning for histopathology image classification. In: IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) Venice, Italy
19. Tom F, Sharma H, Mundhra D, Dastidar TR, Sheet D (2019) Learning a deep convolution network with turing test adversaries for microscopy image super resolution. In: IEEE 16th international symposium on biomedical imaging (ISBI 2019) Venice, Italy
20. Lee JH, Kim K-Y, Shin Y (2019) Feature image-based automatic modulation classification method using CNN algorithm and this research was supported by the MSIT, Korea, under the ITRC support program (2018-0-01424) supervised by the IITP, 978-1-5386-7822-0/19/\$31.00. IEEE

Knowledge-based Extraction of Cause–Effect Relations from Biomedical Text



Sachin Pawar, Ravina More, Girish K. Palshikar, Pushpak Bhattacharyya, and Vasudeva Varma

Abstract We propose a knowledge-based approach for extraction of *Cause–Effect* (CE) relations from biomedical text. Our approach is a combination of an unsupervised machine learning technique to discover causal triggers and a set of high-precision linguistic rules to identify cause/effect arguments of these causal triggers. We evaluate our approach using a corpus of 58,761 Leukaemia-related PubMed abstracts consisting of 568,528 sentences. We could extract 152,655 CE triplets from this corpus where each triplet consists of a cause phrase, an effect phrase, and a causal trigger. As compared to the existing knowledge base—SemMedDB [5]—the number of extractions are almost twice. Moreover, the proposed approach outperformed the existing technique SemRep [7] on a dataset of 500 sentences.

Keywords Information extraction · Cause–effect relations · Linguistic rules

1 Introduction

The immense text present in the biomedical domain is growing day by day in the form of research papers, case reports, patient health records, health-related Question–Answering (QA) forums and even social media. The effective extraction of knowl-

Ravina More was working in TCS Research when the work was carried out.

S. Pawar (✉) · R. More · G. K. Palshikar
TCS Research, Pune, India
e-mail: sachin7.p@tcs.com

G. K. Palshikar
e-mail: gk.palshikar@tcs.com

P. Bhattacharyya
Indian Institute of Technology Bombay, Mumbai, India
e-mail: pb@cse.iitb.ac.in

V. Varma
International Institute of Technology Hyderabad, Hyderabad, India
e-mail: vv@iiit.ac.in

edge from this text is key to find solutions to pressing problems in the medical domain such as cancer [12]. Given the scale of this text, it is important to extract this knowledge automatically and store in some machine-readable knowledge representation (e.g., tables or graphs) so that the knowledge can be indexed, queried, analyzed, or inferred further to generate new knowledge. The state-of-the-art text mining systems are error prone due to the challenges and complexity of Natural Language Processing (NLP). Efficient text mining algorithms will help to improve the knowledge extraction from biomedical text and will be of real value in developing knowledge discovery graphs, QA systems, and knowledge summarization.

Cause–effect relations that denote causal dynamics between entities (e.g., **Bortezomib causes proteasome, Shp-2 is upregulated by p210 bcr/abl oncoprotein**), capture critical knowledge about the domain. Such knowledge can be utilized for correctly answering questions like—**What are the causes for apoptosis of Kasumi-1 cells?**, **Which drugs given to leukaemia patients cause anemia as a side effect?**, and **Tell me all causes for cytotoxicity in tumor cells.**

While popular biomedical knowledge bases such as SemMedDB contain causal predicates such as CAUSES, INHIBITS, STIMULATES, etc., from biomedical papers, they are not able to capture all the causal relations. In this paper, we present an approach to augment the causal predications of SemMedDB through a knowledge-based method by extracting more CE relations. Our proposed approach is a combination of an unsupervised machine learning technique to discover causal triggers and a set of high-precision linguistic rules to identify cause/effect arguments of these causal triggers. We also use simple rules to extract additional arguments of the CE relations: *negation* and *uncertainty*. In our experiments with a Leukaemia-related subset of 58,761 PubMed citations, our approach is able to extract 152,655 cause–effect triplets whereas for the same subset of citations, SemMedDB has only 77,135 causal predications. Also, the precision of our CE triplets which are extracted over and above SemMedDB was evaluated to be around 60% using a random subset.

2 Cause–Effect Relation Extraction

We represent a cause–effect relation mentioned in the form of a triplet which consists of

- **Causal trigger:** A multi-word expression or a verb which invokes a cause–effect relation, e.g., **because**, **due to**, **causes**, and **inhibits**.
- **Cause phrase:** A noun or verb phrase which represents a *cause* argument of the cause–effect relation invoked by the causal trigger.
- **Effect phrase:** A noun or verb phrase which represents an *effect* argument of the cause–effect relation invoked by the causal trigger.

Consider the following sentence: **MMuLV infection of non-transgenic mice-induced primarily mature T-cell lymphomas**. For this sentence, fol-

lowing CE triplet is extracted:

(Causal trigger: **induced**, Cause phrase: **MMuLV infection of non-transgenic mice**, Effect phrase: **primarily mature T-cell lymphomas**)

Here, the headwords of the cause-and-effect phrases are underlined. Intuitively, the headword of a phrase is its most important word, and grammatically it is the ancestor of all the words in a phrase in the sentence’s dependency parse tree.¹ Identification of the headword of a cause/effect phrase is an important step in our proposed approach.

2.1 Proposed Approach

We propose an algorithm for extracting CE relation triplets which works in following phases.

2.1.1 Causal Trigger Identification

In this first phase, a set of causal triggers (words or multi-word expressions) are identified in a given sentence. We observed that the causal triggers can be domain-agnostic (e.g., **due to**, **because**, **caused**) or domain-specific (e.g., **inhibits**, **down-regulated**). For domain-agnostic causal triggers, we used a list proposed by Girju [2]. Since ensuring the correctness and completeness of causal triggers for the biomedical domain requires a lot of domain knowledge, it was not feasible for us to manually compile such a list. Instead, we employed an unsupervised technique for automatically discovering the domain-specific causal verbs as described in [9]. This technique creates a large list of causal verbs in biomedical domain, using only unlabeled domain corpus and does not require any manual supervision. However, once the list is created, we manually curated it to retain only the high-precision causal verbs. The final list consisted of 109 domain-specific causal triggers and 33 domain-agnostic causal triggers.² All the morphological variations (e.g., **induced**, **inducing**) and nominal forms (e.g., **induction**) of the causal verbs are also considered. Given any input sentence, this list is looked up for identifying *candidate* causal triggers. These are referred to as candidate causal triggers because they become a part of a complete CE triplet only if both of the cause-and-effect arguments are also identified in the same sentence.

¹ Throughout the paper, we have used dependency relation types as per SpaCy. For detailed explanation of each dependency type, please refer to: https://github.com/clir/clearnlp-guidelines/blob/master/md/specifications/dependency_labels.md.

² Included in the Appendix.

2.1.2 Cause/Effect Headword Identification

In the second phase of our algorithm, for each candidate causal trigger v , headwords of its cause-and-effect argument phrases are identified. All the words in a given sentence which satisfy following conditions are identified as candidate headwords of a cause/effect phrase:

- All the verbs in the sentence which are not auxiliary of any other main verb, i.e., all the verbs whose dependency relation with their parent is not *aux*.
- All the nouns in the sentence which are headwords of any base noun phrase,³ i.e., all the nouns whose dependency relation with their parent is not *compound*.
- All other words which play a noun-like role in the dependency parse tree, i.e., the words whose dependency relation with their parents is one of the following: *nsubj* (nominal subject), *nsubjpass* (passive nominal subject), *doobj* (direct object), and *pobj* (prepositional object).

For each pair of a causal trigger v and a candidate cause/effect headword u in a given sentence, various types of features are generated. Table 1 describes various types of features with the help of an example $\langle v, u \rangle$ pair. These features are designed to capture various lexical and syntactic characteristics about how any causal trigger v and its corresponding cause/effect argument headword u are mentioned in a given sentence. Then these pairs are classified such that each pair is labeled with any one of the following classes: (i) CAUSE (indicating that u is the headword of the “cause” argument of the causal trigger v), (ii) EFFECT (indicating that u is the headword of the “effect” argument of the causal trigger v), and (iii) OTHER (indicating that u is not a cause/effect argument of the causal trigger v). If manually annotated $\langle v, u \rangle$ pairs are available, then any supervised classifier can be trained for identifying the above classes, using the features described in Table 1. However, our goal was to build a system which requires little or no supervision because creation of such an annotated dataset is quite time and effort intensive. Hence, using the same set of features, we employed a decision list algorithm where the rules in the decision list are designed manually. Each rule consists of three sets of features (described in Table 1) as follows:

- **AND set:** This set should be non-empty and each feature in this set should be present in the features set associated with any $\langle v, u \rangle$ pair for the rule to be satisfied. Here, v is a causal trigger and u is a candidate headword of a cause/effect argument.
- **OR set:** This set may be empty but if it is non-empty, then at least one feature from this set should be present in the features set associated with any $\langle v, u \rangle$ pair for the rule to be satisfied.
- **NEG set:** This set may be empty but if it is non-empty, then none of features from this set should be present in the features set associated with any $\langle v, u \rangle$ pair for the rule to be satisfied.

³ A base noun phrase is the noun phrase which does not contain any other noun phrase within it.

Table 1 Various features generated for a pair of a candidate causal trigger and a candidate cause/effect headword

Sentence: **Three long-term T-cell lines, established from peripheral blood mononuclear cell cultures from three STLV-1-seropositive monkeys, produced HTLV-1 Gag and Env antigens and retroviral particles**

Candidate causal trigger: $v =$ **produced**; **Candidate cause/effect headword:** $u =$ **lines**

Lexical features:

- Actual word tokens corresponding to v and u : $v.text$.**produced**, $u.text$.**lines**
 - Rootwords (lemmas) of words corresponding to v and u : $v.rootword$.**produce**, $u.rootword$.**line**
-

POS tag-based features:

- Part-of-speech (POS) tags of the words corresponding to v and u : $v.POS$.*VBD*, $u.POS$.*NNS*
 - Generalized POS tags of the words corresponding to v and u : $v.POS_gen$.*VERB*, $u.POS_gen$.*NOUN*
-

Dependency-based features:

- Parents/governors in the dependency parse tree for v and u : $v.parent$.*text.root*, $u.parent$.*text.produced*
- Dependency relation with parent in the dependency parse tree for v and u : $v.parent$.*dep.root*, $u.parent$.*dep.nsubj*
- Whether v is an ancestor of u in the dependency parse tree: *ancestor.v.u*
- Rootword of the “Lowest Common Ancestor” (LCA) of u and v in the dependency parse tree: *LCA.root_word.produce*
- Complete path of dependency relations from u to v in the sentence’s dependency parse tree: *dep.path.u<nsubj<v*

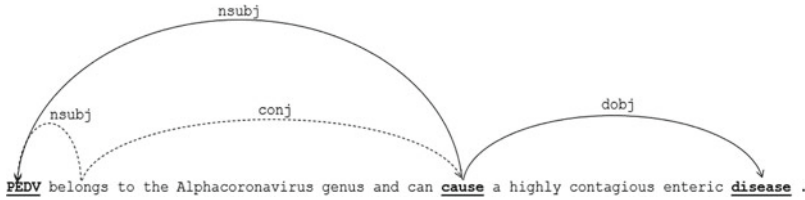
(Note: “<DR<” denotes an edge in the dependency tree labeled with dependency relation DR where the child is on the left and the parent is on the right. Similarly, “>DR>” denotes an edge where parent on the left and the child on the right.)

- Whether there is a direct edge between u and v in the dependency parse tree: *edge.v.u.nsubj*
 - Whether any particular dependency relation type lies on the dependency path connecting u to v : *path.v.u.nsubj*
 - Whether any particular word lies on the dependency path connecting u to v : NA (because, here u is directly connected to v)
-

Overall set of features associated with $\langle v=$ **produced, $u=$ **lines** $\rangle =$ { $v.text$.**produced**, $u.text$.**lines**, $v.rootword$.**produce**, $u.rootword$.**line**, $v.POS$.*VBD*, $u.POS$.*NNS*, $v.POS_gen$.*VERB*, $u.POS_gen$.*NOUN*, $v.parent$.*text.root*, $u.parent$.*text.produced*, $v.parent$.*dep.root*, $u.parent$.*dep.nsubj*, *ancestor.v.u*, *LCA.root_word.produce*, *dep.path.u<nsubj<v*, *edge.v.u.nsubj*, *path.v.u.nsubj*}**

Table 2 An illustration of application of rules to identify headword described in Tables 4 and 5

PEDV belongs to the Alphacoronavirus genus and can cause a highly contagious enteric disease



$\langle v = \text{cause}, u = \text{PEDV} \rangle$: CAUSE (Rule ID 1 in Table 4)

$\langle v = \text{cause}, u = \text{disease} \rangle$: EFFECT (Rule ID 1 in Table 5)

In other words, each rule must specify a conjunction of some features. Optionally, it may specify other sets of features in addition to the conjunction, which behave as a disjunction or negation of some other features. For example, consider the following rules:

- One of the rules to identify CAUSE is

AND: $\{ \text{dep.path.u} < \text{nsbj} < v \}$;

OR: $\{ u.\text{POS_gen.NOUN}, u.\text{POS_gen.PROPN} \}$;

NEG: $\{ v.\text{rootword.result} \}$

This rule will be satisfied for a $\langle v, u \rangle$ pair if the dependency path from the candidate headword u to the causal trigger v is $u < \text{nsbj} < v$ (i.e., u is a nominal subject of v). In addition, the generalized POS tag of the candidate headword u should be either *NOUN* (common noun) or *PROPN* (proper noun). Also, the trigger v should not be any morphological variation of **result**.

- Another rule to identify CAUSE is

AND: $\{ v.\text{POS_gen.VERB}, \text{dep.path.u} > \text{relcl} > v, v.\text{child.which} \}$; **OR:** $\{ \}$;

NEG: $\{ u.\text{parent.dep.attr} \}$

This rule will be satisfied for a $\langle v, u \rangle$ pair if the causal trigger v is a verb, the dependency path from u to v is $u > \text{relcl} > v$ (i.e., a relative clause headed at v modifies u), and **which** is one of the children of v . Moreover, the dependency relation of u with its parent should NOT be *attr*. There are no disjunctive features for this rule.

Our decision list has 33 and 21 rules for identifying CAUSE and EFFECT headwords, respectively. Tables 4 and 5 show the most important rules for identifying CAUSE and EFFECT headwords, respectively.⁴ Table 2 illustrates the application of these rules for an example sentence.

2.1.3 Cause/Effect Phrase Expansion

In the third phase of our algorithm, the cause/effect headwords identified in the second phase are expanded to get the complete phrases. Here, we again use dependency

⁴ All rules are included in the Appendix.

tree and simply consider the complete subtree rooted under a given headword. For example, in Table 2, the headword **disease** is expanded using the subtree rooted at **disease** which is a **highly contagious enteric disease**. We also use a simple rules specification language for specifying some exceptions for phrase expansion, as follows:

- An option to exclude children/dependents having certain dependency relations with their parents. In our experiments, we exclude following dependency relations—*punct* (connects to punctuation symbols), *appos* (connects to an appositive phrase), and *advcl* (connects to an adverbial clause).
- An option to limit the phrase boundaries to the left/right of the trigger word. If the trigger word itself is a descendent of a cause/effect headword, then the cause/effect phrase boundaries will never include the trigger word and exceed beyond it.

2.1.4 Cause–Effect Triplet Formation

In this final phase, CE triplets are formed for each trigger v . Let U_C be the set of CAUSE headwords identified for v and U_E be the set of EFFECT headwords identified for v . Then, the final set of CE triplets associated with the trigger v is $\{ \langle u_1, v, u_2 \rangle \mid \text{s.t. } (u_1, u_2) \in U_C \times U_E \}$.

Thus, the final CE triplet for the sentence in Table 2 will be $\langle \text{PEDV, cause, a highly contagious enteric disease} \rangle$.

2.2 Extraction of Additional Arguments

In addition to cause-and-effect arguments, we extract two more arguments of a cause–effect relation: (i) negation and (ii) uncertainty. If the causal trigger has a child in its dependency tree with dependency relation *neg*, then we extract it as a *negation* argument. For example, [**Overnight incubation with 1 microM safrole**]_{Cause} did [**not**]_{Negation} [**alter**]_{Trigger} [cell **proliferation**]_{Effect}. Here, the causal trigger is **alter** is negated by **not** which is extracted as a *negation* argument.

Similarly, if the causal trigger has a child in its dependency tree with dependency relation *aux* and it is from a set of uncertainty indicating words (such as **may, might, would**), then we extract it as an *uncertainty* argument. For example, [**Glucocorticoids**]_{Cause} [**might**]_{Uncertainty} [**induce**]_{Trigger} [**the apoptosis of some types of AML cells**]_{Effect}, **just like that of some lymphoid leukemia cells**. Here, the causal trigger is **induce** is modified by **might** which is extracted as an *uncertainty* argument.

3 Related Work

Extensive research is going on in the field of text mining in the biomedical domain. Several interesting approaches have been proposed to extract named entities, relationships, summaries, text classification, ontologies, knowledge discovery graphs, and hypothesis generation [6].

A lot of approaches for biomedical relation extraction use some form of Named Entity Recognition (NER) to identify the medical concepts in a sentence first and then find relations between them using rule-based techniques, machine learning algorithms, or a combination of both [8, 10, 11, 13]. Identifying named entities in the first step helps to reduce the number of candidates for relation identification. However, the extraction using this technique suffers if the medical concept is not picked up or identified incorrectly by the NER module. Our technique is different from these techniques as we do not rely on NER for our CE triplet extraction. As a result we are able to extract CE triplets even in those cases when a named entity can be missed by the NER system. The most similar line of research to our work is SemRep and its corresponding knowledge base SemMedDB.

3.1 *SemRep and SemMedDB*

SemRep [7] is a UMLS-based⁵ program that extracts three-part semantic predications, from sentences in biomedical text. These predications are in the following form:

(subject, RELATION, object)

The subject and object arguments are UMLS Metathesaurus Concepts [1], while the relation between them is from the UMLS Semantic Network. SemRep works by identifying UMLS Concepts present in a sentence and then using rules to determine which UMLS relation exists between them. The Semantic MEDLINE Database (SemMedDB) [5] is a repository of predications extracted from all of PubMed citations using SemRep. SemRep extracts relations of various types but only a few of them represent causal relations, which are AFFECTS, CAUSES, STIMULATES, INHIBITS, DISRUPTS, PRODUCES, PRECEDES, COMPLICATES, PREDISPOSES, PREVENTS. Although our approach does not explicitly identify relation types, it extracts all types of cause–effect relations, and the finer interpretation of each relation can be inferred from the corresponding causal trigger.

4 Experimental Analysis

We evaluate our proposed approach for cause–effect relation extraction along multiple aspects.

4.1 *Leukaemia-Related PubMed Abstracts Corpus*

We created a corpus of 58,761 PubMed citations (title as well as abstract) which are related to Leukaemia. We pre-processed this dataset using SpaCy [3] to obtain—tokens, sentences, and dependency parse trees for each sentence. Overall, the dataset consisted of 568,528 sentences, i.e., on an average each citation contains 9.67 sentences. Also, the median sentence length is 27 words, indicating that the sentences in the corpus are fairly complex.

We applied our cause–effect relation extraction approach on this dataset to extract cause–effect (CE) triplets. Overall, 152,655 CE triplets were extracted. As there are no gold-standard annotations available for this data, we use random sampling to estimate precision. The detailed evaluation methods are described in following sections.

⁵ <https://www.nlm.nih.gov/research/umls/index.html>.

Table 3 Examples of scores assigned by the human expert for evaluating headword identification rules. The candidate trigger v and the candidate headword u are underlined and marked within the sentence

Sentence with a (v, u) pair marked inline	RuleID (C/E)	Score	Comment
In particular, we reported the existence of BCR-ABL alternative splicing isoforms, in about 80% of Philadelphia-positive [patients] u , which [lead] v to the expression of aberrant proteins	8 (CAUSE)	0	Incorrect (due to incorrect parsing)
Childhood acute myeloid leukemia with bone marrow eosinophilia [caused] v by [t(16) u ; 21](q24 ; q22)	3 (CAUSE)	1	Partially correct (due to incorrect tokenization)
Perifosine and TRAIL synergized to activate caspase-8 and induce apoptosis, which was [blocked] v by a caspase-8-selective [inhibitor] u	3 (CAUSE)	2	Correct
Monocytic [maturation] u (morphologic and immunologic) was [induced] v in all cases studied, although to different rates, by TNF-alpha and by HTR-9 incubation	3 (EFFECT)	2	Correct

4.1.1 Evaluation of Cause/Effect Headword Identification Rules

We estimate the precision of each linguistic rule used in our decision list to identify cause/effect headwords. Our decision list is used to classify a pair of a causal trigger and a candidate headword of a cause/effect phrase into three different classes—CAUSE, EFFECT, and OTHER. As described earlier, there are 33 rules to identify CAUSE and 21 rules to identify EFFECT. To estimate precision of a CAUSE-predicting rule, we randomly select 10 CE triplets from the 152,655 CE triplets such that the cause headword in those triplets was identified using that particular rule. We also ensured that our random sample contains equal mixture of simple and complex sentences. Out of the 10 CE triplets, we ensure that 5 of them are extracted from a sentence whose length is more than 27 words (median length) and the rest 5 are shorter than 27 words. A human expert then evaluated these triplets manually to assign a score to each triplet. The scale used to evaluate is—0 if completely incorrect, 1 if partially correct, and 2 if completely correct (see Table 3, for example). The precision for the rule is computed by dividing the total score for 10 CE triplets by 20. Table 4 shows precision as well as *coverage* for 10 rules for identifying CAUSE headword. Similarly, Table 5 shows precision as well as coverage for 10 rules for identifying EFFECT headword. In both the cases, the tables show only top 10 rules as per their coverage. Here, the coverage is the total number of extracted CE triplets for which a certain rule was used for identifying a CAUSE or EFFECT headword.

Table 4 Performance of the rules identifying CAUSE headwords (*v* is a candidate causal trigger and *u* is a candidate headword of a cause phrase)

Rule	Rule	Coverage	Prec.	#CE Triplets ∉ SMDB-L
1	AND: { <i>dep.path.u</i> <nsubj< <i>v</i> }; OR: { <i>u.POS_gen.NOUN</i> , <i>u.POS_gen.PROPN</i> }; NEG: { <i>v.rootword.result</i> }	69,041 (45.2%)	0.8	21,104
2	AND: { <i>dep.path.u</i> <npadvmod< <i>v</i> }	20,263 (13.3%)	0.8	6128
3	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> <pobj<agent< <i>v</i> }	15,113 (9.9%)	0.95	4123
4	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> >reicl> <i>v</i> , <i>v.child.that</i> }	5534 (3.6%)	0.85	1564
5	AND: { <i>v.text.due</i> , <i>dep.path.u</i> <pobj< <i>v</i> }	4012 (2.6%)	0.7	1252
6	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> <nsubj<LCA>prep>pcomp> <i>v</i> }	3157 (2.1%)	0.6	931
7	AND: { <i>v.POS_gen.NOUN</i> , <i>dep.path.u</i> <pobj<prep< <i>v</i> , <i>path.by</i> }	2900 (1.9%)	0.95	845
8	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> >reicl> <i>v</i> , <i>v.child.which</i> }; NEG: { <i>u.parent.dep.attr</i> }	2889 (1.9%)	0.6	852
9	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> <nsubjpass<LCA>xcomp> <i>v</i> }	2836 (1.8%)	1.0	880
10	AND: { <i>v.rootword.role</i> , <i>dep.path.u</i> <nsubj<LCA>dobj> <i>v</i> , <i>lca.rootword.play</i> }	2679 (1.7%)	0.9	551

4.1.2 Evaluation of Phrase Expansion Rules

In order to evaluate the performance of the phrase expansion step, the human expert also evaluated correctness of phrases along with annotations obtained for headword identification rules (10 random CE triplets for each rule as explained earlier). The same scale of 0–2 was used here as was used in case of headword identification. The phrase expansion accuracy was observed to be 95.29% and 94.27% for CAUSE and EFFECT phrases, respectively. For computing this accuracy, only those phrases are considered whose headwords were identified correctly.

4.1.3 Comparison with SemMedDB

We applied our proposed approach on this corpus of 58,761 Leukaemia-related PubMed citations and obtained 152,655 cause–effect triplets. For comparison, we also considered a subset of SemMedDB for the same set of 58,761 PubMed citations, which we refer to as *SMDB-L*. Out of 503,183 predications in *SMDB-L*, only 77,135 correspond to the causal predicates.

Our proposed approach is able to extract almost twice the number of CE triplets as compared to *SMDB-L* (152,655 vs. 77,135). We estimate the precision of our CE triplets which are extracted over

Table 5 Performance of the rules identifying EFFECT headwords (*v* is a candidate causal trigger and *u* is a candidate headword of an effect phrase)

Rule ID	Rule	Coverage	Prec.	#CE Triplets \notin SMDB-L
1	AND: { <i>edge.v.u.dobj</i> }	84,072 (55.1%)	0.8	25,783
2	AND: { <i>v.POS_gen.VERB, dep.path.u>amod>v</i> }	18,972 (12.4%)	1.0	5668
3	AND: { <i>edge.v.u.nsubjpass</i> }	9785 (6.4%)	0.9	2692
4	AND: { <i>v.POS.VBN, edge.u.v.acl, u.POS_gen.NOUN, v.POS_gen.VERB</i> }	6300 (4.1%)	1.0	1750
5	AND: { <i>dep.path.u>prep>v</i> }; OR: { <i>v.text.because, v.text.due</i> }	6223 (4.1%)	0.8	1996
6	AND: { <i>v.POS_gen.VERB, v.rootword.lead, dep.path.u<pobj<prep<v, path.to</i> }	4897 (3.2%)	1.0	1441
7	AND: { <i>v.POS_gen.NOUN, dep.path.u<pobj<prep<v, path.of</i> }; OR: { <i>v.rootword.cause, v.rootword.reason, v.child.prep.by, v.child.agent.by</i> }	3850 (2.5%)	1.0	1234
8	AND: { <i>v.POS_gen.VERB, v.rootword.contribute, dep.path.u<pobj<prep<v, path.to</i> }	3438 (2.3%)	1.0	933
9	AND: { <i>v.rootword.role, dep.path.u<pobj<prep<v, path.in</i> }	2365 (1.5%)	1.0	463
10	AND: { <i>dep.path.u>advcl>mark>v, v.text.because</i> }	2128 (1.4%)	0.8	853

Table 6 Examples of scores assigned by the human expert for evaluating CE triplets. The trigger and cause/effect phrases are marked inline in the sentences. The headwords of cause/effect phrases are underlined. Note that each row represents one CE triplet and there may be more CE triplets extracted in the same sentence

Sentence with a CE triplet marked inline	Score	Comment
Furthermore, [the current diagnostic interpretation of flow cytometry readouts] <i>Effect</i> is [influenced] <i>Trigger</i> arbitrarily by individual experience and [knowledge] <i>Cause</i>	0	Incorrect (not causal)
[CsA treatment] <i>Cause</i> [resulted] <i>Trigger</i> in [an increased incidence of hyperbilirubinemia, which rapidly reversed upon conclusion of drug therapy] <i>Effect</i>	1	Partially correct (due to long effect phrase)
We conclude that [GM-CSF] <i>Cause</i> is effective in improving CLL associated chronic neutropenia and also [enhances] <i>Trigger</i> [impaired granulocyte chemiluminescence] <i>Effect</i> .	2	Correct

Table 7 Precision of extracted CE triplets which is estimated by manually evaluating random 100 triplets

Type of CE triplets	Precision (strict)	Precision (lenient)
CE triplets \notin SMDB-L	0.60	0.74

Table 8 Performance of our proposed approach for cause–effect relation extraction as compared to SemRep over the gold-standard dataset

Approach	Precision	Recall	F1-measure
SemRep	58.78	29.84	39.59
Proposed	50.83	35.66	41.91

and above *SMDB-L* using a random sample. We randomly select 100 CE triplets which are extracted from those sentences for which *SMDB-L* does not any predication having a causal predicate. Here, the causal predicates are AFFECTS, CAUSES, STIMULATES, INHIBITS, DISRUPTS, PRODUCES, PRECEDES, COMPLICATES, PREDISPOSES, PREVENTS. Examples of non-causal predicates are PART_OF, TREATS, PROCESS_OF. A human expert then evaluated these randomly selected 100 CE triplets and assigned a score to each triplet. The same scale used earlier to evaluate headword identification rules is used, i.e., 0 if completely incorrect, 1 if partially correct, and 2 if completely correct (see examples in Table 6). We compute two precision values: (i) strict precision is computed by considering partially correct to be incorrect and (ii) lenient precision is computed by considering the sum of scores of all triplets divided by 200. Table 7 shows both of these precision values for the triplets extracted over and above *SMDB-L*. Moreover, the last column in Tables 11 and 10 shows the number of CE triplets extracted by each rule which are not part of *SMDB-L*. These CE triplets are extracted from sentences for which *SMDB-L* does not have any predication with a causal predicate.

4.2 Gold-Standard Dataset

We obtained from SemRep website,⁶ a gold-standard dataset [4] for semantic predications where subject/object arguments are manually annotated. This dataset contains 500 sentences annotated with 1371 semantic predications annotated by human experts. Out of these 1371 predications, only 258 correspond to causal predicates (the same list of predicates considered in *SMDB-L*). We ignored the non-causal predications as the scope of this work is limited to identifying only causal relations.

Each predication in this dataset consists of following important fields:

- Predicate/relation type.
- Subject name, its concept ID from ULMS Metathesaurus, and corresponding text span which is analogous to *Cause* phrase in our CE triplets.
- Object name, its concept ID from UMLS Metathesaurus, and corresponding text span which is analogous to *Effect* phrase in our CE triplets.

⁶ <https://semrep.nlm.nih.gov/GoldStandard.html>.

4.2.1 Baseline: SemRep

In order to compare the performance of our proposed approach on the gold-standard dataset with SemRep, we processed its 500 sentences using the online SemRep Batch Facility.⁷ We set the *Knowledge Source* and the *Lexicon Year as 2018* and selected the *Strict Model*. We obtained the *Full Fielded Model Output* because it provided the text spans of the subject and object in the sentence. These text spans of subject and object correspond to cause-and-effect phrases in our CE triplet format, respectively.

4.2.2 Evaluation

The format of our CE triplet is $\langle \textit{cause phrase}, \textit{trigger}, \textit{effect phrase} \rangle$, whereas each predication in the gold-standard dataset and SemRep output is considered in the form— $\langle \textit{subject text}, \textit{predicate}, \textit{object text} \rangle$. As we are considering only the causal predicates, we ignore the actual trigger/predicate type while evaluating. Hence, a CE triplet is considered to be *matching* a gold-standard predication if there is an overlap of at least one content word between the *cause phrase* and the *subject text* as well as between the *effect phrase* and the *object text*. For example, consider a gold-standard predication $\langle \text{Calcitonin gene-related peptide, INHIBITS, monocyte chemoattractant protein-1} \rangle$. We consider it to be matching with our predicted CE triplet $\langle \text{Calcitonin gene-related peptide, inhibits, interleukin-1beta-induced endogenous monocyte chemoattractant protein-1 secretion in type II alveolar epithelial cells} \rangle$.

Similarly, a SemRep predication is considered to be *matching* a gold-standard predication if there is an overlap of at least one content word between the corresponding *subject texts* and between the corresponding *object texts*. For each gold-standard predication, a true positive (TP) is counted for our proposed approach if there is a *matching* CE triplet and a false negative (FN) is counted otherwise. Also, for each predicted CE triplet, a false positive (FP) is counted if there is no *matching* predication in the gold-standard dataset. Precision, Recall, and F1-measure are then computed as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F1 = \frac{2 \cdot P \cdot R}{P + R}. \quad (1)$$

For the SemRep baseline, the evaluation is carried out in a similar manner. Table 8 shows the performance of our proposed technique as compared with the SemRep baseline.

We analyzed our extracted CE triplets for low precision. One of the reasons is that unlike SemRep, our technique does not restrict cause/effect phrases to be of certain “semantic types”. Hence, we have several False Positive extractions where cause/effect phrases may not be normalized to any UMLS concept. For example, consider the sentence: **IFN-alpha profoundly alters cytoskeletal organization of hairy cells and causes reversion of the hairy appearance into a rounded morphology**. Here, we extract the following CE triplet which is correct but not annotated in the gold-standard dataset and hence is counted as a False Positive: $\langle \text{IFN-alpha, causes, reversion of the hairy appearance into a rounded morphology} \rangle$.

⁷ https://ii.nlm.nih.gov/Batch/UTS_Required/semrep.shtml.

5 Conclusion and Future Work

We proposed a knowledge-based cause–effect relation extraction approach which does not require any kind of supervision or large annotated corpus. The proposed approach is based on: (i) an unsupervised machine learning technique to discover causal triggers and (ii) a set of high-precision linguistic rules to identify cause/effect arguments of these causal triggers. We evaluated our approach using a large corpus of 58,761 Leukaemia-related PubMed abstracts consisting of 568,528 sentences. We extracted 152,655 CE triplets from this corpus where each triplet consists of a cause phrase, an effect phrase, and a causal trigger. As compared to the existing knowledge base—SemMedDB [5]—the number of extractions is almost twice, i.e., for the same set of PubMed abstracts, SemMedDB has only 77,135 predications corresponding to causal predicates. In addition, we also evaluated our approach on a gold-standard annotated corpus of 500 sentences and it outperformed the existing technique SemRep [7] on this corpus. In future, we plan to extend this work in several aspects:

- Sentence simplification: Upon analysis of errors, we observed that a large fraction of errors are occurring due to incorrect dependency parsing. Hence, even if our linguistic rules are correct, due to incorrect dependency paths we get wrong extractions. A major reason behind this is that the sentences in biomedical domain are more complex (median length of 27 words) than the general domain sentences (on which dependency parsers are generally trained). Hence, it would be interesting to simplify sentences before parsing using any sentence simplification tool.
- Distant supervision: Even though our linguistic rules are high precision, it is difficult to scale-up the recall and maintain these rules. Hence, we plan to use these rules to automatically create a large annotated dataset of cause–effect relations with possibly noisy annotations and then train a supervised machine learning model using this dataset.
- Use of sophisticated knowledge validation techniques to automatically improve the quality of extracted CE triplets.

Appendix

Causal Triggers Table 9 shows the complete list of domain-agnostic and domain-specific causal triggers.

Rules for Cause/Effect Headword Identification

Our decision list has 33 and 21 rules for identifying CAUSE and EFFECT headwords, respectively. Tables 11 and 10 show all these rules for identifying CAUSE and EFFECT headwords, respectively.

Table 9 List of cue-phrases and causal verbs used in the first phase (causal trigger identification)

Domain-agnostic cue-phrases and causal verbs:	
cause of, causes of, cause for, causes for, eason for, reasons for, reason of, reasons of, as a consequence, as a result, due, because, activate, bring about, cause, contribute to, create, derive from, effect, elicit, entail, evoke, generate, give rise to, implicate in, lead to, originate in, rovoke, result from, stem from, stimulate, trigger off, role	
Domain-specific causal verbs:	
coadministrate, down-regulate, up-regulate, co-express, re-express, over-express, dysregulate, degranulate, knockdown, ablate, abrogate, accelerate, advance, affect, alter, attenuate, benefit by, benefit from, block, convert, decrease, degrade, delineate, deplete, deregulate, die of, diminish, discharge, disrupt, disseminate, divide, elevate, eliminate, enforce, enhance, enrich, eradicate, exacerbate, exert, expand, extend, fuse, govern, impact, impair, improve, increase, induce, infect, infiltrate, influence, inhibit, inject, intensify, kill, knock down, maximize, mediate, minimize, optimize, originate from, portend, prevent, produce, proliferate, prolong, protect, reactivate, reduce, regain, regulate, relapse, remove, replicate, repress, reproduce, rescue, restore, reverse, revert, sensitize, shorten, stabilize, substitute, suppress, transfer, transform, trigger, transplant, escalate, complicate, express, progress, decline, predispose, translate, secrete, unblock, grow, remit, remove, abolish, drive, modulate, amplify, antagonize, destruct, destroy, lower	

Table 10 The rules for identifying EFFECT headwords (*v* is a candidate causal trigger and *u* is a candidate headword of an effect phrase)

ID	Rule
1	AND: { <i>edge.v.u.dobj</i> }
2	AND: { <i>v.POS_gen.VERB, dep.path.u>amod>v</i> }
3	AND: { <i>edge.v.u.nsubjpass</i> }
4	AND: { <i>v.POS.VBN, edge.u.v.acl, u.POS_gen.NOUN, v.POS_gen.VERB</i> }
5	AND: { <i>dep.path.u>prep>v</i> }; OR: { <i>v.text.because, v.text.due</i> }
6	AND: { <i>v.POS_gen.VERB, v.rootword.lead, dep.path.u<pobj<prep<v, path.to</i> }
7	AND: { <i>v.POS_gen.NOUN, dep.path.u<pobj<prep<v, path.of</i> }; OR: { <i>v.rootword.cause, v.rootword.reason, v.child.prep.by, v.child.agent.by</i> }
8	AND: { <i>v.POS_gen.VERB, v.rootword.contribute, dep.path.u<pobj<prep<v, path.to</i> }
9	AND: { <i>v.rootword.role, dep.path.u<pobj<prep<v, path.in</i> }
10	AND: { <i>dep.path.u>advcl>mark>v, v.text.because</i> }
11	AND: { <i>v.rootword.result, v.child.from, dep.path.u<nsubj<v</i> }
12	AND: { <i>v.rootword.role, dep.path.u<pcomp<prep<v, path.in</i> }
13	AND: { <i>v.rootword.result, v.child.in</i> }; OR: { <i>dep.path.u<pcomp<prep<v, dep.path.u<pobj<prep<v</i> }
14	AND: { <i>v.POS_gen.NOUN, dep.path.u<pobj<prep<v, dep.path.len.1.for>pobj>u, v.rootword.cause</i> }
15	AND: { <i>v.POS_gen.NOUN, dep.path.u<pcomp<prep<v, dep.path.len.1.of>pcomp>u</i> }; OR: { <i>v.rootword.cause, v.rootword.reason</i> }
16	AND: { <i>v.POS.VBN, dep.path.u<nsubj<LCA>attr>acl>v, v.POS_gen.VERB</i> }
17	AND: { <i>edge.u.v, lca.rootword.be, v.text.due, u.copula_verb_with_object</i> }
18	AND: { <i>v.rootword.die, dep.path.u<nsubj<v, v.child.of</i> }
19	AND: { <i>v.POS_gen.NOUN, dep.path.u<pobj<prep<v, dep.path.len.1.for>pobj>u, v.rootword.reason</i> }
20	AND: { <i>v.text.due, dep.path.u<nsubj<LCA>acomp>v, lca.rootword.be</i> }
21	AND: { <i>v.text.due, dep.path.u>amod>v</i> }

Table 11 The rules for identifying CAUSE headwords (*v* is a candidate causal trigger and *u* is a candidate headword of a cause phrase)

ID	Rule
1	AND: { <i>dep.path.u</i> <nsubj< <i>v</i> }; OR: { <i>u.POS_gen.NOUN</i> , <i>u.POS_gen.PROPN</i> }; NEG: { <i>v.rootword.result</i> }
2	AND: { <i>dep.path.u</i> <npadvmod< <i>v</i> }
3	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> <pobj<agent< <i>v</i> }
4	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> >recl< <i>v</i> , <i>v.child.that</i> }
5	AND: { <i>v.text.due</i> , <i>dep.path.u</i> <pobj< <i>v</i> }
6	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> <nsubj<LCA>prep>pcomp> <i>v</i> }
7	AND: { <i>v.POS_gen.NOUN</i> , <i>dep.path.u</i> <pobj<prep< <i>v</i> , <i>path.by</i> }
8	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> >recl< <i>v</i> , <i>v.child.which</i> }; NEG: { <i>u.parent.dep.attr</i> }
9	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> <nsubjpass<LCA>xcomp> <i>v</i> }
10	AND: { <i>v.rootword.role</i> , <i>dep.path.u</i> <nsubj<LCA>dobj> <i>v</i> , <i>lca.rootword.play</i> }
11	AND: { <i>v.text.due</i> , <i>dep.path.u</i> <pobj<prep< <i>v</i> , <i>path.to</i> }
12	AND: { <i>v.POS_gen.NOUN</i> , <i>dep.path.u</i> <nsubj<LCA>attr> <i>v</i> , <i>v.rootword.cause</i> }
13	AND: { <i>v.rootword.cause</i> }; OR: { <i>dep.path.u</i> <nsubjpass<LCA>xcomp>attr> <i>v</i> , <i>dep.path.u</i> <nsubjpass<LCA>prep>pobj> <i>v</i> , <i>dep.path.u</i> <dobj<LCA>prep>pobj> <i>v</i> }
14	AND: { <i>v.POS_gen.VERB</i> , <i>v.POS.VBG</i> , <i>dep.path.u</i> <nsubj<LCA>attr>acl> <i>v</i> }
15	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> <nsubj< <i>v</i> , <i>u.POS_gen.VERB</i> }
16	AND: { <i>v.POS_gen.VERB</i> , <i>edge.v.u.csubj</i> }
17	AND: { <i>v.rootword.die</i> , <i>dep.path.len.1.of</i> >pobj> <i>u</i> , <i>dep.path.u</i> <pobj<prep< <i>v</i> }
18	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> <nsubj<LCA>xcomp> <i>v</i> }
19	AND: { <i>v.rootword.result</i> , <i>v.child.from</i> }; OR: { <i>dep.path.u</i> <pcomp<prep< <i>v</i> , <i>dep.path.u</i> <pobj<prep< <i>v</i> }
20	AND: { <i>dep.path.u</i> <nsubj< <i>v</i> , <i>v.rootword.result</i> , <i>v.child.in</i> }; OR: { <i>u.POS_gen.NOUN</i> , <i>u.POS_gen.PROPN</i> }; NEG: { <i>v.child.from</i> }
21	AND: { <i>v.POS_gen.NOUN</i> , <i>dep.path.u</i> <pobj<prep< <i>v</i> , <i>dep.path.len.1.for</i> >pobj> <i>u</i> }; OR: { <i>v.rootword.consequence</i> , <i>v.rootword.result</i> , <i>v.rootword.effect</i> }
22	AND: { <i>v.text.because</i> , <i>dep.path.u</i> <pobj< <i>v</i> }
23	AND: { <i>v.POS_gen.VERB</i> , <i>v.POS.VBG</i> , <i>dep.path.u</i> <nsubj<LCA>advcl> <i>v</i> }
24	AND: { <i>v.POS_gen.NOUN</i> , <i>dep.path.u</i> <nsubj<LCA>attr> <i>v</i> , <i>v.rootword.reason</i> }
25	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> <nsubj<LCA>acompl>prep>pcomp> <i>v</i> }
26	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> <nsubj<LCA>acompl>xcomp> <i>v</i> }
27	AND: { <i>v.text.because</i> , <i>dep.path.u</i> >mark> <i>v</i> }
28	AND: { <i>v.text.due</i> , <i>dep.path.u</i> <pobj<pcomp< <i>v</i> , <i>path.v.u.pcomp</i> , <i>path.v.u.pobj</i> , <i>ancestor.v.u</i> , <i>u.parent.dep.pobj</i> , <i>path.to</i> }
29	AND: { <i>v.POS_gen.VERB</i> , <i>dep.path.u</i> <nsubj<LCA>attr>recl> <i>v</i> , <i>u.POS_gen.NOUN</i> }
30	AND: { <i>v.POS_gen.NOUN</i> , <i>u.POS_gen.NOUN</i> , <i>dep.path.u</i> >appos> <i>v</i> }; OR: { <i>v.rootword.inhibitor</i> , <i>v.rootword.predictor</i> , <i>v.rootword.marker</i> , <i>v.rootword.cause</i> }
31	AND: { <i>v.POS_gen.NOUN</i> , <i>dep.path.u</i> <nsubj<LCA>attr> <i>v</i> }; OR: { <i>v.rootword.inhibitor</i> , <i>v.rootword.predictor</i> , <i>v.rootword.marker</i> , <i>v.rootword.cause</i> , <i>v.rootword.complication</i> }
32	AND: { <i>dep.path.u</i> <nsubj<LCA>xcomp>attr> <i>v</i> , <i>u.POS_gen.NOUN</i> , <i>u.rootword.cause</i> }
33	AND: { <i>dep.path.u</i> <csbjpass<LCA>xcomp> <i>v</i> }; NEG: { <i>u.POS_gen.PRON</i> }

References

1. Bodenreider O (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 32(suppl_1):D267–D270
2. Girju R (2003) Automatic detection of causal relations for question answering. In: Proceedings of the ACL 2003 workshop on multilingual summarization and question answering, vol 12. Association for Computational Linguistics, pp 76–83
3. Honnibal M, Montani I (2017) spaCy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. <https://spacy.io/>
4. Kilicoglu H, Roseblat G, Fiszman M, Rindflesch TC (2011) Constructing a semantic predication gold standard from the biomedical literature. *BMC Bioinform* 12(1):486
5. Kilicoglu H, Shin D, Fiszman M, Roseblat G, Rindflesch TC (2012) SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics* 28(23):3158–3160
6. Luque C, Luna JM, Luque M, Ventura S (2019) An advanced review on text mining in medicine. *Wiley Interdiscip Rev: Data Min Knowl Discov* 9(3):e1302
7. Rindflesch TC, Fiszman M (2003) The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform* 36(6):462–477
8. Rink B, Harabagiu S, Roberts K (2011) Automatic extraction of relations between medical concepts in clinical texts. *J Am Med Inform Assoc* 18(5):594–600
9. Sharma R, Palshikar G, Pawar S (2018) An unsupervised approach for cause-effect relation extraction from biomedical text. In: International conference on applications of natural language to information systems. Springer, pp 419–427
10. Uzuner O, Mailoa J, Ryan R, Sibanda T (2010) Semantic relations for problem-oriented medical records. *Artif Intell Med* 50(2):63–73
11. Rong X, Wang QQ (2015) Large-scale automatic extraction of side effects associated with targeted anticancer drugs from full-text oncological articles. *J Biomed Inform* 55:64–72
12. Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, Vongsangnak W, Shen B (2013) Biomedical text mining and its applications in cancer research. *J Biomed Inform* 46(2):200–211
13. Zhu X, Cherry C, Kiritchenko S, Martin J, De Bruijn B (2013) Detecting concept relations in clinical text: Insights from a state-of-the-art model. *J Biomed Inform* 46(2):275–285

NyOn: A Multilingual Modular Legal Ontology for Representing Court Judgements



Sarika Jain , Pooja Harde, and Nandana Mihindukulasooriya 

Abstract In this paper, we present the Nyaya Ontology (NyOn) that describes Indian Supreme Court Judgements. It is planned to be used to build legal Knowledge Graphs from Indian Supreme Court Judgements and in downstream tasks such as question answering. NyOn has been developed using a micro-level modular methodology and published adhering to the Semantic Web best practices and FAIR principles. It was evaluated using OOPS! Pitfall Scanner. The NyOn ontology contains 5 modules covering both criminal and civil courts and it is currently available in 5 languages.

Keywords Court judgements · Judicial system · Legal ontology

1 Introduction and Motivation

The Judicial System plays a vital role in¹ forming the society and maintaining the society affairs with the help of law and order. Every country has their own judicial structure and legislation that they follow. These judicial systems contain various documents and information which are called legal resources. Legal Resources in the Judicial System get created intermittently. Therefore it is necessary to properly maintain this data. The legal resources are of many different types viz., case documents, acts, legislation, constitution, court opinions, reports, events, amendments, gazettes, etc. Any legal resource is not considered completed as a whole single document or a resource. Each and every legal resource cites various different legal documents to present the information. Example, consider the legal document of any case. The case document cites various different documents like preceding cases documents,

¹<https://w3id.org/def/NyOnLegal#>.

S. Jain (✉) · P. Harde
National Institute of Technology Kurukshetra, Kurukshetra, India
e-mail: jasarika@nitkkr.ac.in

N. Mihindukulasooriya
IBM Research, Dublin, Republic of Ireland
e-mail: nandana@ibm.com

FIRs (if they belong to criminal case), Rules, Sections, Acts, etc. Therefore in such a scenario it is very important to gather all the information at one place that will help the domain experts and the domain related people to find the information with an ease at a single place for analyzing the data, decision making process, automation of documents, etc. To do this we require the good quality of metadata with good quantity for the different documents that are present in the judicial system, which will help for better processing. There exist many countries that use legal information systems for maintaining their legal data and make the same data available to the common public via their portals. Countries UK, US, and European Union have their own legal portals where the data is made available publicly to the public regarding any judicial work. Beyond these countries it becomes difficult to find the legal portals where the data is made available to the public and is maintained with the state-of-the-art. In India also, there are many portals that are available to search for the legal information. These systems are briefly discussed in the below section with their limitations. In developing countries like India, which is the largest democratic country in the world, the availability of legal information and making the same available to the legal domain experts, common people, bureaucrats is very important. There exists many international online legal research databases available. Some of them are Westlaw International,² HeinOnline,³ LexisNexis,⁴ JSTOR,⁵ Westlaw,⁶ Austrian Rechtsinformationssystem des Bundes (RIS),⁷ etc. The list of Indian Online Legal Research Databases are Manupatra,⁸ Indlaw,⁹ SCC Online,¹⁰ AIR—All India Reporter (AIR),¹¹ Corporate Law Advisor,¹² etc.

The challenges that exists for creating a court decision document are how to identify the entities from such large volume of documents, which vocabulary to use to represent different keywords as per the judges interpretation, kind of structure to follow for ontology development, methodology that best suits to develop such ontologies and so on. The main objective of the paper is to create a multilingual legal ontology for the Indian Court Decision documents. The contributions include creating a multilingual legal ontology by name NyOn in 5 different languages (English, French, Spanish, German, and Hindi), evaluating it using the OOPS! scanner, and publishing it over the web using w3id¹³ permanent ID.

² <https://www.westlawinternational.com/>.

³ <https://home.heinonline.org/>.

⁴ <https://www.lexisnexis.com/en-us/gateway.page>.

⁵ <https://www.jstor.org/>.

⁶ <https://legal.thomsonreuters.com/en/westlaw>.

⁷ <https://www.ris.bka.gv.at/>.

⁸ <https://www.manupatrafast.com>.

⁹ <http://crl.du.ac.in/sub.database/Indlaw.comhtm.htm>.

¹⁰ <https://www.sconline.com/>.

¹¹ <https://www.aironline.in/>.

¹² <https://www.claonline.in/>.

¹³ <https://w3id.org/>.

2 Current Judicial System Scenario in India

According to the cited data from National Judicial Data Grid¹⁴ and Supreme Court¹⁵ there are in total 4.5 million cases pending in India. From which 4.1 crore cases (as on 13.02.2022) pending in the district and subordinate courts, 56.7 lakh cases (as on 13.02.2022) in the various high courts, and 70,101 cases (as on 03.02.2022) in the Supreme Court. According to the Ministry of Law and Justice of India, there are 21.03 Judges per million population. These figures are huge problems as it delays timely justice to the citizens of the country. On the top of the same, searching for legal information, citations of the case, preceding case information, reading case documents for relevant information findings makes it more time consuming and overhead to the domain experts. Considering all the above scenarios, it has become very important to make the legal research information available to the legal experts, in a precise format with all the preceding information related to the case, citations, elements of the cases, participants involved, and relation of all these entities which will save their time for studying the cases.

Complexities finding in the above systems While studying the above mentioned databases we encounter various complexities that are present in these databases. First, despite the fact that most of these web portals or databases provide good depth of linking of the documents for legal research, their annual subscriptions are too expensive for the common people and sometimes also for the domain expert people to use such portals on a daily basis. Second, although there are many free legal research databases or web portals available in different countries including India, the documents are not at all linked with other documents. One has to search for the different documents to meet their requirements. Sometimes the search results also do not provide the complete list of the available documents due to the poor metadata of the documents. Thus missing links of the documents makes the tasks of the domain related people time-consuming and tedious. As the legal domain is the sector where extensive documents are created in the regular intervals, it becomes very important to link such data or information together for better maintainability and the availability for the reuse or referring purpose. This scenario motivates us to create a metadata in a good quantity which will help us to bind the legal documents together.

Proposed Solution Much work has been done till date with the help of Semantic Web in the Legal Domain. Semantic Web helps to convert the data (unstructured, semi-structured, etc.) in the machine readable format which helps for better data processing. As the legal domain contains huge amounts of data that need to be maintained and also the same amount of large data gets created in the regular intervals, researchers found the semantic web the most profound area for research in the legal domain. It is a complex task to interlink all the required legal documents internally and make them available to the machines in the machine-readable format to generate the expected outputs from the system. Semantic Web can help in various legal proceedings like advanced case law search engines, online dispute resolution, assistance

¹⁴ <https://njdg.ecourts.gov.in/>.

¹⁵ <https://main.sci.gov.in/statistics>.

in drafting needs, analysis that is predictive, categorisation of contracts according to different criteria and detection of divergent or incompatible contractual clauses, etc.

The legal documents we are considering for now are the court decision documents. The purpose of using only the court decision is, the judicial system is a hierarchy that works in a certain flow. There is an Apex Court or Supreme Court of India which is the supreme judicial body and the highest court in various countries, more than half of the total cases of this court are re-appealing cases (appellate jurisdiction) meaning where the appealing party due to the unsatisfactory decision from the lower courts re-appeals the case in the upper court. Such cases have various preceding hearings from the different lower courts in the hierarchy which creates the huge amount of data for a particular case.

The paper discusses the NyOn (Nyaya Ontology means Legal Ontology) which creates the metadata for the court decision documents in the extensive manner in which the existing ontologies fall short.

3 NyOn—The Nyaya Ontology

3.1 Use Case Scenario

We capture the requirements for the ontology based on the competency questions. We have a list of 15 T-box questions, 30+ more A-box questions depending on the data. The list of few T-box competency questions are listed below:

1. What are the different courts that can refer cases to the Supreme Court?
2. What are the different jurisdictions of SCI?
3. Number of crimes according to the location in a specific year?
4. What are the different jurisdictions of HC?
5. What properties does a court judgment have?
6. What are the various sections of IPC cited in crimes against women?
7. Cases of a specific Bench?
8. When does a case belong to appellate jurisdiction for Higher Courts?
9. When does a case belong to review jurisdiction for Supreme Courts?

3.2 Methodology

There are many methodologies available for the ontology designing process. The methodologies provide the clear concept about the flow one needs to follow while creating the ontology. Some of the stages or phases in methodologies work in iterative processes for the better enhancement of the ontology development. The names of some methodologies are TOVE (Toronto Virtual Enterprise) [12] methodology contains 6 stages, developed for Enterprise Engineering taking into consideration

the activities, time and cost as the modules. This methodology mainly focuses on the axioms creation for ontology development. Enterprise-Model Approach describes 4 stages that need to be considered for ontology development but falls short to describe in detail the process of ontology development while encoding. Methontology [9] is the most commonly used methodology for the ontology development process. It describes 7 phases of the development process with its documentation, KBSI-IDEF5 [15], OntoSpec [16] designed using DOLCE ontology for structuring application ontologies, NeON [24] Methodology, YAMO [8] Methodology. Most of the methodologies are designed by keeping in mind the specific domain and are developed according to that. Such ontologies are DiDON [17] using biomedical domain, TDDonto [18] is a Test-Driven Development for SROIQ language features.

After studying the existing methodologies for ontology development, we observe that every methodology has some strong as well as weak points. We cannot consider the existing one for our ontology development as a whole. Therefore we came up with our methodology which is named D2MD (Data to Metadata Methodology) containing 4 phases described below: **1. Purpose Identification and Requirement Specification:** The first stage in building ontology from scratch is to identify the purpose and requirements. During this phase, the scope of the application is identified using sources such as literature, surveys, motivational factors, or software engineering processes. After determining the scope, the necessary materials are acquired for the knowledge acquisition step in the form of documents, existing ontologies, literature, etc. After performing the scope determination and knowledge acquisition step, the concept extraction is performed. The entities and their appropriate relationships are decided upon in this step. **2. Ontology Development Phase (ODP):** ODP phase is the 2nd phase in the methodology which is responsible for deciding the Ontology Architecture, Designing Conceptual Map, Encoding in the OWL format, and the final step adding the formalization i.e. adding the rules for the inference purpose to discover new knowledge. **3. Validation & Evaluation:** Validation step is performed to validate the ontology based on different criteria like content validity, application-based validity, etc. For the evaluation step, correctness, completeness, accuracy, consistency, etc are the features that are considered. **4. Deployment:** This is the final phase where the publication of the ontology is taken care of along with the maintenance, updates, etc.

3.3 NyOn Modules

Figure 1 illustrates the 5 main modules of NyOn:

1. People Role

People Role module consider all the different persons that are involved in the case. This module consists of 5 major classes, namely, Party, Court Official, Author, Bench, Bench, Witness. Party class contains 4 different subclasses viz., Appellant, Respondent, Petitioner, Plaintiff. These class helps to identify the parties who

filed the case and against whom. PartyType subclass depicts whether the parties involved in the case are the individuals, organization, government, mixed type and so on. Court Official subclass determines the legal persons involved in the case. Author subclass determines one of the Judge present in the case as an Author of the case document. Witness subclass delineate who are the witness involved in the case. Bench subclass incite the type of bench present for the case hearing as it depicts the number of judges present in the court of law for the case hearing.

2. Documents

Document modules delineate the types of documents are presented in front of the court for a particular case. It also determines the case document type i.e. whether the case document itself is an appeal, petition, etc.

3. Court Decision

Court Decision module consists of 2 classes. Court Decision and Date of Judgment. The Court Decision class has two subclasses. One is Judgment which talks about the judgment given to the case in the court of law and another is Order which talks about the type of order given by the court of law which determines

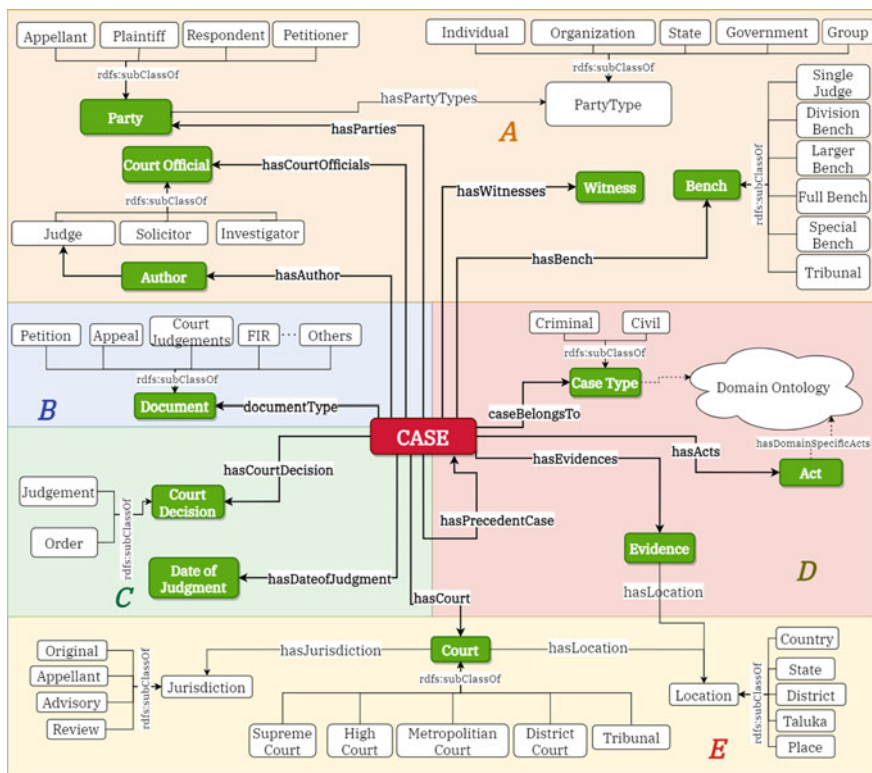


Fig. 1 NyOn ontology modules: a People role, b documents, c court decision, d litigation, e court

the action party needs to take. The Date of Judgment class is used to extract the date on which the final hearing of the case gets completed.

4. *Litigations*

Litigation module talks about the process of taking the legal action in the court of law. Here in this module we determine all the Acts, Case Type, Evidences presented, Domain Ontology concepts are considered. Acts, Evidence and Case Type are the major classes of the litigation module. This module also uses the Domain Ontology which contains the specific concepts related to the case type. The “hasPrecedentCase” is the referring the CASE class as a self loop as all the previous cases are also the court decisions from the lower courts.

5. *Court*

Court module covers the type of courts involved in the case, the jurisdiction the specific case follows and location of the court. Court is the major class of the module which is linked with the Jurisdiction and Location class. As the ontology is specific to Indian Judicial System, the subclasses of court class contains the Indian Judicial Court Names, viz., Supreme Court, High Court, Metropolitan Court, District Court and Tribunal.

4 Related Work

There is a lot of work done in the semantic web for the legal domain. Collecting the metadata from such a wide domain is a difficult task. Legal Domain contains numerous different vocabulary, different types of documents, evidences (proofs), grounding truth and so on. The section below describes the existing ontologies with their gaps and explains how NyOn addresses these problem by including an exhaustive list of entities for Indian Legal System.

Existing Ontologies grouped using ontology levels

Upper Ontology: LKIF Ontology [14], FOLaw (Valente 1995) [2], ELI (European Legislation Identifier, EU 2015) [1], NRV (Jurix 2017) [10], Top Ontology of the Law (Hage et al. 1999) [13], NM-L (Shaheed et al. 2005) [23], LRI-Core (Breuker et al. 2005) [3], CLO (Gangemi et al. 2005) [11].

Domain Ontology: LOTED (Tenders) [7], PPROC (Public Procurements) [20], GDPRtEXT (General Data Protection Regulation—EU, Pandit et al. 2018) [21], LDR (Data Rights) [19], UFO (Criminal, Oliveria Rodrigues et al. 2019) [5], IPRonto (Digital Rights) [6], JCO (Criminal Domain—Thomas et al. 2019) [25], JuDo (Core + Domain) [4].

Application Ontology: Legal Case Ontology (Wyner 2010) [26] considering single legal case document

The above listed ontologies contain either core legal concepts or the domain specific legal concepts. The ontology structure varies as every country has their own judicial structure that it follows. Almost all the ontologies are only designed for a particular region (i.e. country) or for a specific country law. Even if there exists core ontologies which can be used as the metadata for the other regions but still there

remains some gaps while considering the relations between entities, information extraction as some of the information is not being considered while designing the ontologies. Therefore we came to the conclusion to design our own ontology from scratch.

5 Conclusion

In this paper, we presented Nyay Ontology (NyOn), a multilingual modular legal ontology for representing court judgements for Indian Supreme Court Court Decision Document driven by legal use cases with the guidance of domain experts. NyOn is open for extension and in its present form, it succeeds in covering all the concepts of the criminal and civil court decision documents. The NyOn is evaluated using the OOPS! [22] scanner to identify the different levels of issues in the ontology. As future work, we plan to add more concepts from the Court Decision documents which will cover the maximum data from such documents and use the NyOn ontology for constructing a legal KG and a question answering system.

References

1. Boella G, Di Caro L, Graziadei M, Cupi L, Salaroglio CE, Humphreys L, Konstantinov H, Marko K, Robaldo L, Ruffini C et al (2015) Linking legal open data: breaking the accessibility and language barrier in European legislation and case law. In: Proceedings of the 15th International conference on artificial intelligence and law, pp 171–175
2. Breukers J, Hoekstra R (2004) Epistemology and ontology in core ontologies: FOLaw and LRI-core, two. In: Proceedings of EKAW workshop on core ontologies [Internet]. Sun SITE Central Europe, Northamptonshire, UK. Citeseer (2004)
3. Breuker J, Hoekstra R et al (2004) Core concepts of law: taking common-sense seriously. In: Proceedings of formal ontologies in information systems (FOIS-2004), pp 210–221
4. Ceci M, Gangemi A (2016) An owl ontology library representing judicial interpretations. *Semant Web* 7(3):229–253
5. de Oliveira Rodrigues CM, De Freitas FLG, De Azevedo RR (2016) An ontology for property crime based on events from UFO-B foundational ontology. In: 2016 5th Brazilian conference on intelligent systems (BRACIS). IEEE, pp 331–336
6. Delgado J, Gallego I, Llorente S, García R (2003) IPRonto: an ontology for digital rights management. In: 16th Annual conference on legal knowledge and information systems, JURIX, vol 106. Citeseer
7. Distinto I, d’ Aquin M, Motta E (2016) LOTED2: an ontology of European public procurement notices. *Semant Web* 7:267–293. <https://doi.org/10.3233/SW-140151>
8. Dutta B, Chatterjee U, Madalli DP (2015) YAMO: yet another methodology for large-scale faceted ontology construction. *J Knowl Manag*
9. Fernández-López M, Gómez-Pérez A, Juristo N (1997) Methontology: from ontological art towards ontological engineering
10. Gandon F, Governatori G, Villata S (2017) Normative requirements as linked data. In: JURIX 2017-The 30th international conference on legal knowledge and information systems, pp 1–10
11. Gangemi A, Sagri MT, Tiscornia D (2005) A constructive framework for legal ontologies. In: *Law and the semantic web*. Springer, pp 97–124

12. Grüninger M, Fox MS (1995) Methodology for the design and evaluation of ontologies
13. Hage J, Verheij B (1999) The law as a dynamic interconnected system of states of affairs: a legal top ontology. *Int J Human-Comput Stud* 51(6):1043–1077
14. Hoekstra R, Breuker J, Di Bello M, Boer A et al (2007) The LKIF core ontology of basic legal concepts. *LOAIT* 321:43–63
15. Jones D, Bench-Capon T, Visser P (1998) Methodologies for ontology development
16. Kassel G (2005) Integration of the dolce top-level ontology into the OntoSpec methodology. arXiv preprint [cs/0510050](https://arxiv.org/abs/cs/0510050)
17. Keet CM (2012) Transforming semi-structured life science diagrams into meaningful domain ontologies with DIDOn. *J Biomed Inform* 45(3):482–494
18. Lawrynowicz A, Keet CM (2016) The TDDonto tool for test-driven development of DL knowledge bases
19. Leone V, Di Caro L, Villata S (2020) Taking stock of legal ontologies: a feature-based comparative analysis. *Artif Intell Law* 28(2):207–235
20. Muñoz J, Esteban G, Corcho O, Serón F (2016) PPROC, an ontology for transparency in public procurement. *Semant Web* 7:295–309. <https://doi.org/10.3233/SW-150195>
21. Pandit H, Fatema K, O’Sullivan D, Lewis D (2018) GDPRtEXT—GDPR as a linked data resource, pp 481–495. https://doi.org/10.1007/978-3-319-93417-4_31
22. Poveda-Villalón M, Gómez-Pérez A, Suárez-Figueroa MC (2014) OOPS! (OntOlogy Pitfall Scanner!): an on-line tool for ontology evaluation. *Int J Semant Web and Inf Syst (IJSWIS)* 10(2):7–34
23. Shaheed J, Yip A, Cunningham J (2005) A top-level language-biased legal ontology. In: Workshop proceedings, legal ontologies and artificial intelligence techniques, international association for artificial intelligence and law, workshop series, vol 4. Wolf Legal Publishers, pp 13–24. Citeseer
24. Suárez-Figueroa MC, Gómez-Pérez A, Fernández-López M (2012) The neon methodology for ontology engineering. In: *Ontology engineering in a networked world*. Springer, pp 9–34
25. Thomas ASS (2017) A legal case ontology for extracting domain-specific entity-relationships from e-judgments
26. Wyner A (2008) An ontology in owl for legal case-based reasoning. *Artif Intell Law* 16(4):361–387

The Applications and Deployment Track

Technologies for AI-Driven Fashion Social Networking Service with E-Commerce



Jinseok Seol, Seongjae Kim, Sungchan Park, Holim Lim, Hyunsoo Na, Eunyoung Park, Dohee Jung, Soyong Park, Kangwoo Lee, and Sang-goo Lee

Abstract The rapid growth of the online fashion market brought demands for innovative fashion services and commerce platforms. With the recent success of deep learning, many applications employ AI technologies such as visual search and recommender systems to provide novel and beneficial services. In this paper, we describe applied technologies for AI-driven fashion social networking service that incorporate fashion e-commerce. In the application, people can share and browse their outfit-of-the-day (OOTD) photos, while AI analyzes them and suggests similar style OOTDs and related products. To this end, we trained deep learning-based AI models for fashion and integrated them to build a fashion visual search system and a recommender

J. Seol (✉) · S. Kim · H. Lim · H. Na · S. Lee

Department of Computer Science and Engineering, Seoul National University, Seoul, South Korea
e-mail: jamie@europa.snu.ac.kr

S. Kim

e-mail: sjkim@europa.snu.ac.kr

H. Lim

e-mail: ihl7029@europa.snu.ac.kr

H. Na

e-mail: monchana@europa.snu.ac.kr

S. Lee

e-mail: sglee@europa.snu.ac.kr

S. Park · E. Park

IntelliSys Co., Ltd., Chennai, India

e-mail: scpark@intellisys.co.kr

E. Park

e-mail: eypark@intellisys.co.kr

D. Jung · S. Park · K. Lee

LOTTE Homeshopping Inc., Seoul, South Korea

e-mail: conan94@lotte.net

S. Park

e-mail: soyong.park19@lotte.net

K. Lee

e-mail: kangwoo.lee@lotte.net

system for OOTD. With aforementioned technologies, the AI-driven fashion SNS platform, *iTOO*, has been successfully launched.

Keywords Fashion AI · AI-driven SNS · Visual search · Recommender system

1 Introduction

With the development of the internet and computer technologies, the size of the e-commerce market has been growing steeply. Moreover, the social distancing environment of the COVID-19 pandemic has brought growth and demands for innovative e-commerce platforms [1]. To meet the demands and achieve benefits, leading commerce companies such as Amazon [2], eBay [3], and Alibaba [4] are introducing distinctive and novel services, including visual search and product recommendations [5, 6]. Meanwhile, in the fashion industry, a variety of innovative applications have emerged. For example, ViSENZE [7] provides fashion image processing solutions including image search and attribute prediction, and Zalando research team has been working on fashion product recommendation [8]. There are more interesting applications such as Intelistyle [9], which provides a chatbot-based AI stylist, and a wardrobe-based AI stylist Fitzme [10] (Fig. 1).

Fashion-focused social networking is another large portion of consumer activities [11]. We claim that people look for the outfit-of-the-day (OOTD) of other people to acquire insights and trends in fashion. To achieve this, consumers browse online applications such as Lookbook, Polyvore, Pinterest, etc. Moreover, users commonly share their OOTD photos through general social networking services (SNS) (e.g., Facebook, Instagram, and Twitter). In this environment, connecting fashion SNS with e-commerce is undoubtedly beneficial, as in the case of Instagram and Styleshare. Therefore, it is reasonable to come up with a new kind of service that incorporates e-commerce with fashion SNS by applying AI technologies.

Normally, to connect user-uploaded OOTD photos with retail products, the uploader must directly attach a link to the product, which is a cumbersome task, leading to the automation challenge. We applied the fashion visual search system to overcome the challenge, thus the users can easily share their OOTD in the form of a common fashion SNS. It provides an opportunity for other users to purchase retail products similar to the OOTD, without burdening the uploader. To this end, we implemented AI components including a fashion object detector, a fashion category classifier, and a fashion attribute tagger. Moreover, we deployed a personalized recommender system that can handle OOTD data to engage more users to the application. These AI technologies are combined and enabled launching the *iTOO*.

2 Related Work

2.1 *Deep Learning for Fashion*

The most important media type in the fashion domain is photographic images. However, it contains major challenges that make it difficult to process images in the fashion domain [12]. Typically, self-occlusions may occur in the target of interest, and when a person is wearing fashion garments, the image variance can be amplified due to the viewpoint, posture, lighting, and scale of the subject [13]. In addition, several fashion items may appear in a single image simultaneously. Therefore, localization procedure is essential. To this end, region-of-interest (RoI) detection, landmark detection [14], parsing with a fashion component [15], or pose estimation [16] is often employed. Moreover, classifying the categories of fashion products [17], predicting colors and detailed attributes including latent fashion style, is also a major component to recognize fashion item [18, 19].

2.2 *Visual Search System*

Image retrieval, or a visual search system, has been successfully applied in various areas such as face recognition [20] and product search [3]. Recently, many studies implemented a visual search model by comprehending images through CNN and learning through deep metric learning [21]. Academic datasets for fashion visual search are publicly available [22, 23], but to apply in real-world application, datasets should cover a broader range of categories. Therefore, collecting and refining the dataset is also an essential task [24]. Techniques using proxies by employing latent embeddings for each instance are the current state-of-the-art models for the visual search [25], to the best of our knowledge. However, when the number of items becomes millions, training causes another challenge and requires complex techniques [26].

2.3 *Recommender System*

The recommender system is a core technology in contents platforms and e-commerce as Amazon [27], YouTube [28], and Netflix [29] have shown. Starting with Collaborative Filtering [30, 31], advanced methods including implicit-based approaches [32], content-based filtering [33], and more recently, deep learning-based recommendation algorithms [34] have been studied. In the fashion domain, models using visual information have been mainly studied [35, 36]. Furthermore, there are many cases that multiple images form a single outfit, thus outfit recommendation techniques that consider multiple images simultaneously are being studied [37, 38].

3 Service Overview

In this section, we introduce *iTOO* from LOTTE. *iTOO* is a service that integrates Fashion SNS and commerce, where you can share and browse your OOTD photos, look for related products and styles, get recommendations, and even purchase fashion products in place.

3.1 Share and Browse OOTD

One of the main purposes of the application is to share a picture of your OOTD. A user can add a brief description and hashtags when uploading the picture. Immediately, the fashion items that comprise the OOTD are detected and analyzed automatically by AI. Therefore, users can easily share extensive information by simply taking a picture and leaving a short description. Moreover, users can browse through OOTDs posted by other users through curation or exploration. As shown in Fig. 2, the home screen recommends OOTDs that fit the preference, style, and body shape of the user.

3.2 Look into OOTD

By examining the OOTD detail view, a user can check out purchasable retail products that are similar to the comprising items of the OOTD as illustrated in Fig. 1. This function benefits users who want to buy fashion products through OOTD curation in place, and retail shops can merchandise their products through viral marketing. In addition, other OOTDs with a similar style are recommended. Such curations help users to drill down OOTD pools that fit their preferences.

3.3 Get More Recommendations

Besides aforementioned OOTD recommendations, the “style leaders” who often post trending and decent OOTDs are also recommended to a user as who-to-follow. To get more accurate curations, a user can provide detailed information about the fashion persona such as demographic information, body shape, and preference style tags. All information including OOTD interactions, profiles, and following user list is gathered to the recommender system and provides a personalized recommendation.

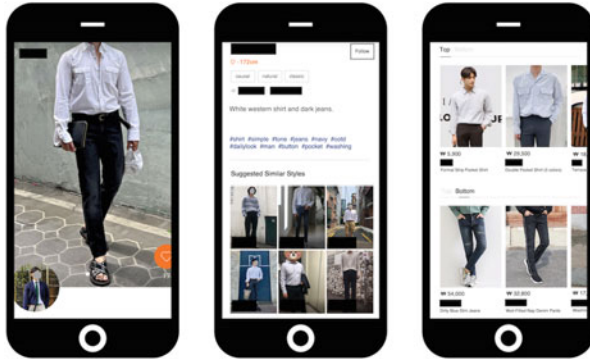


Fig. 1 Actual usage screen of *iTOO*, a Fashion SNS from LOTTE. Whenever users upload their outfit-of-the-day (OOTD) photos, AI analyzes them and suggests similar style OOTDs and related products

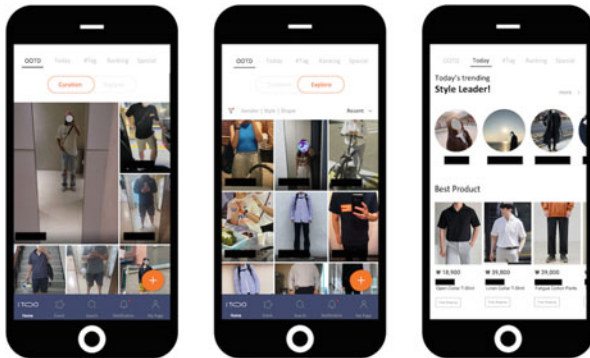


Fig. 2 The home screen of our AI-driven fashion SNS, *iTOO*. Users can browse and interact with the OOTDs of other users just like in ordinary SNS. Most of the contents are curated by the recommender system

4 AI Components

This section introduces the AI technologies that enable fashion SNS focused on OOTD images. We used the best-performing deep learning models in our knowledge, and they were fine-tuned to be suited for the fashion domain and the application. The core models of the application, visual search, and OOTD recommender system, are described in Sects. 5 and 6, respectively. Note that the part of the AI component is also used to construct datasets for training the other AI components.

4.1 Fashion Object Detector

A fashion image may present a single product, but in many cases, it comes with a person wearing several fashion garments. Therefore, localizing or detecting where the fashion items are in the image is a process that must be preceded. We considered pose estimation and human parsing methods, however, we adopted a model that predicts region of interest (RoI) for practicality because human information does not always come in. As a training dataset, we mixed and reorganized Street2Shop [39], ModaNet [40], and DeepFashion [22, 23] datasets. In detail, we mapped category information to 6 super-categories (top, bottom, outer, dress, shoes, bag) in order to combine datasets from different sources. Due to the throughput performance issue, YoloV4 [41] was selected as our detector model. Note that, in the application situation, we could assume that each OOTD image has at least one fashion item. Moreover, top/bottom items and dresses are mutually exclusive, so considering these properties, we added a post-processing module and achieved performance gain in terms of recall. Furthermore, we also included the fashion category classifier module from Sect. 4.2 to increase the precision. The predicted RoI is cropped and inferred by the category classifier, and filtered out if the super-category is different from the detector model. By combining the YoloV4 with the post-processing module, we could build a fast and accurate fashion object detector.

4.2 Fashion Category Classifier

Classifying the category of a fashion item is another basic element of fashion item recognition. We constructed an integrated dataset using ModaNet, DeepFashion, iMaterialist [42], and crawled data from YOOX and Polyvore. To increase category coverage, we crawled the data from the top popular 30 online fashion malls, summing up to 1.7M images in total. Since the crawled data does not have RoI labels, we used the fashion object detector model from Sect. 4.1 and filtered out RoIs with a super-category label parsed from metadata provided by the malls. It is necessary to reorganize the category hierarchy to integrate multiple datasets, thus we designed a category hierarchy consisting of 6 super-categories and 32 sub-categories: 6 from outer, 6 from top, 6 from bottom, 2 from dress, 7 from shoes, and 5 from bag. As a classification model backbone, EfficientNet [43] was employed under consideration of inference speed and memory usage. We also leverage the training techniques such as cosine annealing and label smoothing, etc.

4.3 Fashion Attribute Tagger

We build a fashion attribute tagger model to find detailed attributes of fashion items such as color, style, and length. Specifically, we defined 18 attribute groups, where 11 are categorical and 7 are multi-label. Since some attribute groups are only limited according to the sub-category, outputs are filtered out through the post-processing module. Similar to other AI component models, we merged and reorganized multiple datasets from different sources: DeepFashion, iMaterialist, Fashion550k [44], and MVC [45]. Adopted CNN backbone and training methods are the same as fashion category classifier.

5 Fashion Visual Search

To train the visual search model, the same-class labels denoting different images of the same item are necessary. In the fashion domain, the image variance especially the gap between the product image provided by the shopping malls and the image uploaded by consumers is large. Therefore, it is important to construct a model and datasets that can cover the cross-domain image retrieval task. To this end, we collected multiple datasets and pre-processed them to build an in-house dataset suitable for the fashion visual search model in the application. Note that since it is common to learn through negative sampling rather than learning the distribution of all items, dataset quality is sensitive to false positives rather than false negatives.

5.1 Dataset Construction

5.1.1 Collecting Data

Academic datasets (e.g., DeepFashion) are often not complete and cannot be directly applied to real-world applications due to the limitation of category coverage. To fill this gap, we selected and crawled the top popular 30 online fashion malls and collected a total of 0.3 M items with 1.3 M images, including consumer photo review data. We conducted a small experiment to confirm that adding more data affects search performance in terms of category coverage.

5.1.2 Preprocessing

Similar to the case of the fashion category classifier from Sect. 4.2, crawled data cannot be used for training without preprocessing. We used the fashion object detector

model from Sect. 4.1 and acquired RoI crops for the localization, and filtered out the crops that do not match the super-category information parsed from target malls.

Meanwhile, when crawling the data from online malls, the abundant image data is often located in the “descriptive image”, which consists of multiple photos of a fashion item, description texts, and even irrelevant images like advertisements. To gather meaningful data, we first separated the descriptive image with the connected components algorithm, then removed duplicate images using perceptual hashing [46]. The detector model and post-process procedures are applied then after. Additionally, to reduce false-positive errors, we use the fashion category classifier and select images with the sub-category of the majority.

5.2 Color Separation

An easy-to-miss aspect when building a dataset for a fashion visual search model is to separate fashion items that have multiple color variants. Many online fashion malls, including DeepFashion dataset, treat item images that differ only in color as the same item. However, this scheme can lead CNN to neglect the color information of the input image, and a “shortcut” by color information cannot be used. In our settings, it is beneficial to use this shortcut because the precision of the search result is more important than the recall, and by conducting a benchmark experiment, we confirmed that separating the color variants into different items helps to improve precision. In the case of the DeepFashion dataset, the color information labels are provided with fine granularity, so we re-adjusted the same-class label using the color tag of our fashion attribute tagger from Sect. 4.3. Again, when it comes to precision, only the false-positives of the dataset matter so the inaccuracy of the attribute tagging model does not affect critically.

5.3 Model

Most of the recent state-of-the-art methods on image retrieval tasks are based on metric learning. When the model is trained, we can obtain a representation vector from the input image by feeding it into the model, and similar items can be retrieved through cosine similarity. We considered basic metric learning [20], AP learning [47], and proxy-based methods [25]. However, methods that require item embeddings are often difficult to deal with numerous or variable item pool. Moreover, we use an under/over-sampling scheme to balance the datasets from different sources, which means, the whole item pool is changed on every epoch. Therefore, for the flexibility of the training, we adopted simple N -pair contrastive learning [48]. Although the basic metric learning cannot match the state-of-the-art performance, it still serves as a decent baseline with advantages from other aspects.

In concrete, to train the N -pair loss, we sample one positive (same item) image per input image and gather N negative image samples from the training batch. The metric learning is performed using normalized-temperature cross-entropy (NT-Xent) loss [49]. The rest of the training detail including backbone CNN is similar to the category classifier from Sect. 4.2. The dimension of the representation vector was set to 128 for memory efficiency, and although a larger dimension was under consideration, the performance improvement compared to memory usage was not significant. The under/over-sampling of datasets are empirically adjusted considering the image types, characteristics, and category distribution of each dataset.

5.4 Experimental Results

5.4.1 Performance on Benchmark Dataset

To show the precision gain on the color separation scheme, we experimented on DeepFashion In-shop dataset, which has 52,712 images with 7,982 items. Note that this dataset provides RoI crop data, so we use the box coordinates with 20-pixel margin, then resized it into 256×256 images. Table 1 shows the results in top- k accuracy, which checks whether the positive image is within the top- k items retrieved. Although the N -pair baseline model cannot reach state-of-the-art performance, it is still a decent baseline compared to older and complex models. On the other hand, we can see the significant performance gain in top-1 when the color separation scheme is applied.

Table 1 Performance comparison on DeepFashion In-shop dataset, using top- k accuracy. Our N -pair based model may not be state-of-the-art, but it can be easily scaled out to millions of items. The suggesting color separation training scheme (last row) shows that even with simple label modification, the top-1 accuracy can be improved by a large margin

Model	$k = 1$	$k = 5$	$k = 10$	$k = 20$
Liu et al. [22]	53.0	–	73.0	76.0
Park et al. [10]	–	82.6	–	90.9
Cakir et al. [47]	90.9	–	97.7	98.5
Kim et al. [25]	92.6	–	98.3	98.9
Baseline	77.9	91.6	94.5	96.5
Color separation	83.4	92.4	94.0	95.6



Fig. 3 Examples of results from visual search models, trained in different dataset compositions. Results from the second query show that by adding more review data, robustness to cross-domain retrieval can be improved. The third query shows that the model cannot accurately deal with unseen categories. We can conclude that in the visual search model, dataset composition is critical as model architecture

5.4.2 Dataset Influence

To see the influence of the dataset constitutions, we trained a N -pair model with three different dataset settings: using only DeepFashion dataset, adding more consumer photo review data from crawled online fashion malls, and adding shoes and bags which DeepFashion does not have. The examples of visual search results are shown in Fig. 3. In the results from the first query image, all three settings produced similar query image results. In the second case, since the query image involves a partially human shape, the setting with an additional consumer review image shows more robustness in terms of cross-domain image retrieval. In the final case, where the query image represents shoes, it can be seen that settings without shoes and bags could not maintain the sub-category of the query image. As a result, we argue that constructing a well-tempered dataset is just as important as selecting the model.

5.5 Inference Pipeline

In the application, the visual search model has to be combined with other AI components. As shown in Fig. 4, when a user uploads an OOTD image, the fashion object detector first finds RoIs. Then, the fashion category classifier and the fashion attribute tagger are applied to each cropped ROI in parallel. Finally, the visual search model extracts representation vectors and store them into the vector index, corresponding to the super-category.

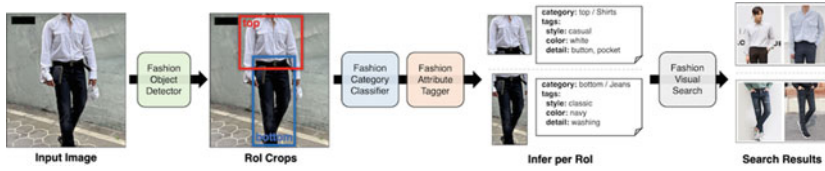


Fig. 4 Inference pipeline for OOTD. After the fashion object detector model finds region of interest (RoI), the fashion category classifier and the fashion attribute tagger are applied to acquire more detailed information for each cropped RoI. After that, the visual search model extracts the representation vectors and stores them to the corresponding vector index

6 Recommender System

With the recommender system, users receive personalized OOTD recommendations, similar styled OOTDs for each OOTD, and style leaders to follow.

6.1 Personalized OOTD Curation

On the first screen of the service, users can get the recommendation of OOTDs that suits their preferences. The recommendation basically leverage CF-CBF, and the final recommendation list is generated by mixing up with the weekly best products and best products by demographic-based user segment. In the case of CF-CBF, both user-based and item-based CF are used.

6.1.1 Style Vector

A fashion item vector \mathbf{v}_i of item I_i consists of a concatenation of representations of the category classifier, the attribute tagger, and the visual search model: for an item image x_i , $\mathbf{v}_i = \text{concat}(f_C(x_i), f_A(x_i), f_S(x_i))$. For each fashion item, the item style vector $\tilde{\mathbf{v}}_i$ is obtained by subtracting the average of item vectors of the sub-category which the given item belongs: for $c(i)$ a sub-category index of an item I_i , $S_k = \{j \mid c(j) = k\}$, $\bar{\mathbf{v}}_{c(i)} = |S_{c(i)}|^{-1} \sum_{j \in S_{c(i)}} \mathbf{v}_j$, $\tilde{\mathbf{v}}_i = \mathbf{v}_i - \bar{\mathbf{v}}_{c(i)}$. The OOTD style vector is defined as the average of the item style vector of the following items: $\mathbf{o}_t = |o_t^*|^{-1} \sum_{i \in o_t^*} \tilde{\mathbf{v}}_i$, where o_t^* is a set of indices of comprising items in the OOTD o_t .

6.1.2 Semantic OOTD Similarity

Given two OOTDs, we define semantic OOTD similarity as a weighted sum of the cosine similarity between the OOTD style vectors of each OOTD and the Jaccard

similarity of the hashtags that are dependent on the two OOTDs: for given OOTDs o_{t_1} and o_{t_2} ,

$$\text{sim}_o(o_{t_1}, o_{t_2}) = \lambda_o \left(\frac{\mathbf{o}_{t_1} \cdot \mathbf{o}_{t_2}}{\|\mathbf{o}_{t_1}\| \|\mathbf{o}_{t_2}\|} \right) + (1 - \lambda_o) \frac{|a_o(t_1) \cup a_o(t_2)|}{|a_o(t_1) \cap a_o(t_2)|}, \quad (1)$$

where $a_o(t)$ denotes a set of hashtags of an OOTD o_t . Note that we use this similarity to make similar styled OOTD recommendations.

6.1.3 Semantic User Similarity

Similar to semantic OOTD similarity, we define a user style vector by aggregating style vectors of H OOTDs that the user has recently viewed or liked. We use a weighted average to reflect the preferences of recent interaction more strongly: for user u_n ,

$$\mathbf{u}_n = \left(\sum_m^H w_m \right)^{-1} \sum_{m=1}^H w_m \mathbf{o}_{t_m}, \quad (2)$$

where $w_m = \left(\frac{H-m+1}{H} \right)^\alpha$, and $u_n^* = \{t_1, t_2, \dots, t_H\}$ is a set of indices of the user's recent OOTD views or likes, and $0 < \alpha < 1$ is a recency decay hyper-parameter. Cosine similarity between two user style vectors and the Jaccard similarity between the preference tags in the user profiles are used to measure the semantic user similarity: for user u_{n_1} and u_{n_2}

$$\text{sim}_u(u_{n_1}, u_{n_2}) = \lambda_u \left(\frac{\mathbf{u}_{n_1} \cdot \mathbf{u}_{n_2}}{\|\mathbf{u}_{n_1}\| \|\mathbf{u}_{n_2}\|} \right) + (1 - \lambda_u) \frac{|a_u(u_{n_1}) \cup a_u(u_{n_2})|}{|a_u(u_{n_1}) \cap a_u(u_{n_2})|}, \quad (3)$$

6.1.4 CF-CBF for OOTD

We use Collaborative Filtering (CF) as the basis for our recommendation algorithm. We first calculate the TF-IDF values from user-OOTD interactions. The value of TF-IDF is decayed to reflect the recency using time decay coefficient β^d , where d is days passed since the interaction has occurred, and $0 < \beta < 1$ is a decay rate hyper-parameter. We use both item-based and user-based CF and combine it with other recommendation results. In the case of item-based, the recommended OOTD list is obtained through similarity of the OOTDs that the user has recently viewed. In the case of user-based, the recommendation list is constructed by joining users obtained through user similarity and the OOTD list that the user has recently viewed. In both cases, CF-CBF can be implemented by considering TF-IDF as a CF part and semantic similarity as a Content-Based Filtering (CBF) part. Let \mathbf{r}_n be a TF-IDF vector of a user u_n , treating the user as a document when calculating the TF-IDF

values. Then, the final similarity between two users u_{n_1}, u_{n_2} can be calculated as follows:

$$\text{sim}_{\text{CF-CBF}}(u_{n_1}, u_{n_2}) = \lambda_{\text{CF}} \left(\frac{\mathbf{r}_{n_1} \cdot \mathbf{r}_{n_2}}{\|\mathbf{r}_{n_1}\| \|\mathbf{r}_{n_2}\| + h} \right) + (1 - \lambda_{\text{CF}}) \text{sim}_{\text{u}}(u_{n_1}, u_{n_2}), \quad (4)$$

where h is a shrinkage term for the case with relatively small interactions [50]. A similar methodology is applied to item-based CF-CBF.

6.1.5 OOTD Curation

With user-based and item-based CF-CBF, we mix up the weekly best OOTD list with the best OOTD list by segment based on demographic information. Note that since a decay term is used, a result closer to global taste is provided rather than personalized content to those who have not used the application for a long time. This reflects the characteristics of the fashion domain where trends change over time and provides exploration opportunity and serendipity.

6.2 Style Leader Suggestion

A style leader means a person who can be subscribed, and a user can receive better OOTD curation when they follow the style leader. For style leader recommendation, both the latent method and the graph-based method are used. In the case of the latent method, recommendation candidates are determined using the modified semantic user similarity. Here, we use cosine similarity between the user style vectors of the recent view/like OOTD history of the follower and the user style vectors of the recent *upload* OOTD history of the following candidates. On the other hand, when using the graph-based algorithm, recommendation candidates are obtained by performing a random walk twice in the following/follower relationship graph. Finally, we recommend a mixture of latent-based, graph-based, similar segment users using demographic information, and popular users. Segment and the weekly best serve as exploration and baseline at the same time.

7 System Deployment

7.1 Overall Architecture

Serving deep learning models for real-world application requires high-cost and complex infrastructure. To minimize the burden, we adopt AWS Cloud Service, mainly

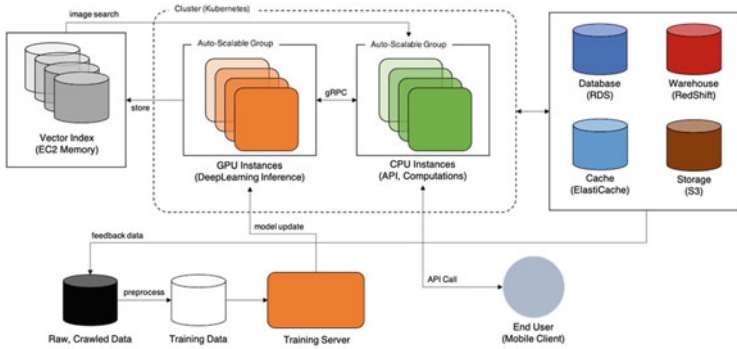


Fig. 5 Overall system architecture. To serve deep learning models in a real-world application, we adopted Kubernetes that can orchestrate complex infrastructure

orchestrated through Kubernetes. Deep learning models are loaded on auto-scalable GPU pods to adapt to the variable traffic. The overall architecture is illustrated in Fig. 5.

7.2 *Serving Deep Learning Models*

The development of deep learning models is usually done using frameworks such as PyTorch or TensorFlow in an experimental environment. To serve the trained model for the inference, we used NVIDIA Triton Inference Server since it can accommodate all types of neural network models exported in ONNX format, independent from the deep learning framework. For the communication between the service API and the deep learning models, we created an in-house gRPC client library. Each inference step is divided into CPU-heavy parts, such as image preprocessing or data loading, and the core GPU-consuming part so that each component can be scaled out in parallel.

7.3 *Vector Indexing*

The throughput of the visual search system heavily relies on similar vector search algorithms [51]. We considered well-known approaches including Deep Hashing [51] and hierarchical search methods [52]. Empirically, vectors from visual search models form intrinsically clustered spaces, thus separating the hashing stage from the model does not degrade search accuracy compared to learning to hash methods. Therefore, HNSW [52] was adopted in consideration of implementation difficulty, search time complexity, and memory used. In our situation, hundreds of fashion items are added

every day, so the vector index is rebuilt every dawn to include such items. In the case of HNSW, the memory consumption increases linearly for the items in the database. Therefore, whenever the index cannot be afforded by a single computing instance, we apply the sharding technique [4, 26] and rearrange the search results through post-processing. Note that since the super-category of a fashion item is almost always accessible, vector indexes are built separately according to the super-categories.

7.4 Data Warehouse and DAG Management

When it comes to the recommender system, it is necessary to analyze logs and identify the user-item relationships from large-scale data. To this end, we implemented data processing modules and a basic CF model using AWS RedShift, the data warehouse instance. Moreover, both the visual search system and the recommender system are a pipeline of relatively small modules. In this structure, task parallelism can be applied to improve throughput. We adopted Argo as a Directed Acyclic Graph (DAG) task management tool to implement task parallelism. Through Argo and Kubernetes configurations, we can automatically scale out the bottlenecks in the DAG.

8 Conclusion

In this paper, we describe technologies for AI-driven fashion SNS that incorporate fashion e-commerce. Users can share and browse their OOTD, while detailed fashion attribute analysis, similar products search, and getting recommendations are all automatically provided by AI. To this end, we built a fashion object detector, a fashion category classifier, a fashion attribute tagger, a fashion visual search system, and an OOTD recommender system. With all these techniques, the fashion SNS platform *iTOO* from LOTTE has been launched. Future work is to tune the hyperparameters of the AI models and improve model architecture with user feedback.

Acknowledgements This work was made in collaboration with Seoul National University, IntelSys Co., Ltd., and LOTTE Homeshopping Inc. Also, this work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2021-0-00302, AI Fashion Designer: Mega-Trend and Merchandizing Knowledge Aware AI Fashion Designer Solution). Special thanks to Jeeseung Han.

References

1. Silvestri B (2020) The future of fashion: how the quest for digitization and the use of artificial intelligence and extended reality will reshape the fashion industry after covid-19. *ZoneModa J* 10(2):61–73
2. Shrestha N, Nasoz F (2019) Deep learning sentiment analysis of amazon.com reviews and ratings. [arXiv:1904.04096](https://arxiv.org/abs/1904.04096)
3. Yang F, Kale A, Bubnov Y, Stein L, Wang Q, Kiapour H, Piramuthu R (2017) Visual search at ebay. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. pp 2101–2110
4. Zhang Y, Pan P, Zheng Y, Zhao K, Zhang Y, Ren X, Jin R (2018) Visual search at alibaba. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. pp 993–1001
5. Mohanty SN, Chatterjee JM, Jain S, Elngar AA, Gupta P (2020) Recommender system with machine learning and artificial intelligence: practical tools and applications in medical, agricultural and other industries. Wiley
6. Jain S (2020) Exploiting knowledge graphs for facilitating product/service discovery. [arXiv:2010.05213](https://arxiv.org/abs/2010.05213)
7. Chokshi S, Bhattacharya L (2020) Transforming the vision of retail with ai: Visenze
8. Freno A (2017) Practical lessons from developing a large-scale recommender system at zalando. In: *Proceedings of the eleventh ACM conference on recommender systems*. pp 251–259
9. Zou X, Wong W (2021) fashion after fashion: a report of ai in fashion. [arXiv:2105.03050](https://arxiv.org/abs/2105.03050)
10. Park S, Han J, Kim JY, Lim H, Kim S, Jung J, Park E, Lee SG, Lee Y, Rha JY (2019) A deep learning based architecture for personal a.i. fashion stylist services. In: *The 2nd artificial intelligence on fashion and textile international conference (AIFT 2019)*
11. Nelson DW, Moore MM, Swanson KK (2019) Fashion and social networking: a motivations framework. *J Fash Mark Manag: Int J*
12. Cheng WH, Song S, Chen CY, Hidayati SC, Liu J (2021) Fashion meets computer vision: a survey. *ACM Comput Surv (CSUR)* 54(4):1–41
13. Zhan H, Shi B, Kot AC (2017) Cross-domain shoe retrieval using a three-level deep feature representation. In: *2017 IEEE international symposium on circuits and systems (ISCAS)*. IEEE, pp 1–4
14. Liu Z, Yan S, Luo P, Wang X, Tang X (2016) Fashion landmark detection in the wild. In: *European conference on computer vision*. Springer, pp 229–245
15. Liang X, Liu S, Shen X, Yang J, Liu L, Dong J, Lin L, Yan S (2015) Deep human parsing with active template regression. *IEEE Trans Pattern Anal Mach Intell* 37(12):2402–2414
16. Toshev A, Szegedy C (2014) Deeppose: human pose estimation via deep neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp 1653–1660
17. Cho H, Ahn C, Min Yoo K, Seol J, Lee SG (2019) Leveraging class hierarchy in fashion classification. In: *Proceedings of the IEEE/CVF international conference on computer vision workshops*. p
18. Li Y, Huang C, Loy CC, Tang X (2016) Human attribute recognition by deep hierarchical contexts. In: *European conference on computer vision*. Springer, pp 684–700
19. Lee H, Seol J, Lee SG (2017) Style2vec: Representation learning for fashion items from style sets. [arXiv:1708.04014](https://arxiv.org/abs/1708.04014)
20. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp 815–823
21. Chen W, Liu Y, Wang W, Bakker E, Georgiou T, Fieguth P, Liu L, Lew MS (2021) Deep image retrieval: a survey. [arXiv:2101.11282](https://arxiv.org/abs/2101.11282)
22. Liu Z, Luo P, Qiu S, Wang X, Tang X (2016) Deepfashion: powering robust clothes recognition and retrieval with rich annotations. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp 1096–1104

23. Ge Y, Zhang R, Wang X, Tang X, Luo P (2019) Deepfashion2: a versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 5337–5345
24. Corbiere C, Ben-Younes H, Ramé A, Ollion C (2017) Leveraging weakly annotated data for fashion image retrieval and label prediction. In: Proceedings of the IEEE international conference on computer vision workshops. pp 2268–2274
25. Kim S, Kim D, Cho M, Kwak S (2020) Proxy anchor loss for deep metric learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 3238–3247
26. Zhao K, Pan P, Zheng Y, Zhang Y, Wang C, Zhang Y, Xu Y, Jin R (2019) Large-scale visual search with binary distributed graph at alibaba. In: Proceedings of the 28th ACM international conference on information and knowledge management. pp 2567–2575
27. Linden G, Smith B, York J (2003) Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Comput* 7(1):76–80
28. Covington P, Adams J, Sargin E (2016) Deep neural networks for youtube recommendations. In: Proceedings of the 10th ACM conference on recommender systems. pp 191–198
29. Bennett J, Lanning S, et al (2007) The netflix prize. In: Proceedings of KDD cup and workshop, vol 2007. New York, NY, USA, p 35
30. Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J (1994) Grouplens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM conference on computer supported cooperative work. pp 175–186
31. Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web. pp 285–295
32. Hu Y, Koren Y, Volinsky C (2008) Collaborative filtering for implicit feedback datasets. In: 2008 Eighth IEEE international conference on data mining. IEEE, pp 263–272
33. Basu C, Hirsh H, Cohen W, et al (1998) Recommendation as classification: Using social and content-based information in recommendation. In: *Aaai/iaai*. pp 714–720
34. He X, Liao L, Zhang H, Nie L, Hu X, Chua TS (2017) Neural collaborative filtering. In: Proceedings of the 26th international conference on world wide web. pp 173–182
35. He R, McAuley J (2016) Vbpr: visual bayesian personalized ranking from implicit feedback. In: Proceedings of the AAAI conference on artificial intelligence, vol 30
36. Yin R, Li K, Lu J, Zhang G (2019) Enhancing fashion recommendation with visual compatibility relationship. In: The world wide web conference. pp 3434–3440
37. Lin Y, Moosaei M, Yang H (2020) Outfitnet: Fashion outfit recommendation with attention-based multiple instance learning. In: Proceedings of the web conference 2020. pp 77–87
38. Lu Z, Hu Y, Chen Y, Zeng B (2021) Personalized outfit recommendation with learnable anchors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 12722–12731
39. Hadi Kiapour M, Han X, Lazebnik S, Berg AC, Berg TL (2015) Where to buy it: matching street clothing photos in online shops. In: Proceedings of the IEEE international conference on computer vision. pp 3343–3351
40. Zheng S, Yang F, Kiapour MH, Piramuthu R (2018) Modanet: a large-scale street fashion dataset with polygon annotations. In: Proceedings of the 26th ACM international conference on Multimedia. pp 1670–1678
41. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: optimal speed and accuracy of object detection. [arXiv:2004.10934](https://arxiv.org/abs/2004.10934)
42. Guo S, Huang W, Zhang X, Srihanta P, Cui Y, Li Y, Adam H, Scott MR, Belongie S (2019) The materialist fashion attribute dataset. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. p 0
43. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR, pp 6105–6114
44. Inoue N, Simo-Serra E, Yamasaki T, Ishikawa H (2017) Multi-label fashion image classification with minimal human supervision. In: Proceedings of the IEEE international conference on computer vision workshops. pp 2261–2267

45. Liu KH, Chen TY, Chen CS (2016) Mvc: A dataset for view-invariant clothing retrieval and attribute prediction. In: Proceedings of the 2016 ACM on international conference on multi-media retrieval. pp 313–316
46. Zauner C (2010) Implementation and benchmarking of perceptual image hash functions
47. Cakir F, He K, Xia X, Kulis B, Sclaroff S (2019) Deep metric learning to rank. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 1861–1870
48. Sohn K (2016) Improved deep metric learning with multi-class n-pair loss objective. In: Advances in neural information processing systems. pp 1857–1865
49. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR, pp 1597–1607
50. Bell RM, Koren Y (2007) Improved neighborhood-based collaborative filtering. In: KDD cup and workshop at the 13th ACM SIGKDD international conference on knowledge discovery and data mining. Citeseer, pp 7–14
51. Liu H, Wang R, Shan S, Chen X (2016) Deep supervised hashing for fast image retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 2064–2072
52. Malkov YA, Yashunin DA (2018) Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans Pattern Anal Mach Intell* 42(4):824–836

Deep Learning-Based Classification of Customer Communications of a German Utility Company



Jinghua Groppe, René Schlichting, Sven Groppe, and Ralf Möller

Abstract The Germany utility company, Zweckverband Ostholstein (ZVO), receives each year a large number of customer communications that are read and classified by its employees. To automate the processing of customer communications, University of Lübeck and ZVO collaborate to model the classification problem of customer communications over a challenging real-world dataset that is multi-label, imbalanced, extremely varied in length and full of noise. To find an optimal model, we first identify suitable classification algorithms and text encoding techniques and then develop 12 deep learning models that combine the different capabilities of neural networks with classical feature-extraction methods and modern word embedding techniques. These models are extensively tested and evaluated, and the results of evaluation are in-depth analyzed and discussed. Finally, we not only find an optimal model for the challenging dataset but also obtain interesting findings and insights that we believe will benefit other deep learning-based text classification projects.

1 Introduction

Given a set of text documents and their classes, text classification technologies are used to discover the hidden patterns between the texts and their classes and the patterns are then employed to automatically predict classes of new texts. Text classification has a wide range of applications in data mining, information retrieval, content

J. Groppe (✉) · S. Groppe · R. Möller
Institute of Information Systems (IFIS), University of Lübeck, Ratzeburger Allee 160, 23562
Lübeck, Germany
e-mail: groppej@ifis.uni-luebeck.de

S. Groppe
e-mail: groppe@ifis.uni-luebeck.de

R. Möller
e-mail: moeller@uni-luebeck.de

R. Schlichting
Zweckverband Ostholstein (ZVO), Wagrienring 3-13, 23730 Sierksdorf, Germany
e-mail: R.Schlichting@ZVO.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
S. Jain et al. (eds.), *Semantic Intelligence*, Lecture Notes in Electrical Engineering 964,
https://doi.org/10.1007/978-981-19-7126-6_16

205

management and business automation. Text classification can be divided into three types:

- *binary classification* where each text belongs to one of two classes
- *multi-class classification* where each text belongs to one of k classes with $k > 2$
- *multi-label classification* where a text belongs to one or more classes

A dataset is imbalanced if the numbers of texts of classes vary significantly. A multi-label classification over imbalanced datasets is most difficult among all classification problems.

In machine learning, text classification is the problem of finding a classification model from a given dataset, which is a dataset-specific labor-intensive task. Although a number of models of classification over real-world datasets have been reported, they cannot be directly applied to other datasets. In order to address challenges in text classification and to find optimal classification models over specific datasets, Kaggle¹ from Google hosts a number of competitions of classification over different datasets. Widely reported text classification tasks include sentiment analysis on movie reviews of the Rotten Tomatoes [2], the classification of 20 newsgroups of Usenet [12], the classification of toxic online comments of Wikipedia [13], and spam filtering over email datasets [20].

In this work, we report a project of text classification over a real-world dataset of customer communications of a German company, Zweckverband Ostholstein (ZVO). ZVO is a German utility company whose business includes the supply of energy, drinking water, waste water and Internet broadband, and the management of waste and recycling. It serves around 100,000 customers and each year receives around 150,000 customer communications via various communication channels. Each customer communication is read and labeled by its employees, and the labeled communication is then forwarded to the responsible departments and dealt with there.

The dataset of communications contains 18 classes, and each communication can be assigned with one or multiple classes. Although most communications belong to one class, a small number of them has up to 15 class labels. Some classes own a very big number of communications, meanwhile the others have very less communications. Furthermore, this dataset varies extremely in length from one word until 15,471 words and contains also a large amount of noises in association to the characteristics of the business and the process of data obtaining. In order to automate the processing of customer communications and address the challenges that the dataset of customer communications imposes, University of Lübeck and ZVO work together to model the text classification over the multi-label, imbalanced and noisy dataset. We are of the opinion that the dataset and its corresponding classification problem are typical for companies of this size and with frequent customer communications. Hence our results are transferable also to other similar companies.

The aims of this work are to find an optimal classification model and answer the following research questions:

¹ <https://www.kaggle.com/>.

1. What is the best classification algorithm, which can optimally address the challenges imposed by this dataset?
2. Which text encoding model is optimal for the dataset of customer communications given a classification algorithm?
3. Is the performance of the optimal model optimal under the given dataset?

The rest of this paper is organized as follow. Section 2 describes in detail the features of the dataset of customer communications. In Sect. 3 we study and identify appropriate techniques of classification, and based on it we develop in Sect. 4 different classification models. In Sect. 5 we perform an extensive evaluation and discussion and report interesting findings and insights. Section 6 concludes the work.

2 Dataset of Customer Communications

The dataset contains the communications between the company ZVO and its customers, which took place between 2015 and 2020 by email. Each communication was read by the employees and assigned to one or more classes. The dataset has following properties:

- **Real-world business data:** The content of the dataset covers the supply of energy, drinking water and internet broadband, waste management and recycling.
- **German:** The communications are mainly written in German, and so preprocessing of the dataset needs to consider the specific features of the German language. Furthermore, some communications are appended with a footer in different languages, including English, Russian, and Arabic.
- **Noisy:** Specific to the characteristics of the business, the dataset contains a large number of numbers (e.g. utility consumption), person names, street addresses, emails, and URL addresses. Such content typically does not own predictive power. Furthermore, the dataset also contains a large amount of garbage strings such as HTML codes and URIs with messy query parameters. The reason for this is that the content of the communications is not cleanly extracted from HTML files. Figure 1 provides a sample of raw noise data.
- **Length of communications:** The communications vary extremely in length from only one word until more than 10,000 words. This characteristic imposes a big difficulty for any ML algorithm to make a good generalization from such a dataset. Figure 2 depicts the distribution of length of communications.
- **Multi-label:** Each communication can belong up to 18 classes. Table 1 lists the number of classes that a communication can belong to and the corresponding number of communications.
- **Imbalance:** The biggest class has more than 23,000 communications, and the smallest one owns only 43 communications. Table 2 presents the characteristics of imbalance.

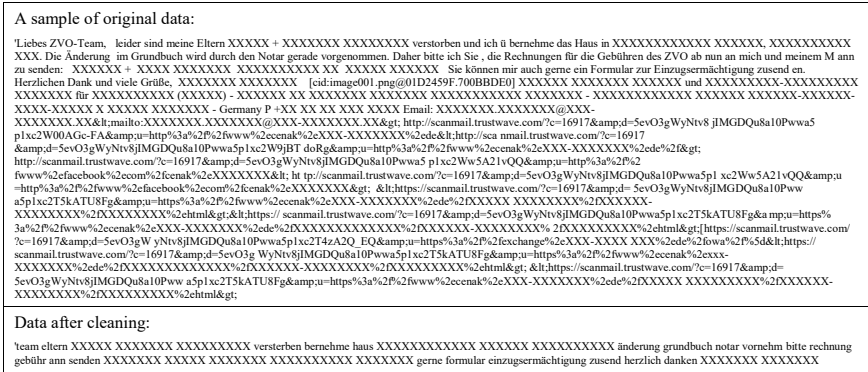


Fig. 1 The dataset is full of noises. In the example, the business-related and personal data are masked

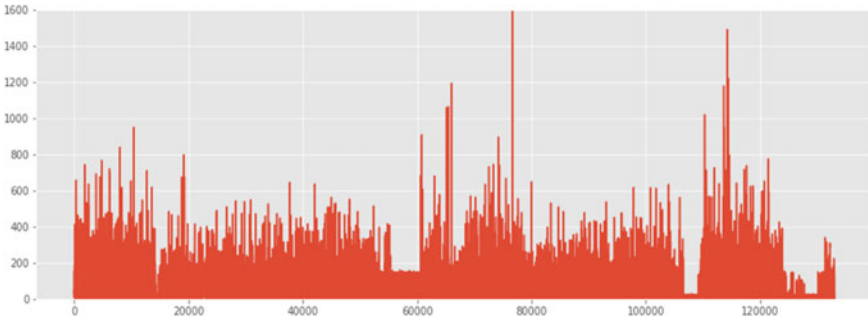


Fig. 2 Length of communications versus the corresponding number of communications. The longest communication has 15,471 words, the shortest one has only 1 word, and average length is 59 words

Table 1 Number of classes that a communication can belong to the corresponding number of communications

#Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
#Sample	119655	12031	1079	199	31	16	6	7	2	3	5	3	2	2	4

Data cleaning: Most noises can be detected and cleaned using the technique of regular expressions. In the process of cleaning, some non-words are created, which can be removed by means of their document frequency. German and the dataset-specific stop-words are filtered out. We use only top 20,000 most frequent words as the features of the dataset (and we will explain this selection later on). Therefore, we further remove the words with document frequency ≤ 10 . After the data preparation, we get 133,045 communications and 20,979 unique words.

Table 2 Number of communications per class

Class	C0	C1	C2	C3	C4	C5	C6	C7	C8
#Comm	9770	21062	14245	235	1307	251	4009	3610	9410
Class	C9	C10	C11	C12	C13	C14	C15	C16	C17
#Comm	23533	20676	43	2866	9617	8144	12126	4748	2700

3 Selection of Modeling Techniques

A number of techniques have been developed for the problem of classification. The first step in finding an optimal model for a dataset is to identify and select appropriate modelling techniques.

3.1 Classification Algorithms

Classical machine learning classification algorithms such as SVM [4], Random Forests [19], and XGBoost [6] can achieve good performance with less training data and need less computational resources. However, these algorithms are designed for multi-class or binary classification problems. To apply them to multi-label datasets, methods of problem transformation [29–32] or techniques of algorithm adaptation [8, 22, 36] are needed. They convert multi-label problems into a number of binary or multi-class classification problems where the relevance of labels must also be considered. This increases the computational complexity, and the advantages of classical ML algorithms are therefore greatly weakened or disappear.

In comparison, neural network-based deep learning classification algorithms are one for all approaches. The differences are only the number of neurons and the activation function of the output layer. Various types of neural networks have been developed for classification problems and they own different capability and computational complexity.

- *Multilayer Perceptrons (MLP)*: MLPs are a fully connected network (FCN), where every single neuron in a layer is connected to each neuron in the following layer and each neuron has an activation function that performs nonlinear transformation over input data. Such an architecture makes MPLs universal function approximators [10]. The FCN is a main component in all types of neural networks.
- *Convolutional Neural Networks (CNN)*: CNNs consist of a convolution network and a FCN. The convolution part is a filter-based feature extraction model and is capable to find the location-independent features that are important for classification. CNNs are originally developed for image classification and also show good results in text classification [7].
- *Long Short-Term Memory Networks (LSTMs)*: They introduce sophisticated memory cells into the architecture of recurrent neural network (RNN) and so they

address the gradient vanishing problem of traditional RNNs and have capability to remember data over long time intervals. Gated Recurrent Units (GRUs) [9] are a variant of LSTM, which is developed for reducing the complexity of LSTM cells. A GRU cell uses two gates and a LSTM cell uses three gates, so a GRU is only slightly less complex than a LSTM. We will adopt the architecture of bidirectional LSTM, and it can capture sequential information and word dependency in both directions backwards and forwards.

- *Transformers*: The technique was developed in 2017 by Google Brain [34] also for sequence learning problem. Transformers do not process the data in order like RNN-based models, rather adopting the self-attention mechanism to decide the importance of each word in a sentence. This feature allows for more parallelization than RNN-based models. A big weakness of Self-attention is that it is very computationally intensive. In our current experimental environment, training a BERT-Transformer over a batch of 32 instances took 40s and 1 Epoch 37h, while other models needed only 40–350m over the same size of batch and 1 Epoch 136–1160s. In a very preliminary test, the BERT-Transformer did not show a good performance. Therefore, in our work we will not adopt this model.

3.2 Text Encoding

All ML algorithms do not directly work with text data, and they must be transformed into numerical forms. Bag of Words (BoW) and Term Frequency Inverse Document Frequency (TF-IDF) are classical text encoding and feature extracting models, and the word embedding is a technique of text encoding used in deep learning techniques.

The BoW model encodes a text document by counting the occurrences of words in the text. The more often a word appears in a text, the more important it is to the text. In our work, word counts are regularized using the length of documents,

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

where $f_{t,d}$ is the raw count of a term t in a document d . TF-IDF refines the TF model by additionally considering word occurrences in the whole corpus. It weights the counts of a word using the inverse document frequency of the word and so gives the document-specific words more importance. We will use a smoothed version of IDF, $idf(t, D) = \log\left(\frac{N}{1 + n_t}\right) + 1$, where N is the number of documents in the corpus D and n_t is the number of documents where the term t appears.

Word embeddings represent a word as a real-valued vector and a text as a sequence of vectors. They can be learned by a classification model itself during training, or they can be pre-trained using a word-embedding technique. Pre-trained word embeddings over a big corpus can capture the context information of words and words with similar context have similar vectors. GloVe [28] and Word2Vec [26] are popular methods to learn embedding vectors for individual words. FastText [3] is built on Word2Vec and can learn word representations from the level of characters, such that it can

generate embedding vectors for words out of vocabulary. Word vectors can have arbitrary dimensions and pre-trained word embeddings typically have 100, 200 and 300 dimensions. The more the dimensions of embeddings, the more they contain the information of context. We will use self-learning embeddings and a pre-trained GloVe embedding model² and a FastText embedding model.³ All embeddings have 300 dimensions.

BoW and TF-IDF are simple and easy to understand, but they do not capture the sequence of words and their context information. In contrary, encoding text using word embeddings preserves the sequence of words and captures the context information of words, but the meaning of word embeddings is difficult to explain and computationally intensive. MLP networks work typically with BoW and TF-IDF models, and CNNs and LSTMs with word embeddings.

3.3 Approaches to Class Imbalance

The imbalance of classes could lead to two main problems: (1) Unequal prior probabilities bias the decision in favor of the a priori more likely class as shown by the Bayesian decision theory [1, 24]. (2) Minority classes might not be able to exhibit enough patterns. Therefore, different techniques have been developed to address the problems of class imbalance. Undersampling (dropping samples from big classes) is not a good strategy because each data collected is valuable. Simple oversampling (duplicating examples from the minority class) will cause the problem of overfitting. SMOTE [5] and Generative Adversarial Networks (GANs) [14] are sophisticated oversampling techniques that generate new samples from existing samples. Since we are processing a multi-label problem and each sample could have more than one label, the application of these methods will typically not make the classes balanced. Therefore, in our work, we are not going to adopt oversampling techniques.

We will use the strategy of class weights to mitigate the decision bias caused by class imbalance. We give more weightage to minority classes in the cost function of the algorithm so that it could provide a higher penalty to misclassification for the minority classes and such “pay more attention” on reducing their errors. We will give the class weights inversely proportional to their respective frequencies. The weight W of a class C in a dataset D is calculated as $W_C = \frac{\#samples_D}{\#samples_C} * \frac{1}{\#classes}$ where the second part is a normalizing factor.

² The Glove embeddings are trained by deepset.ai over a German wikipedia corpus of 500 GB around 2018.

³ The FastText embeddings are trained by fasttext.cc over the data collected from commoncrawl.org and wikipedia.org in 2018. They are computed using CBOW with position-weights and character n-grams of length 5, a window of size 5 and 10 negatives.

4 Models

Based on the study of modeling techniques, we will develop 12 models and from them find the optimal one. They are two MLP models (MLP-TF with TF encoding and MLP-TFIDF with TF-IDF), five CNN-based models (CNN-sl with self-learning embeddings, CNN-G with the pre-trained GloVe embeddings where the embeddings are not updated during training, CNN-G-u with the update of embeddings, CNN-F and CNN-F-u with the pre-trained FastText embeddings) and five bidirectional LSTM models (LSTM-st, LSTM-G, LSTM-G-u, LSTM-F and LSTM-F-u). In order to define an optimal architecture for each model, a large amount of tests are performed. The starting points of our tests are based on the theoretical characteristics of neural networks and rules-of-thumb reported by researchers and practitioners [16, 23].

Vocabulary: We will use top 20,000 most frequent words in the dataset as its features. The research in the linguistic field has found that typically native speakers of a language know 15,000–20,000 word families [15]. We also tested some models with more features and results showed that using more features than the threshold contributes very little to the performance of classification and could even lead to overfitting. For CNNs and LSTMs, we also need to regularize each text to a fixed length. Since 97% customer communications have a length <200, the length of 200 is a reasonable choice. As a result, for MLPs, each text is encoded as a vector with 20,000 elements and for CNNs and LSTMs, each text is encoded as a two-dimensional data of 200 * 300 elements.

Architecture: We use a shallow fully connected network with only one hidden layer. Theoretically, one hidden layer can approximate almost any function. We test some FCNs with one, two and three hidden layers and experimental results showed that FCNs with two and three hidden layers have no obvious advantages. The number of neurons in the hidden layer is decided based on the geometric pyramid rule proposed by Masters [23]: $\sqrt{m * n}$ where m is the number of input neurons and the number of classes. As a result, MLPs have 600, CNNs 84 and LSTMs 68 neurons in their hidden layer respectively.

With CNNs, we use three convolution layers of 128 filters each and each layer is stacked under the input layer and extracts 128 features from each text. The filters in the three layers have small perceptive fields of 1, 2, 3 respectively, and this means that CNNs will consider 1-gram, 2-gram and 3-gram tokens. The more grams, the more they can capture context around each word. However, using 4-grams or higher will produce a much larger and sparser feature set and also cause the problem of overfitting [11].

LSTMs have a bidirectional layer, which propagates the input forwards and backwards through the LSTM layer with 128 nodes, and this helps LSTM to learn long term dependencies. The output from the bidirectional layer are 256 values, because it doubled what we put in LSTM.

Regularization: We use the dropout [17, 18] as regulation mechanism to improve the generalization ability of models and reduce overfitting. The 50% of the output

of the hidden layer at MLPs and the 50% of the output of the bidirectional layer in LSTMs are randomly dropped out. For CNNs, we use the global max-pooling method [35] to perform regularization.

5 Evaluation

Given a training dataset, a model learns the parameter settings, which optimally generalizes the patterns of classes. Optimal model parameters are learned by minimizing the cross-entropy loss [27], which especially penalizes those predictions that are confident but wrong. We use 80% of the dataset of customer communication as training data and 20% for evaluation. The models are trained with the batch size of 32 and learning rate of 0.001, and the Adam optimizer [21] is used to minimize the loss function. Each model is trained 5 times and the training stops after 5 further training epochs when the validation loss reaches its minimal value. From the 5 times of training, the models with best exact match rate are used to predict and therefore only their results of evaluation are compared and presented here.

5.1 Overall Performance

Table 3 presents the overall performance of the 12 models averaged over samples.

Accuracy measures the rate of exact match, where all labels predicted for a sample must exactly match the given labels of this sample. This is a harsh metric since it ignores samples with partially correct and consider them to be incorrect. LSTM-G-u shows the best exact match rate and CNN-G has the lowest exact match rate and the difference between them is 10.6%. However, not all LSTMs are better than their CNN counterparts. In fact, CNN-sl is the second best model. At first glance, the results seem quite stochastic. However, at a closer look, we do discover some interesting phenomenons: (i) All models with word embeddings updated during training are better than ones where word embeddings are not co-trained. (ii) Pre-trained embeddings over a huge corpus are not necessarily superior to the embeddings that are randomly initialized. (iii) Since not all LSTMs are better than their CNN counterparts, the type of neural network (LSTM or CNN) is not the only performance-boosting factor in itself.

Precision, $\frac{1}{n} \sum_{i=1}^n \frac{y_{pred} \cap y_{true}}{|y_{pred}|}$, measures the ability of the model not to predict a negative sample as positive. A low precision means that a model makes a lot of False Positive predictions. **Recall**, $\frac{1}{n} \sum_{i=1}^n \frac{y_{pred} \cap y_{true}}{|y_{true}|}$, measures the ability of the classifier not to predict a positive sample as negative. A low recall indicates that a model predicts too many False Negative predictions, and this means that the model is lack of the ability of finding positive samples. The results of evaluation over the

Table 3 Performance of 12 models in terms of sample averages

Model	MLP-TF	MLP-TFIDF	CNN-sl	CNN-G	CNN-G+u	CNN-F	CNN-F+u	LSTM-sl	LSTM-G	LSTM-G+u	LSTM-F	LSTM-F+u
Accuracy	55.1	53.2	58.7	50.6	57.5	55.1	56.4	57.4	53.5	61.2	51.2	58.0
Precision	61.6	59.2	66.3	57.2	64.8	62.4	64.4	64.4	60.0	69.0	57.3	65.2
Recall	60.4	58.2	65.8	56.4	64.1	62.1	64.3	62.6	58.3	67.9	55.6	63.7
AUC	79.7	78.7	82.2	77.6	81.5	80.4	81.5	80.6	78.5	83.2	77.2	81.1
Time (s)	961	285	1388	1805	1674	2282	1005	5728	14817	7973	22757	5775

two metrics are similar to the results of exact match and show the above findings once again.

AUC measures the ability of a model to distinguish between the positive class and the negative class. AUC is computed from Receiver Operating Characteristic (ROC) curves and summarizes the curves in a single value between 0 and 1. When AUC is 1, the model is perfectly able to distinguish between positive and negative classes. When AUC is approximately 0.5, the model has no discrimination capacity. When AUC is approximately 0, the model is actually reciprocating the classes and it means the model is predicting a negative class as a positive class and vice versa. The results of evaluation over this metric also show the above findings.

Training time is measured to quantify the computational complexity of different models. What is unexpected is that the models without updating embeddings need much more training time in order to reach their best performance (which is however obviously, although not significant, lower than their counterparts with the update of embeddings).

5.2 Performance Over Individual Classes

To understand why our models get such results, let's dig deeper and see how they perform in each category. Table 4 presents the evaluation over individual classes of LSTM-G-u, which has the best overall performance.

Our models are completely unable to predict C11. This is no wonder, because the class only has 34 training data, which is far too little for training any models. C3 and C5 have also very less training data, which are not enough to yield (reliable) generalizable results. In order to have a reasonable capability of classification for the models, much more training data for these classes needs to be collected. In comparison, C10 has a large training data (16,559) and thus our models have a strong prediction ability for the class.

Conceptually, we should expect that the more training data a class has, the better it can be predicted. This is generally true, and however, we also see several expectations. For example, C9 has 2,209 training data more than C10 and even 10,968 more than C0. However, the evaluation results for C9 were clearly lower than for C0 and C10. The reason for this should be the quality of data. Although C9 has the largest training data, they do not show a good pattern. In contrary, the training data of C0 have more in common and so they have a better quality. Therefore, in addition to the quantity, the quality of the training data also plays a very important role.

The biggest exception is C1. It has 16,824 training data and is the second largest class after C9. However, our models are hardly able to predict it. The recall and AUC values are just slightly better than C12 that has however only 34 training data. This indicates a very low commonality and the reason for this should be in the class itself, which is labeled as 'General Correspondence'. As the name suggests, the customer communications in the group could be very diverse and varied. This means that these

Table 4 Performance of LSTM-G-u over individual classes

Class	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17
Precision	87.3	63.5	88.4	21.0	46.9	17.8	74.7	52.9	78.9	81.7	85.3	0.00	45.5	58.5	69.9	69.5	64.3	44.7
Recall	80.0	27.4	81.0	46.6	56.2	42.8	76.6	58.5	78.3	65.9	86.9	0.00	55.0	50.8	66.7	69.4	55.5	52.5
AUC	89.5	62.2	89.8	73.1	77.8	71.2	87.9	78.5	88.4	81.3	92.1	50.0	76.8	73.9	82.4	83.2	77.2	75.6
#eval	1970	4238	2858	45	265	49	809	719	1881	4765	4117	9	547	1948	1581	2417	939	518
#train	7800	16824	11387	190	1042	202	3200	2891	7529	18768	16559	34	2319	7669	6563	9709	3809	2182

communications have little in common, even though they belong to the same group. This could imply that collecting more data would not bring much for this class.

5.3 Separability of Classes

The results of evaluation have indicated that different classes have different quality. The quality of data can be reflected from two aspects. When the data of a class has a large stochastic behavior, any ML algorithms will not be able to find a pattern in order to predict it, no matter how large the dataset is. If two classes have many similar data, it is difficult for a ML algorithm to distinguish one from another. We can observe the stochastic nature and the separability of classes by the visualization of data. There are a number of techniques of dimension reduction for this purpose, among them t-SNE [33] is widely used for visualizing high-dimensional data. However, we adopt a recently developed technique, UMAP [25], because it can preserve more global structure and is computationally more efficient than t-SNE. Figure 3 presents the results of the visualization of applying UMAP to the dataset of the customer communications of ZVO.

From these scatter plots, we can see positive samples (red points) interleave together with negative samples (gray points) and most points (either red or gray) are concentrated in a small area. They show that the classes have a low separability and it is really hard to find a reasonably good boundary to separate them. Though the low separable characteristic, nevertheless, we can still see some patterns. Among all classes, C10 shows the strongest pattern, our models can best distinguish it as indicated in Table 4. After C10 are C0 and C2. C0 has more similar samples as seen in several intense red areas, and the strong point of C2 is its most points located at a small area. C1 shows a big degree of randomness and the reason for this lies in the nature of this class as discussed earlier. However, no matter how strong or weak these patterns are, they are all submerged in negative samples. Given these characteristics of our dataset, we have reasons to believe that our models have achieved a quite good performance.

5.4 Further Discussion

LSTMs versus CNNs: CNNs are a feature extraction model, which can find a feature no matter where it is located. They are efficient when using small receptive fields, which basically gives us the information about short-term dependency of forward and backward. For considering long-term dependency, we need large receptive fields. In the extreme, the receptive field is the size of the entire input. In this case a CNN essentially becomes fully connected, and such stopping being a CNN.

In contrary, bidirectional LSTMs are designed to support long-term and short-term dependency of sequences with two directions. However, not all LSTM models are

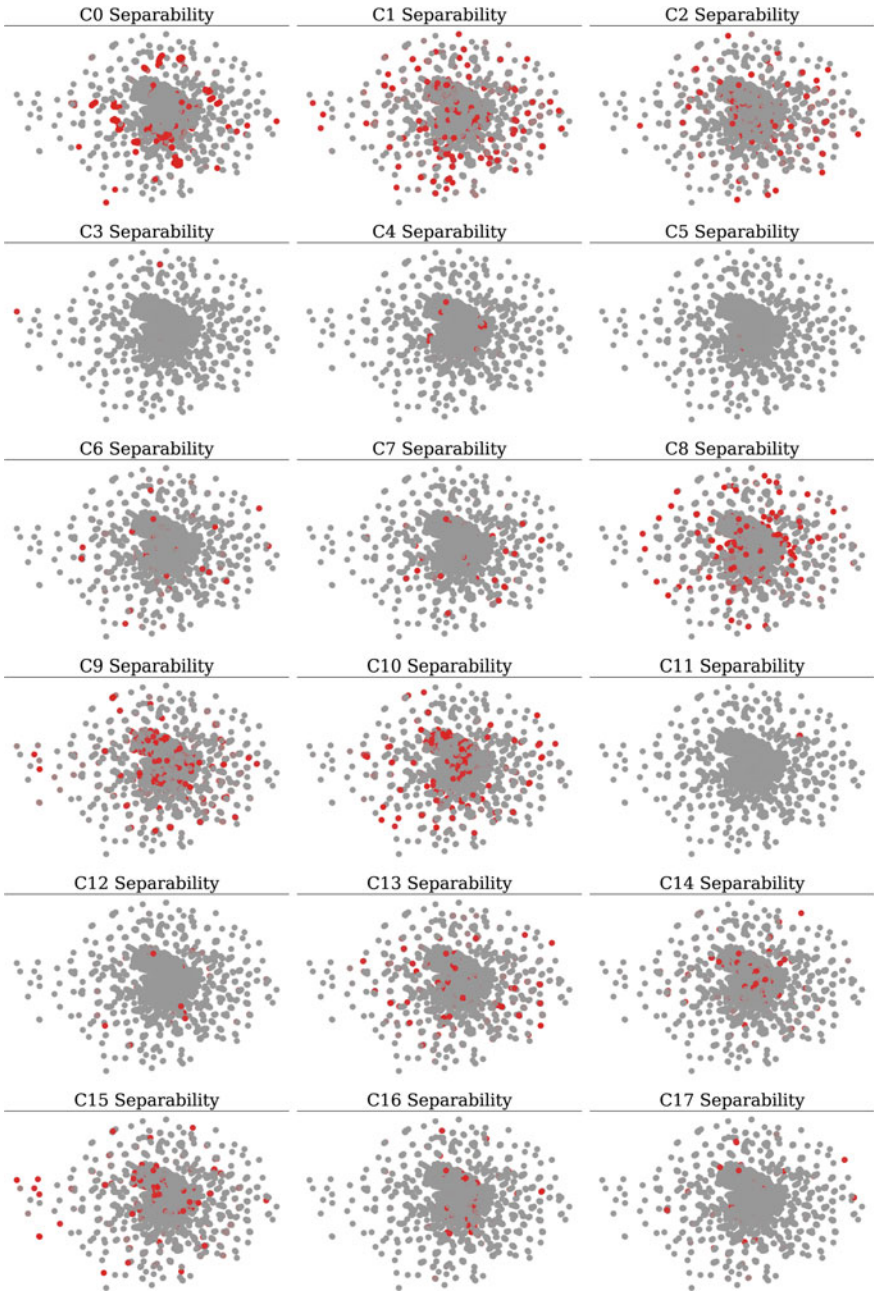


Fig. 3 Distribution of data illustrating the separability of classes and the stochastic degree of sample distribution, where positive samples (red points represents positive samples and gray points negative samples)

better than their CNN counterparts. A possible explanation for this is that LSTMs will show more advantages than CNNs only when the long-term dependency of data plays a more important role than the short-term dependency in determining the meaning of words. The long-term dependency of text sequences is important for language translation, but not necessarily for classification tasks as shown in our evaluation. Our CNNs consider 1, 2, 3-word dependency, and this can identify the context of most words. As shown in language modelling, using 4-word dependency could already lead to overfitting [11]. Considering the computational complex of LSTMs, it would be a good strategy that first to train CNN models and then LSTMs.

Pre-trained versus self-learning embeddings: The GloVe and FastText embeddings that we used are trained over large corpus and contain the information of 300 semantic contexts for each word. The pre-trained embeddings can be co-trained (updated) or not (frozen) when training a model. If we want to use the context information contained in these embeddings, we need to freeze them. It turns out that freezing pre-training embeddings is a bad idea both in terms of the performance and training time. Several models in our work do not use any pre-trained embeddings and instead they learn an embedding model themselves when they learn a solution to classification, and the self-learning embedding models are typically initialized randomly. Our test results show that pre-training embeddings with update is not necessary superior to self-learning embeddings with random initialization, which do not capture any semantics.

When we allow pre-trained embeddings to be re-trained, in fact we are having a model with self-learning embeddings, the only difference is the self-learning embeddings are initialized with a pre-trained embedding model. This initialization could be better or inferior to the random initialization as shown in our evaluation results. No matter what the case is, we must have a pre-training embedding model and we also need to integrate it into the classification model. When using the embedding self-learning approach, we only need to do a random initialization. Therefore, it might be a good idea to start with self-learning of embeddings and then use pre-trained embeddings if any.

That LTSM-G-u is better than LTSM-sl and LSTM-F-u could mean that the GloVe embeddings provide a better initialization of embeddings for the LSTM architecture. Likely, the random initialization of embeddings used in CNN-sl is a better initialization of embeddings than the pre-trained GloVe and FastText embeddings for the CNN architecture. This could imply that the semantic information contained in the pre-trained embeddings might not be really important for the classification problem (at least in our project), and what matters is a good initialization of embeddings.

MLP models: Theoretically, TF-IDF is a refinement of TF. However, MLP with TF is slightly better than with TF-IDF in terms of performance, and this result was a little bit unexpected. The reason could be the special characteristics of our dataset: A large number of very short texts and extremely varied length of texts. Our MLP models are 8–10% lower in performance than the best model, but they can be trained very quickly. Therefore, a classification project could start with MLP models to quickly

get a first feeling about what level of performance classification models can achieve for a dataset, and then proceed to more sophisticated but also more computationally-intensive models.

6 Conclusions

This work aims to find an optimal model of classification over a challenging real-world dataset from a German utility company. Based on the study on the capability of existing classification techniques, we develop multiple deep learning-based classification models with both classical and embedding-based text encoding and feature extracting techniques, an extensive evaluation of these models are performed, and the results of evaluation are in-depth analysed and discussed. The observations and findings from this work not only answer the questions specific to the real-world dataset, but also provide general insights for deep learning-based text classification problems.

Pre-training embeddings integrate the information of contexts, and the higher the embedding dimension, the more refined the context. However, our work shows that using pre-training embeddings are not necessarily better than random initialization of embeddings and keeping the context information of pre-training embeddings does not make sense. This implies that a low-dimensional embedding, e.g. 50 instead of 300 dimensions, might be enough for classification tasks. LSTMs are capable to remember long-term as well as short-term dependencies and CNNs are only efficient for short-term dependencies. However, LSTM models are not necessarily better than CNN counterparts, and this could imply that long-term dependencies, although important for language translation, do not play a big role for classification problems. From the results, we also see that an embedding model is optimal for a classification model but not necessarily also optimal for others. Summarizing all these observations, we can conclude that it is important to find an optimal combination of a model and good initial embeddings. However, finding an optimal initial embedding model is a challenging task.

Furthermore, this work also provides interesting insights from this project for the practical operation in enterprise. At the time of the initial set-up of the digital workflow in 2011 for the distribution of customer communication, machine learning was not yet thought of in the company. With today's insights, the categories could be designed differently without having to significantly change the required workflow. This will offer the opportunity to increase the probability of hits in the future for the models identified as suitable in the project, because stringent work will be being done on the data quality of the individual categories. In the past, a miscast had no major consequence, and a set of communications such as "General Correspondence" collected everything for which there was no suitable category at first glance. In everyday work, speed is of the essence in business. If a clear classification cannot be made immediately for a communication, the employee will select the collection set when classifying and the final classification is done afterwards for archiving

purposes. However, once it is made clear to employees that their actions during classification may in the future reduce the workload during distribution through the use of deep learning models, the motivation of performing this step accurately is significantly increased due to its greater importance. In summary, there are thus two adjusting screws for practical operation in business, a change in categorization and greater care in classification. This is countered by the trend toward ever shorter formulated customer communications, which requires a great deal of knowledge in the distribution process to interpret what the customer really wants, so that the customer's request can be processed completely to the satisfaction of the customer at the very first attempt.

References

1. Bayes T (1763) LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philos Trans R Soc Lond* 370–418
2. Beineke P, Hastie T, Manning C, Vaithyanathan S (2004) Exploring sentiment summarization. In: *Proceedings of the AAAI spring symposium on exploring attitude and affect in text: theories and applications*, vol 39. The AAAI Press Palo Alto, CA
3. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. *Trans Assoc Comput Linguist* 5:135–146
4. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on computational learning theory*, pp 144–152
5. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. *J Artif Intell Res* 16:321–357
6. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp 785–794
7. Chen Y (2015) Convolutional neural network for sentence classification. Master's thesis, University of Waterloo
8. Chen Y-L, Hsu C-L, Chou S-C (2003) Constructing a multi-valued and multi-labeled decision tree. *Expert Syst Appl* 25(2):199–209
9. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*
10. Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst* 2(4):303–314
11. Daniel J, James HM (2000) *Speech and language processing*. Prentice-Hall
12. Dasgupta A, Drineas P, Harb B, Josifovski V, Mahoney MW (2007) Feature selection methods for text classification. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp 230–239
13. Georgakopoulos SV, Tasoulis SK, Vrahatis AG, Plagianakos VP (2018) Convolutional neural networks for toxic comment classification. In: *Proceedings of the 10th Hellenic conference on artificial intelligence*, pp 1–6
14. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: *Advances in neural information processing systems*, vol 27
15. Goulden R, Nation P, Read J (1990) How large can a receptive vocabulary be? *Appl Linguist* 11(4):341–363

16. Heaton J (2008) Introduction to neural networks with Java. Heaton Research, Inc
17. Hertz J, Krogh A, Palmer RG, Horner H (1991) Introduction to the theory of neural computation. *Phys Today* 44(12):70
18. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint [arXiv:1207.0580](https://arxiv.org/abs/1207.0580)
19. Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol 1. IEEE, pp 278–282
20. Hu W, Du J, Xing Y (2016) Spam filtering by semantics-based text classification. In: 2016 eighth international conference on advanced computational intelligence (ICACI). IEEE, pp 89–94
21. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)
22. Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S (2012) An extensive experimental comparison of methods for multi-label learning. *Pattern Recognit* 45(9):3084–3104
23. Masters T (1993) Practical neural network recipes in C++. Morgan Kaufmann
24. McGrayne SB (2011) The theory that would not die. Yale University Press
25. McInnes L, Healy J, Melville J (2020) Umap: uniform manifold approximation and projection for dimension reduction, pp 1–63 (arXiv)
26. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781)
27. Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press
28. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
29. Read J, Pfahringer B, Holmes G, Frank E (2011) Classifier chains for multi-label classification. *Mach Learn* 85(3):333–359
30. Soufan O, Ba-Alawi W, Afeef M, Essack M, Kalnis P, Bajic VB (2016) Drabal: novel method to mine large high-throughput screening assays using Bayesian active learning. *J cheminformatics* 8(1):1–14
31. Spolaôr N, Cherman EA, Monard MC, Lee HD (2013) A comparison of multi-label feature selection methods using the problem transformation approach. In: *Electronic notes in theoretical computer science*, vol 292, pp 135–151
32. Tsoumakas G, Vlahavas I (2007) Random k-labelsets: an ensemble method for multilabel classification. In: *European conference on machine learning*, pp 406–417. Springer
33. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(11):2579–2605
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: *Advances in neural information processing systems*, pp 5998–6008
35. Yamaguchi K, Sakamoto K, Akabane T, Fujimoto Y (1990) A neural network for speaker-independent isolated word recognition. In: *First international conference on spoken language processing*
36. Zhang M-L, Zhou Z-H (2007) ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recognit* 40(7):2038–2048

The Trends and Perspectives Track

Short Analysis of the Impact of COVID-19 Ontologies



Sven Groppe, Sanju Tiwari, Hanieh Khorashadizadeh, Jinghua Groppe, Tobias Groth, Farah Benamara, and Soror Sahri

Abstract During the COVID-19 pandemic, researchers started to develop technical approaches to solve the numerous challenges imposed by the new pandemic. One fundamental precondition for research is to make relevant data about the COVID-19 pandemic available in a machine-processable way. For this purpose, COVID-19 ontologies and knowledge graphs have been developed and proposed for many different subareas of COVID-19 applications and research. In this paper, we provide a short analysis of the impact of COVID-19 ontologies.

Keywords COVID-19 · Knowledge graphs · Ontology impact

S. Groppe (✉) · H. Khorashadizadeh · J. Groppe · T. Groth
Institute of Information Systems (IFIS) Universität zu Lübeck, Lübeck, Germany
e-mail: groppe@ifis.uni-luebeck.de

H. Khorashadizadeh
e-mail: khorashadizadeh@ifis.uni-luebeck.de

J. Groppe
e-mail: groppej@ifis.uni-luebeck.de

T. Groth
e-mail: groth@ifis.uni-luebeck.de

S. Tiwari
Universidad Autonoma de Tamaulipas, Reynosa, Tamaulipas, Mexico
e-mail: sanju.tiwari@uat.edu.mx

F. Benamara
IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
e-mail: farah.benamara@irit.fr

S. Sahri
Université Paris Cité, Paris, France
e-mail: soror.sahri@parisdescartes.fr

1 Introduction

The world has abruptly encountered the COVID-19 pandemic, and no one has been prepared to fight against it. It had many adverse consequences, including economic, social, and many more. Many researchers worldwide have been conducted studies on the COVID-19 pandemic [5, 14, 35]. There must be a common vocabulary to share information between researchers. Ontologies are models representing structured and unstructured information with entities, their properties, and relations with each other. Numerous ontologies have been developed distinctively for COVID-19 since then. Some existing ontologies also tried to include COVID-19 data to cover this concept.

This paper has considered several COVID-19 related ontologies, often published together with large instance data forming a knowledge graph. All the outspread COVID-19 associated data and technologies can be organized and integrated using knowledge graphs.

We also devised a first-step approach to measure the impact of ontologies. Our approach is based on ontologies' direct and indirect reuses on other projects or ontologies. The results of this analysis of the impact of COVID-19 ontologies might help data scientists to choose those ontologies promising the best interoperability with other applications and tools. Furthermore, we have reviewed a couple of COVID-19 knowledge graphs and datasets in this paper.

2 Short Analysis of COVID-19 Ontologies

In this section, we want to analyze the properties of ontologies dealing with the COVID-19 virus and pandemic. Numerous ontologies have been developed to represent data and knowledge about the COVID-19 pandemic. In order to restrict the analysis to the most important ones, we consider here only those ontologies, which are contained in a comprehensive repository like BioPortal [38].

2.1 COVID-19 Ontologies in BioPortal

We enumerate the result for searching for “COVID-19” in the class label in Table 1 (based on a search¹ on 8.3.2022): It contains the extracted ontologies,² their domains and corresponding references to scientific literature or web resources of the

¹ The search can be repeated with the following url: <https://bioportal.bioontology.org/search?q=COVID-19>.

² This analysis is only meant as short analysis and does not include all available ontologies related to COVID-19. For example, the search result differs for “COVID-19” or “SARSCoV2”, i.e., BioPortal does not provide any semantic search capabilities. However, in a rigorous future analysis, synonyms of COVID-19 and other related terms should be included in the search query and also other sources for ontologies and knowledge graphs should be considered to get a more complete picture of the

Table 1 The investigated ontologies, their domains and references

Ontology	Domain	Refs.
Medical Dictionary for Regulatory Activities Terminology (MedDRA)	All phases of drug development, health effects and Malfunction of devices	[10]
SNOMED CT (SNOMEDCT)	Clinical terms	[54]
COVID-19—Medical Subject Headings (MESH)	Subject headings in MEDLINE/PubMed, NLM Catalog and Other	[27]
Mapping of Drug Names and MeSH 2022 (MDM)	Links between DrugBank and MESH	[34]
International Classification of Diseases, Version 10—Clinical Modification (ICD10CM)	International comparability for mortality statistics	[55]
Human Disease Ontology (DOID)	Human disease	[52]
Coronavirus Infectious Disease Ontology (CIDO)	Coronavirus infectious diseases: etiology, transmission, pathogenesis, diagnosis, prevention and treatment	[1]
COVID-19 Ontology (COVID-19)	Virus-host-interactions, virus life cycle	[22]
Neuroscience Information Framework (NIF) Standard Ontology (NIFSTD)	Neuroscience data, tools and information	[18]
Homeostasis imbalance process ontology (HOIP)	Homeostasis imbalance (triggered by COVID-19)	[25]
Obstetric and Neonatal Ontology (ONTONEO)	Electronic health records of pregnant woman and baby	[9]
Assessment of Indian Economy During COVID-19 (INBANCIDO)	Extension of CIDO for impact of COVID-19 on Indian economy	[40]
COVID-19 Surveillance Ontology (COVID-19)	COVID-19 surveillance in primary care	[28]
ZonMW COVID-19 (ZONMW-CONTENT)	Predictive diagnostics, treatment, care, prevention, societal dynamics	[29]
Mondo Disease Ontology (MONDO)	Harmonizes disease definitions across the world	[61]
Experimental Factor Ontology (EFO)	Biological process, experimental conditions	[30]
Cell Culture Ontology (CCONT)	Cell lines and culture conditions	[11]
The COVID-19 Infectious Disease Ontology (IDO-COVID-19)	Extension of IDO and VIDO for COVID-19	[1]
COVID-19 Impact on Banking Ontology (COVID-19-IBO)	Impact of the COVID-19 on the banking sector of India	[41]
National Cancer Institute Thesaurus (NCIT)	clinical care, translational and basic research	[19]
Mass Spectrometry Ontology (MS)	Mass spectrometry experiments	[32]
Logical Observation Identifier Names and Codes (LOINC)	Medical laboratory observations	[56]
Veterans Health Administration National Drug File (VANDF)	Clinical drugs, drug classes, ingredients	[37]
MedlinePlus Health Topics (MEDLINEPLUS)	Health terms	[49]
An Ontology for Collection and Analysis of COVID-19 Data (CODO)	Cases on daily basis with geolocation, patient data: symptom, treatment facility, patient's travel history, transmission reason, tracking of patient test results, ...	[8]
Gender, Sex, and Sexual Orientation Ontology (GSSO)	Connecting terms from biology, medicine, psychology, sociology and gender studies	[23]
Vaccine Ontology (VO)	Biomedical ontology in the vaccine domain	[26]
Intelligence Task Ontology (ITO)	Artificial intelligence tasks, benchmarks and benchmark results, including the biomedical domain	[4]
International Classification of Diseases Ontology (ICDO)	International Classification of Diseases (ICD), ICD-10	[62]
VODANA-COVIDTERMS (VODANA-COVID)	COVID-19 generic template used within different facilities in VODANA	[48]

ontologies. There are ontologies especially created for the COVID-19 context like COVID-19, HOIP, INBANCIDO, COVID-19, ZONMW-CONTENT, IDO-COVID-19, COVID-19-IBO, CODO and VODANACOVID, but 70% are ontologies with a general domain, which integrate COVID-19 classes for updating their content to integrate the new virus. One might be astonished that the ontology domains do not overlap so much and many ontologies occupy their own niche. Section 3 discusses also other related ontologies, knowledge graphs and datasets, which are not found in BioPortal.

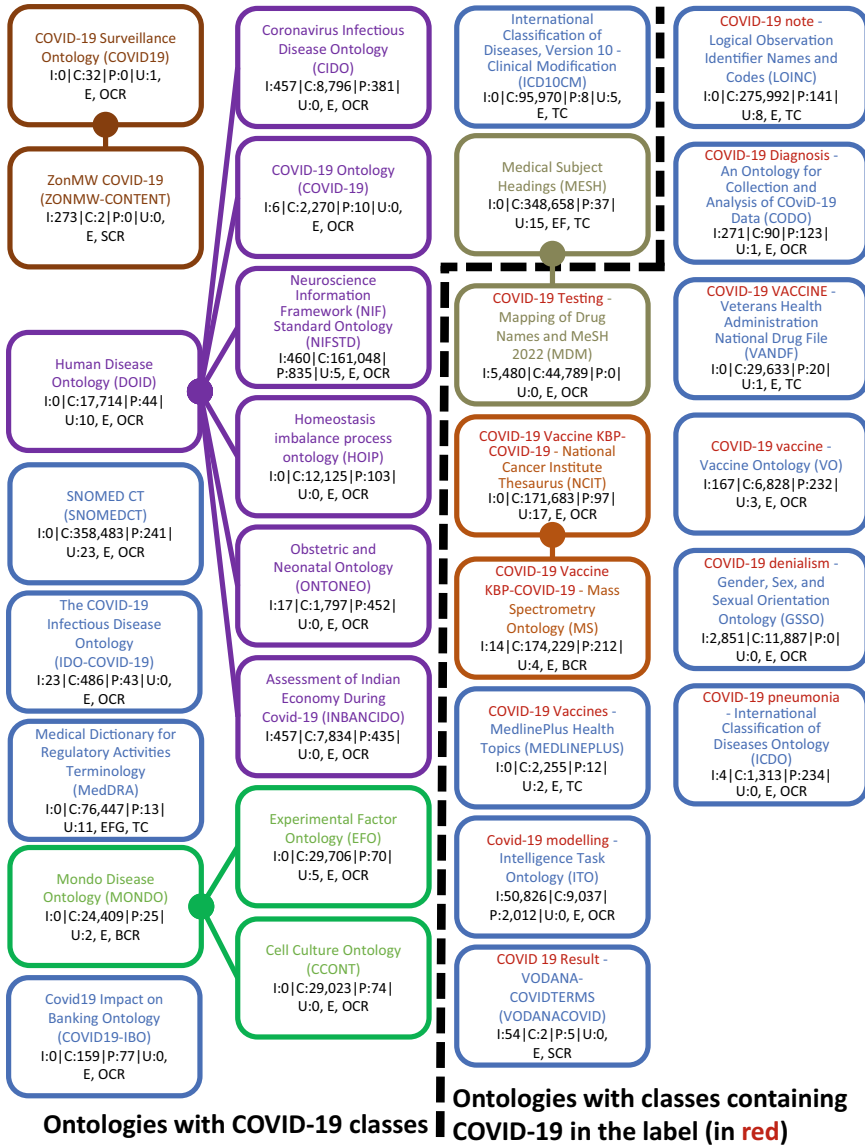
The search result includes 30 ontologies, although only 19 of them integrate classes with a label containing “COVID-19” (see Fig. 1) and 7 of them are exactly matching “COVID-19”. Figure 1 visualizes also ontologies reusing the COVID-19 class of another ontology. We discuss this issue in Sect. 2.2 in more detail. Figure 1 also provides statistics over the number of individuals, classes, properties and usages in projects (as added to BioPortal). It shows also the used ontology language and natural language variants: All ontologies support english labels, but 2 have French versions and one a German version of the ontology. The number of individuals declared in these ontologies is often 0 (50%), but can be up to 50,826 (ITO). The minimum number of classes is 2 (VODANACOVID and ZONMW-CONTENT) and the maximum 358,483 (SNOMED CT). Few have less than 1,000 classes (27%) and between 1,000 and 10,000 classes (20%), Most ontologies have between 10,000 and 100,000 classes (37%) and some few are larger than 100,000 classes (17%). The number of properties is much less than the number of classes ranging from 0 to 2,012 (ITO). 20% of the ontologies have less than 10 properties, 40% have between 10 and 100 properties, 37% have between 100 and 1,000 properties and only 4% (i.e., one ontology) have more than 1,000 properties. We extensively discuss the number of usages in projects in Sect. 2.2.

2.2 *Measuring Impact of Ontologies*

We discuss ways to measure the impact of the ontologies in this section to figure out the most relevant ontologies in the context of COVID-19.

In general, to determine which of several ontologies are better suited for a particular purpose, some criteria must first be available for a *good* ontology. Many ontology metrics and measures have been suggested and several methods tackling these metrics were proposed to evaluate and measure the quality of ontologies. In [15], quality metrics in ontologies were identified and classified into categories. The existing evaluation methods that examine the quality metrics differ on how many of these metrics are targeted, and their main motivation behind evaluating both taxonomic and factual information. Raad and Cruz overview these ontology evaluation methods based on

ontology and knowledge graph landscape related to COVID-19. According to our experiences, there are only few more ontologies in BioPortal search results, such that we focus here on only the keyword “COVID-19” delivering the most and the most related results.



B reuses A I: #Individuals C: #Classes P: #Properties
 U: #Projects using this ontology
 E/F/G: English/French/German version available
 O: OWL S: SKOS B: OBO C: CSV R: RDF/XML T: RDF/TTL

Fig. 1 Ontologies containing a class with label COVID-19 according to a search on 8.3.2022 in the BioPortal [38]

their addressed quality metrics [46]. Once an ontology is checked whether it meets basic quality standards and ensures it fits for the purpose, we tackle other metrics for the ontology impact in the context of COVID-19.

Number of Reuses in other Ontologies: The search result of the BioPortal website already provides information in which ontologies the found class is reused. This number of reusings can be seen as metric for the ontology impact, because a high number of reusings

- is a sign for the popularity of the reused ontology and its classes,
- let ontology users stumble over the reused ontology when applying the reusing ontology, and
- makes more reusings more likely.

For example, the class Person of the widely-known friend-of-a-friend (foaf) ontology is reusing the class Person from schema.org core and is reused in 37 ontologies,³ where we also count the number of indirect usages, i.e., a reused class is reused in another ontology as well.

With 6 times, the class COVID-19 of the Human Disease Ontology (DOID) has been most often reused among the ontologies related to COVID-19 in BioPortal. The class COVID-19 of the Mondo Disease Ontology (MONDO) has been reused 2 times, the ones of COVID-19 Surveillance Ontology (COVID-19), of Medical Subject Headings (MESH) and of National Cancer Institute Thesaurus (NCIT) each have one reuses in other ontologies. All other related ontologies have not been reused so far.

According to this metric, hence only DOID seems to have a high impact, MONDO a moderate, and COVID-19, MESH and NCIT a low impact for COVID-19 related ontologies.

Number of Usages in Projects: This number is an obvious measure for the ontology impact. However, the given number of usages in projects in BioPortal is incomplete⁴ and more a sign for a good maintenance of the ontology itself and motivated ontology developers and project members pointing out these usages of the ontology. Nevertheless, even considering these issues the number of usages in projects remains as a good metric for the ontology impact.

According to the number of usages in projects, we have the following ranking of ontologies related to COVID-19:

1. 32 projects: SNOMED CT
2. 17 projects: NCIT
3. 15 projects: MESH
4. 11 projects: MedDRA
5. 10 projects: DOID
6. 8 projects: LOINC

³ According to a search for Person classes in BioPortal on 9.3.2022.

⁴ For example, according to BioPortal (accessed on 9.3.2022) there are no projects using foaf, which is obviously incomplete, because many projects are using the foaf ontology (please see [https://en.wikipedia.org/wiki/FOAF_\(ontology\)](https://en.wikipedia.org/wiki/FOAF_(ontology)) accessed on 10.3.2022).

7. 5 projects: NIFSTD, EFO, ICD10CM
8. 4 projects: MS
9. 3 projects: VO
10. 2 projects: MONDO, MEDLINEPLUS
11. 1 project: COVID-19 Surveillance Ontology, CODO, VANDF
12. 0 projects: all other ontologies of Table 1

Not surprisingly, established ontologies with a more general domain than COVID-19 are used in many projects and hence are better-ranked according to the metric of project usage. Up so far, ontologies specialized for the COVID-19 domain have at most one usage in projects.

Number of Direct and Indirect Usages in Projects: We define the usage of a given ontology in projects as direct usage and the usage of an ontology reusing the given ontology as indirect usage of the given ontology. In indirect uses, the given ontology itself is also applied, and hence we should take this number for the overall ontology impact into consideration.

As there are only a few ontologies reusing a class containing COVID-19 in the label, also considering the indirect usages (obtained from the result of our search in BioPortal) changes the rankings only slightly: 21 projects are directly or indirectly using NCIT, but NCIT remains in second place with still a big gap to 32 projects of SNOMED CT being on the first place. DOID now achieves 15 (direct and indirect) usages of projects passing MedDRA and gets the same ranking as MESH. There are 7 projects till date directly or indirectly using MONDO, such that MONDO overtakes MEDLINEPLUS, VO, MS, NIFSTD, EFO, and ICD10CM.

Weighted Combinations of Number of Reuses and Projects: Both metrics—the number of reuses in ontologies and number of usages in projects—are independent, at least in theory, because a high number in using projects often results in an increased number of reusing ontologies and vice versa in practice. Nevertheless, there are also examples of a low number of using projects and a high number of reusing ontologies (e.g., according to BioPortal (accessed on 9.3.2022), no projects are using foaf, but the high number of 37 ontologies are reusing the Person class of foaf), and vice versa. Hence, overall a weighted combination of both metrics in one metric seems to be a good idea to provide a good way for calculating a balanced metric for these extreme cases.

For the COVID-19 related ontologies, there are some differences in the rankings according to reused classes and usages in projects, such that a balanced set of weights would lead to a few changes in the rankings. Finding a balanced set of weights is an open question for research and should be based on rigorous analysis of ontologies and their usages in projects and reuses in other ontologies.

3 Related Work

In this section, several studies are considered to analyze the COVID-19 related ontologies, knowledge graphs, and datasets. The central aspect is representing the pandemic data with modeling techniques and methods.

COVID-19 Ontologies: According to the conducted literature, different kinds of research have been done about the COVID-19 pandemic ontologies and knowledge graphs such as Coronavirus Infectious Disease Ontology (CIDO) [16], Covid-19 Ontology (CODO) [8], COVID-19 Surveillance Ontology [28], COVID-19 Ontology [51], Virus Infectious Disease Ontology (VIDO) [1, 3], Controlled Vocabulary for COVID-19 (COVoc)⁵ [42], COVIDCRFRAPID7 ontology [1], Tepuy-COVID ontology [12], iOntoBioethics Ontology [39].

The CIDO ontology was designed under the OBO Foundry approach. The main aim of this ontology is to collectively bring several models to represent pandemic aspects such as common symptoms, similarity to other viruses and drugs offered to treat and diagnose the virus, etc. The CIDO ontology is based on the FAIR (Findable, Accessible, Interoperable, Reusability) principle.

The CODO (Covid-19 Ontology) has been designed to represent and publish the information of the COVID-19 pandemic. This ontology describes the clinical test, patient, current need, trend study, available sources, and growth projections by representing real pandemic cases.

The COVID-19 Surveillance Ontology has been developed as an application ontology for supporting surveillance in initial care. The primary aim of this ontology is to keep COVID-19 data and related respiratory conditions by collecting data from medical record systems.

The COVID-19 ontology has been designed to represent and capture the fundamental entities and concepts needed for COVID-19 research. This ontology also assists semantic interoperability, and text mining approaches in the COVID-19 domain. This ontology creates the basis for developing the referenced namespace in the context of the COVID knowledge graph. This ontology characterizes the roles of cellular and molecular entities in the virus life cycle and virus-host interactions along with medical and epidemiological concepts associated with COVID-19.

The Virus Infectious Disease Ontology (VIDO) includes a virus-neutral extension of IDO Core and terminological content to represent essential viral diseases concepts. VIDO offers a common language to cover viral infectious diseases like CIDO and IDOFLU. This ontology has been designed to bridge the gap between IDO Core and extension ontologies by representing particular diseases and specific causative pathogens.

The COVIDCRFRAPID ontology has been introduced as the semantic data model of the World Health Organization (WHO) COVID-19 rapid version case report form (CRF). It captures the COVID-19 data in semantic context fitting to the CRF of the WHO.

⁵ <https://github.com/EBISPOT/covoc>, visited on 16.3.2022.

Table 2 Examples of COVID-19 datasets

Datasets	Description	Refs.
Kaggle	Kaggle hosts many different datasets and notebooks (containing code for analyzing the hosted datasets) for data scientists including large datasets about the COVID-19 pandemic	[20]
COVID-19 Open Research Dataset (CORD-19)	>500 K scholarly articles, including >200 K with full text about COVID-19, SARS-CoV-2, and related coronaviruses. 17 tasks like “What is known about transmission, incubation, and environmental stability?” >1.6 K Notebooks on Kaggle	[63]
Novel Corona Virus 2019 Dataset	Daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. 7 tasks like “Can We Correlate weather conditions and Corona virus Spread through Data?” >1.5 K Notebooks on Kaggle	[57]
Yahoo COVID-19	Implementation of a COVID-19 KG by Yahoo Knowledge Graph team	[36]
Open COVID-19 Data Working Group ^a	cases of the novel coronavirus with >71 M confirmed cases worldwide	[13]
ECDC-COVID-19	The “European Center for Disease Prevention and Control’s” (ECDC) has acquired various COVID-19 cases and causalities on daily basis from health reports	[44]
Lens COVID-19	Free and open datasets of patent records, scholarly articles metadata, and biological sequences	[24]
2019-nCOV	daily updates of confirmed COVID-19 cases and deaths, active and recovered patients, incident rates, number of people hospitalized and hospitalization rate per nation	[58]
GeoCoV19	A multilingual Twitter dataset based-on a gazetteer-based approach to derive geolocation of tweets	[45]

^a <https://github.com/beoutbreakprepared/nCoV2019> (visited on 16.3.2022)

The COVoc [42], designed as a vocabulary by the European Bioinformatics Institute, is primarily designed for curating and navigating the COVID-19 literature.

The Tepuy-COVID ontology combines several COVID-19 domain ontologies. This ontology models COVID-19 knowledge in different aspects such as treatments, symptoms, socio-cultural elements.

The iOntoBioethics [39] research framework has offered a Bioethics COVID-19 Pandemic Ontology to provide a general denominator to extend and utilize it in a specific healthcare context.

COVID-19 Knowledge Graphs: KG-COVID-19 [47] is specified as a framework to create a knowledge graph for different applications along with machine learning tasks, browsable user interfaces, and hypothesis-based querying to address the issue of harmonizing COVID-19 data and explore relationships. It is also called a KG-hub to form a knowledge graph for SARS-COV-2 and COVID-19.

The COVID-19 Knowledge Graph [7], also termed as COVID-19 Pathophysiology Knowledge Graph⁶ is an extended cause-and-effect network derived from COVID-19 scientific literature to provide its comprehensive aspect by a web application.

DRUGS4COVID195 [2] has been designed as a knowledge graph and a search engine to explore the COVID-19 dataset by identifying the relations among disease, drugs, and texts. This knowledge graph⁷ describes the medication entities like drug, disease, effect, cause, disorder, chemical substance, and symptoms.

COVID-19 KG [21] is a domain-specific knowledge graph for doctors, epidemiologists, policymakers, and other domain experts. There are many more knowledge graphs about COVID-19 available. Please see, e.g., [5] for a survey about COVID-19 knowledge graphs.

COVID-19 Datasets:

There are promising COVID-19 datasets [6, 60] available to curate one or more knowledge graphs. Table 2 has represented some example datasets of different domains.

These datasets play a significant role during the curation of COVID-19 knowledge graphs as these data sources are helpful to disease forecasting and surveillance during the pandemic outbreak. Shuja et al. [53] have presented an extensive survey about COVID-19 datasets.

COVID-19 knowledge graphs construction: Textual contents are important sources for KG acquisition, where natural language processing techniques are used to identify entities, link them to the semantic web and extract semantic relations between entity mentions [31]. Approaches to extract relations at the sentence level are often casted into a multiclass classification problem and range from standard feature-based approaches to deep learning methods where both knowledge-agnostic (relying on the dense representation of words along with additional syntactic information) [66] and knowledge-informed models (that account for semantic ambiguities about entities)

⁶ <https://github.com/covid19kg>, visited on 16.3.2022.

⁷ <https://github.com/oeg-upm/drugs4covid19-kg>, visited on 16.3.2022.

have been proposed [43]. Semi-structured contents are also widely used such as table extraction that targets web tables (but also lists, links, etc.) [17].

These techniques have been used to capture relevant relations in the context of COVID-19. For example, [64] extract fine-grained multimedia knowledge elements (entities, relations and events) from scientific articles and then exploit the constructed knowledge graphs to answer questions in drug repurposing reporting. Michel [33] propose the Covid-on-the-Web project that aims to allow biomedical researchers to access and query COVID-19 related literature relying on named entity recognition techniques coupled with argument extraction to help clinicians analyze clinical trials and make decisions. Relying on the COVID-19 Open Research Dataset, [65] focus on complex relationships between COVID-19 scientific articles that occur across documents. In order to check the validity of the extracted relations, [50] propose COVID-Fact, a tool to detect true claims and their source articles and then generate counter-claims using automatic methods.

Reuses and Impact of Ontologies: To the best of our knowledge the impact of ontologies has not been explored so far in the scientific literature. However, some research already provides statistics about reusing ontologies in other ontologies like the water domain [59].

4 Summary and Conclusions

There is a need for standardized data schemes and knowledge representation for COVID-19 research. Hence numerous efforts are developing and proposing COVID-19 ontologies and knowledge graphs for different subareas of COVID-19 applications and research. We describe some important COVID-19 ontologies in this paper and try to measure their impact by a short analysis based on their usages in projects and other ontologies. In this way, we identify essential ontologies which might support the choice of supported ontologies in an application for optimal interoperability with other applications and tools.

Acknowledgements This work is jointly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 490998901, and the French National Agency Quality-Ont ANR-21-CE23-0036-01.

Funded by



Deutsche
Forschungsgemeinschaft

German Research Foundation

References

1. Babcock S et al (2021) The infectious disease ontology in the age of COVID-19. *J Biomed Semant* 12(1). <https://doi.org/10.1186/s13326-021-00245-1>
2. Badenes-Olmedo C et al (2020) Drugs4Covid: drug-driven Knowledge exploitation based on scientific publications. In: arXiv preprint [arXiv:2012.01953](https://arxiv.org/abs/2012.01953)
3. Beverley J et al (2020) Coordinating coronavirus research: the COVID-19 infectious disease ontology. Open science framework. <https://doi.org/10.17605/OSF.IO/7EJ4H>
4. Blagec K et al (2021) A curated, ontology-based, large-scale knowledge graph of artificial intelligence tasks and benchmarks. arXiv 2110.01434. <https://doi.org/10.48550/ARXIV.2110.01434>. <https://arxiv.org/abs/2110.01434>
5. Chatterjee A et al (2021) Knowledge graphs for COVID-19: an exploratory review of the current landscape. *J Personal Med* 11(4):300. <https://doi.org/10.3390/jpm11040300>
6. Dogan O et al (2021) A systematic review on AI/ML approaches against COVID-19 outbreak. *Complex Intell Syst* 7(5):2655–2678
7. Domingo-Fernández D et al (2021) COVID-19 knowledge graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics* 37(9):1332–1334
8. Dutta B, DeBellis M (2020) CODO: an ontology for collection and analysis of COVID-19 data. arXiv preprint [arXiv:2009.01210](https://arxiv.org/abs/2009.01210)
9. Farinelli F (2021) Obstetric and neonatal ontology. <https://bioportal.bioontology.org/ontologies/ONTONEO?p=summary>
10. Fescharek R et al. (2004) Medical dictionary for regulatory activities (MedDRA). *Int J Pharm Med* 18(5). ISSN: 1179-1993. <https://doi.org/10.2165/00124363-200418050-00001>
11. Ganzinger M et al (2012) On the ontology based representation of cell lines. *PLoS ONE* 7(11) Kannan N (ed). <https://doi.org/10.1371/journal.pone.0048584>
12. González-Eras A et al (2022) Ontological engineering for the definition of a COVID-19 pandemic ontology. *Inform Med Unlocked* 28:100816
13. Open COVID-19 Data Working Group. Detailed Epidemiological Data from the COVID-19 Outbreak
14. Gruenwald L, Jain S, Groppe S (eds) *Leveraging artificial intelligence in global epidemics*. Elsevier. <https://www.elsevier.com/books/leveraging-artificial-intelligence-in-global-epidemics/gruenwald/978-0-323-89777-8>
15. Mc Gurk S, Abela C, Debattista J (2017) Towards ontology quality assessment. In: Joint proceedings of the 3rd workshop on managing the evolution and preservation of the data web (MEPDaW 2017) and the 4th workshop on Linked Data Quality (LDQ 2017) co-located with 14th European Semantic Web Conference (ESWC 2017), Portoroz, Slovenia, May 28th–29th, 2017, vol 1824. CEUR Workshop Proceedings. CEUR-WS.org, pp 94–106
16. He Y et al (2020) CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Sci Data* 7(1):1–5
17. Hogan A et al (2021) Knowledge graphs. *ACM Comput Surv* 54(4). ISSN: 0360-0300. <https://doi.org/10.1145/3447772>
18. Imam FT et al (2012) Development and use of ontologies inside the neuroscience information framework: a practical approach. *Front Genet* 3. <https://doi.org/10.3389/fgene.2012.00111>
19. Jouhet V et al (2017) Building a model for disease classification integration in oncology, an approach based on the national cancer institute thesaurus. *J Biomed Seman* 8(1). <https://doi.org/10.1186/s13326-017-0114-4>
20. Kaggle (2020) Help us better understand COVID-19. www.kaggle.com/covid19
21. Kejriwal M (2020) Knowledge graphs and COVID-19: opportunities, challenges, and implementation. *Harv Data Sci Rev* 1. <https://hdsr.mitpress.mit.edu/pub/xxl0yk6ux>
22. Alpha Tom Kodamullil (2020) COVID-19 ontology. <https://bioportal.bioontology.org/ontologies/COVID-19?p=summary>

23. Kronk C, Tran GQ, Wu DTY (2019) Creating a queer ontology: the gender, sex, and sexual orientation (GSSO) ontology. In: MEDINFO 2019: health and wellbeing e-networks for all. IOS Press, pp 208–212
24. The Lens (2020) Human coronaviruses data initiative. <https://about.lens.org/covid-19/>
25. Lin AY et al (2021) A community effort for COVID-19 ontology harmonization. In: The 12th international conference on biomedical ontologies
26. Lin Y, He Y (2012) Ontology representation and analysis of vaccine formulation and administration and their effects on vaccine immune responses. *J Biomed Semant* 3(1):17. <https://doi.org/10.1186/2041-1480-3-17>
27. Lipscomb CE (2000) Medical subject headings (MeSH). *Bull Med Libr Assoc* 88(3):265–266. <https://pubmed.ncbi.nlm.nih.gov/10928714>
28. de Lusignan S et al (2020) COVID-19 surveillance in a primary care sentinel network: in-pandemic development of an application ontology. *JMIR Pub Health Surveill* 6(4). <https://doi.org/10.2196/21434>
29. Magagna B (2022) ZonMW COVID-19. <https://bioportal.bioontology.org/ontologies/ZONMW-CONTENT?p=summary>
30. Malone J et al (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics* 26(8):1112–1118. <https://doi.org/10.1093/bioinformatics/btq099>
31. Martinez-Rodriguez, JI, Hogan A, Lopez-Arevalo I (2018) In-formation extraction meets the semantic web: a survey. *Semant Web Preprint*, pp 1–81
32. Mayer G et al (2013) The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. In: Database volume. <https://doi.org/10.1093/database/bat009>
33. Michel F et al (2020) Covid-on-the-Web: knowledge graph and services to advance COVID-19 research. In: *The Semantic Web—ISWC 2020*. Springer International Publishing Cham, pp 294–310. ISBN: 978-3-030-62466-8
34. Müller B (2021) Mapping of drug names and MeSH 2022. <https://bioportal.bioontology.org/ontologies/MDM?p=summary>
35. Müller M, Salathé M, Kummervold PE (2020) COVID-Twitter-BERT: a natural language processing model to analyse COVID-19 content on twitter. arXiv 2005.07503. <https://arxiv.org/abs/2005.07503>
36. Nagpal A (2020) Yahoo knowledge graph announces COVID-19 dataset, API, and dashboard with source attribution. Blogpost
37. Nelson SJ et al (2002) A semantic normal form for clinical drugs in the UMLS: early experiences with the VANDF. In: *Proceedings of the AMIA symposium*
38. Noy NF et al (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 37(2):170–173. <https://doi.org/10.1093/nar/gkp440>
39. Odeh M et al (2021) iOntoBioethics: a framework for the agile development of bioethics ontologies in Pandemics, Applied to COVID-19. *Front Med* 8:530
40. Patel A, Debnath NC (2022) Development of the InBan CIDO ontology by reusing the concepts along with detecting overlapping information. *Inventive Comput Inf Technol* 349–359. <https://doi.org/10.1007/978-981-16-6723-7>
41. Patel A et al (2021) Covid19-IBO: a Covid-19 impact on indian banking ontology along with an efficient schema matching approach. *New Gener Comput* 39(3–4):647–676. <https://doi.org/10.1007/s00354-021-00136-0>
42. Pendlington ZM et al (2020) COVoc: a COVID-19 ontology to support literature triage. Workshop
43. Peters ME et al (2019) Knowledge enhanced contextual word representations. In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pp 43–54. <https://doi.org/10.18653/v1/D19-1005>
44. EU Open Data Portal (2020) COVID-19 coronavirus data. Retrieved 29 May 2020. <https://data.europa.eu/data/datasets/covid-19-coronavirus-data?locale=en>

45. Qazi U, Imran M, Ofli F (2020) GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special* 12(1):6–15
46. Raad J, Cruz C (2015) A survey on ontology evaluation methods. In: *KEOD 2015—proceedings of the international conference on knowledge engineering and ontology development, part of the 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3K 2015)*, vol 2, Lisbon, Portugal, November 12–14, 2015. SciTePress, pp 179–186
47. Reese JT, Unni D, Callahan TJ (2021) KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19. *Patterns* (New York, NY), 2(1)
48. Reisen M et al (2021) Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. *Adv Genet* 2(2). <https://doi.org/10.1002/ggn2.10050>
49. Rodríguez-González A et al (2018) Extracting diagnostic knowledge from MedLine Plus: a comparison between MetaMap and cTAKES approaches. *Curr Bioinf* 13(6):573–582
50. Saakyan A, Chakrabarty I, Muresan S (2021) COVID-Fact: fact extraction and verification of real-world claims on COVID-19 pandemic. *ACL/IJCNLP* (1):2116–2129. <https://doi.org/10.18653/v1/2021.acl-long.165>
51. Sargsyan A et al (2020) The COVID-19 ontology. *Bioinformatics* 36(24):5703–5705
52. Schriml LM et al (2011) Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res* 40(D1):D940–D946. <https://doi.org/10.1093/nar/gkr972>
53. Shuja J et al (2020) COVID-19 open source data sets: a comprehensive survey. *Appl Intell* 51(3):pp. 1296–1325. <https://doi.org/10.1007/s10489-020-01862-6>
54. SNOMED International (2022) SNOMED Home-SNOMED international. <https://www.snomed.org/>
55. Steindel SJ (2010) International classification of diseases, 10th edn, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. *J Am Med Inf Assoc* 17(3):274–282. <https://doi.org/10.1136/jamia.2009.001230>
56. Stram M et al (2019) Logical observation identifiers names and codes for laboratorians: potential solutions and challenges for interoperability. In: *Archives of pathology & laboratory medicine*, vol 144, no 2, pp. 229–239. <https://doi.org/10.5858/arpa.2018-0477-RA>
57. Rajkumar S (2019) Novel corona virus 2019 dataset. <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>
58. Hopkins J (2020) University Center for Systems Science and Engineering. Novel coronavirus (COVID-19) cases. <https://github.com/CSSEGISandData/COVID-19>
59. Tiwari S, Garcia-Castro R (2022) A systematic review of ontologies for the water domain. In: Mehta S et al (ed) *Tools, languages, methodologies for representing semantics on the web of things*. ISTE Science Publishing Ltd
60. Tiwari SM, Gaurav D, Abraham A (2020) COVID-19 outbreak in India: an early stage analysis. *Int J Sci Rep* 6(8):332. <https://doi.org/10.18203/issn.2454-2156.intjsci20203117>
61. Vasilevsky N et al (2020) Mondo disease ontology: harmonizing disease concepts across the world. In: *CEUR workshop proceedings*, vol 2807. CEUR-WS
62. Wan L et al (2021) Development of the international classification of diseases ontology (ICDO) and its application for COVID-19 diagnostic data analysis. In: *BMC bioinformatics*, vol 22, no 6. <https://doi.org/10.1186/s12859-021-04402-2>
63. Wang LL et al (2004) CORD-19: the COVID-19 open research dataset. In: *arXiv* 2004.10706. <https://arxiv.org/abs/2004.10706>
64. Wang Q et al (2021) COVID-19 literature knowledge graph construction and drug repurposing report generation. In: *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies: demonstrations*. Association for Computational Linguistics, June 2021, pp 66–77. <https://doi.org/10.18653/v1/2021.naacl-demos.8>
65. Wise C et al (2020) COVID-19 knowledge graph: accelerating information retrieval and discovery for scientific literature. In: *Proceedings of knowledgeable NLP: the first workshop on integrating structured knowledge and neural networks for NLP*. Suzhou, China: Association for Computational Linguistics, pp 1–10. <https://aclanthology.org/2020.knlp-1.1>

66. Wu S, He Y (2019) Enriching pre-trained language model with entity information for relation classification. arXiv preprint [arXiv:1905.08284](https://arxiv.org/abs/1905.08284)

Demystifying Semantic Intelligence for Enabling Intelligent Applications



Sarika Jain

Abstract An intelligent agent is required to fuse heterogeneous sources of information together for which it should be equipped with both the data-driven (statistical) and knowledge-driven (symbolic) AI disciplines. Semantic Technologies make it possible by creating links between disparate and heterogeneous data. When the data is linked as well as open, it is termed as Linked Open Data (LOD). The Symbolic AI and sub-symbolic AI have to go together. The symbolist approach nowadays is manifested as a knowledge graph that advanced statistics and machine learning can run on top of.

Keywords Artificial intelligence · Machine learning · Knowledge graph · Linked open data · Semantic technologies

1 Introduction

Artificial Intelligence (AI) is intelligence exhibited by a computer software/application/machine. AI is not just one thing but a mesh of technologies that aim to create intelligent machines for mimicking human behavior. Such machines tend to have multiple traits with diverse abilities like to solve complex problems, to perceive, to discover meaning, to reason, to generalize, or to learn from past experience.

1.1 *The Trajectory of Technological Progress*

Let us have a look on the trajectory of technological progress starting from the ninetieth century BC (Fig. 1). Pre-ninetieth century BC, the human race started to use fire and wheel, and the natural language developed too. From ninetieth century BC

S. Jain (✉)

National Institute of Technology Kurukshetra, Kurukshetra, Haryana, India
e-mail: jasarika@nitkkr.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
S. Jain et al. (eds.), *Semantic Intelligence*, Lecture Notes in Electrical Engineering 964,
https://doi.org/10.1007/978-981-19-7126-6_18

241

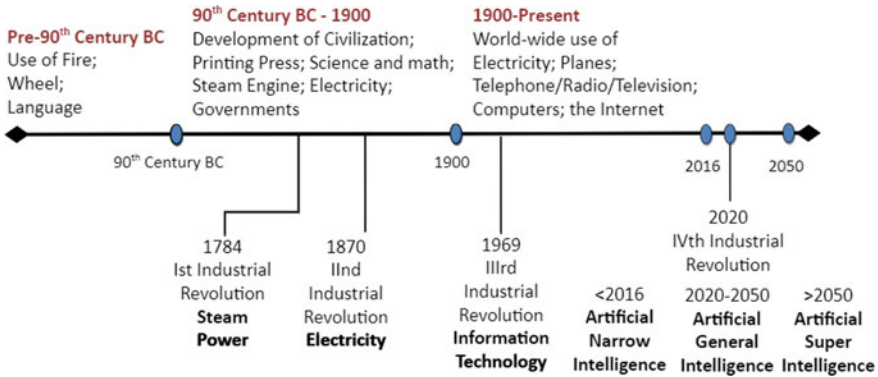


Fig. 1 The trajectory of technological progress

to 1900s, there was development in Civilization, Printing Press, Science and Math, Steam Engine, Electricity, Governments, and more. During the same reign, came the Ist Industrial Revolution in 1784 (Steam Power), and the IInd Industrial Revolution in 1870 (Electricity). From 1900 till the present time, there is worldwide use of Electricity, Airplanes, Radio, Television, Telephone, Computers, and the Internet. This is the time when IT—Information Technology came during the 3rd Industrial Revolution in 1969. Today is the age of the 4th Industrial Revolution in 2020 with three levels of Technological Progress.

- Before 2016, it was **Artificial Narrow Intelligence (ANI)**: It is goal-oriented intelligence with abilities to carry out a specific task at a time like play chess; spot spam email; driver less cars; voice assistants (SIRI or Alexa); play the game Go (alphaGo)). ANI is in fact, the current state of the art and is also termed as Weak AI or Narrow AI.
- From 2020, AI’s Golden Age has begun with the **Artificial General Intelligence (AGI)**, i.e., intelligence with abilities to carry out any task (and not any specific task) just like a human being.
- **Artificial Super Intelligence (ASI)** to come from 2050 onward is the intelligence with abilities not just to mimic human behavior but to surpass even the human capabilities. ASI is able to generate new computer code on its own to help achieve its objectives.

1.2 Intelligent Applications

An intelligent application is an AI incorporated application that is self-aware and enhances user’s experience. It sorts through massive data sets for self training, looking for patterns, and improving upon itself. As a matter of course, Smart or Intelligent applications are developing along two distinct functional use cases [4]:

1. Automating simple routine tasks to save the value time. Example: Personal Virtual Assistants manage schedules and coordinate meetings without user intervention.
2. Providing right data at right time and in proper context. Example: Online Doctor provides decision support.

The software entity that enables *artificial intelligence (AI)* to be put into action by conducting operations in the place of users or programs after sensing the environment is called an *intelligent agent (IA)*. Intelligent agents work autonomously by seeking necessary (present, relevant, and authentic) information, understanding the contents, coordinating with each other, and taking necessary actions to make life simple for human beings. There are three information aspects for an intelligent agent, viz., *Syntax* (sentence construction, grammatical correctness), *Semantics* (human-level interaction), and *Pragmatics* (intention behind the communication. While the semantics refer to the meaning of the sentence on the face (as determined by the language, word sense, and syntax); the pragmatics refer to the intention behind the communication (context) and the action desired from the listener. Pragmatics requires knowledge about the environment (social norms, rules, etc.) and highly depends on the actors.

1.3 Semantic Intelligence

An intelligent agent is required to fuse heterogeneous sources of information together for which it should be equipped with both the data-driven (statistical) and knowledge-driven (symbolic) AI disciplines. We need a representation of our data that not only includes the data itself but where the interactions in it is a first-class citizen.

We have seen in the past decade that statistical models have revolutionized the world. Though the Statistical models have already proved themselves, they are not a Universal Solvent but only a tool as others. Deep learning (DL) is pre-eminent at learning in a static world and executing low-level patterns, provided it is fed with a lot of data. More deep, more intelligent, and of course more black box in nature. The question is “Is the AI of today the Artificial Super Intelligence (ASI)/Artificial General Intelligence (AGI)/Artificial Narrow Intelligence (ANI)? Is the AI of today the AI that we are craving for?” In fact, today’s artificial intelligence is weak AI. There are a number of instances where DL has produced delusional and unrealistic results. Accuracy alone is not sufficient. We require exploring ways of opening the black box of statistical models. When DL researchers are asked to open the black box, this today implies less intelligent models to them (limited capability). In addition to increased performance, AGI aims to build trust.

Symbolic AI and statistical AI have to go together to achieve the AGI. This type of AI is able to generalize and is succinctly referred to as Contextual Computing. The symbolist approach is nowadays manifested as a knowledge graph that advanced statistics and machine learning can run on top of Jain and Murugesan in [5] have defined such a hybrid model as the one that combines machine intelligence with

human intelligence to reach conclusions faster than possible by humans alone along with the explanations needed for trust in the decisions and results; while requiring far fewer data samples for training and conversing in natural language. In addition, semantics have come a long way in enhancing explainability in AI systems.

1.4 Contributions and Organization

Many researchers worldwide have been conducting studies on the efficacy of Semantic Intelligence. Andreas Blumauer, the CEO of Semantic Web Company in his talk “Semantic AI for Legal Experts”¹ at Semantics 2019 asked “Will Artificial Intelligence make Subject Matter Experts obsolete? He asked that while identifying patient-treatment pairings, what will you prefer? An IA solely based on Machine Learning, or an IA solely based on experts’ knowledge (doctors in this case), or an IA that is combination of both. The IA of Minerva Intelligence² thrives in the presence of complex data structures and in domains where training data sets are insufficient for the application of machine learning. Minerva Intelligence merges the modern computational speed and accuracy with the vast knowledge human beings have accumulated to provide cognitive AI solutions.

Semantic Intelligence is a high-impact research field and the current state-of-the-art in AI. Though a plenty of research has already been carried out in this field but by a handful of researchers; thereby it is in its preliminary stage of development. This paper is a vision paper that aims to establish the framework of semantic intelligence to the readers. The document outlines the needs and significance of a contextual system and list down the tech-stack required for its successful realization. It may form the basis of future research in AGI systems.

The paper is further organized in the following manner where Sect. 2 talks in detail regarding how we started extracting knowledge from the information. Section 3 describes the Linked Open Data Cloud and its requirements. While Sect. 4 briefly explains the Semantic Technologies with Sect. 5 about its discussion and insights. Lastly we conclude the paper in Sect. 6 with the summary of the paper and the conclusions.

2 From Information to Knowledge

Knowledge Management is a discipline for the systematic management of the knowledge assets (explicit as well as implicit) with the goal to capture insight and better understanding for meeting the tactical and strategic requirements (achieve wisdom). Issues involved in Knowledge Management:

¹ <https://2019.semantics.cc/semantic-ai-legal-experts>.

² <https://minervaintelligence.com/>.

1. *Knowledge Types*

- Explicit (codified knowledge found in databases or documents)
 - Tacit (Experience-based and context-dependent)
2. **Representation:** Providing high-level descriptions of the world that can be effectively used by computer.
 3. **Storage:** Storing the knowledge as the data is stored.
 4. **Reasoning (Processing):** Finding implicit consequences from explicitly represented knowledge.
 5. **Learning:** Learning from the captured knowledge.

Figure 2 demonstrates the famous DIKW (Data-Information-Knowledge-Wisdom) pyramid. There is a paradigm shift from Data to Knowledge. Data refers to the raw facts, numbers, and symbols. Context and value adds meaning to this data. Consider the data 20062020. If the context is “date”, we can recognize its value as 20th June 2020. It is not necessary for the above data to be recognized as a date only. According to the context it might be called as a ordinal number, or if it is in context of amount or statistics then meaning of the data changes. So in this way it gets meaning and becomes an information. Then comes Knowledge up in the hierarchy and is said to be achieved when the information, that is derived from the collected data, becomes relevant to our goals. The pieces of information connect to add more meaning and value; and can then be applied to achieve our goal as if we have written some rules. Meaning that, knowledge uncovers relationships that are not explicitly stated thereby making the information useful. Understanding is some form of stored knowledge as the mathematical table of 2. When understanding is added, the student can answer difficult questions like $1230 * 45$. In a nutshell, Knowledge is memorizing and Understanding is Learning (synthesizing new knowledge from old knowledge). Wisdom encompasses various said and also unsaid things like the ancient preaching or sayings. Wisdom may even involve intuition.

Data: Tweety, Chichi, Bird, Wings, Fly

Information: Tweety has Wings

Knowledge: If Wings Then Bird; If Bird Then Fly

Wisdom: Tweety has Wings, so Tweety is a Bird, so Tweety can Fly.

We don’t have to look back into the past but look into the future. We are moving from Less Meaning to More Meaning; Less Applicability to more Applicability; Data Management to Knowledge Management; Data Science to Knowledge Science, thus deriving value equivalent to Human. We need to climb up the mountain of wisdom. Semantic Technologies (as discussed in Sect. 4) make it possible by creating links between disparate and heterogeneous data.

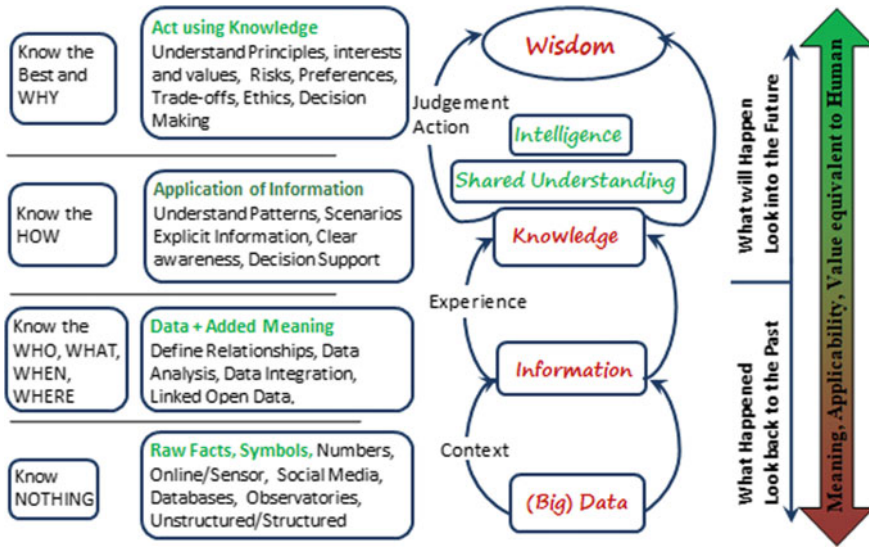


Fig. 2 The DIKW pyramid

3 Linked Open Data (LOD) Cloud

Linked Open Data as the name suggests where all the data will be publicly available in a proper structured format allowing all the metadata to connect and enriched for the ease of the use that too in the machine-readable format. Using Linked Open Data, instead of maintaining different structures for data representation and storage which makes it harder to access and process the data of different representations, we can make a single structured format for the data representation which will be accessible worldwide in a the same format after publishing it on the cloud. Thus if all the datasets are published openly and in the same structuring format it will be possible to interrogate all the datasets at once thus increasing the potential power of analyzing the huge volume of data on the web than the data which is currently available in the form of information silos. This section further briefly describes about the Linked Data, Open Data, and Linked Open Data concepts.

3.1 Linked Data

There are a multitude of disparate sources and formats for every resource on the web. In 1998, the idea of Linked Data was given by Sir Tim Berners-Lee. Linked data interlinks these silos and that too in a machine-readable format forming the **Web of Data**. Understanding graphs is more natural and matches the human like thinking. Graphs helps machine better understand any of the related queries and provide helpful

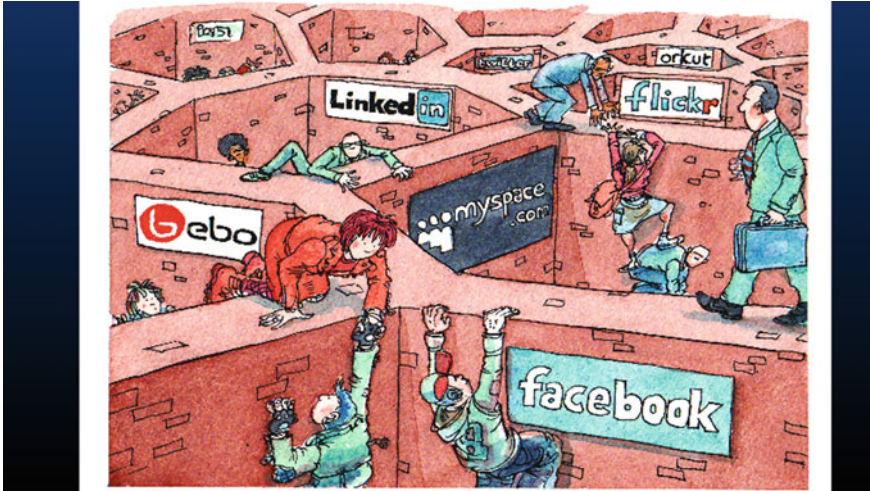


Fig. 3 Closed Data Islands

search cues. The graph data model help the heterogeneous data held in data silos to get mobilized. **Knowledge Graphs (KGs)** are a large network of entities and their semantic types. Knowledge Graphs are based on Linked Data Principles. There is a transition from data silos to interconnected knowledge graphs. Figure 3 shows the Closed Data Islands³ where every website on the web is cooking its own porridge. We need a representation of our data that not only includes the data itself but where the interactions in it is a first-class citizen.

Standard web technologies like HTTP and URIs are used to build Linked Data. Rather than using them as a server web pages for human readable format, it extends itself to share information that can be ready by the computers. This way of reading data automatically by computers enables data from different sources to be connected and queried.

3.2 Open Data

The wealth of data produced by the public sector can be made available to citizens and businesses, creating new business opportunities. When the data is freely accessible to the widest range of users for the widest range of purposes, it is called Open Data.

³ Taken from the slide set presented at the TED 2009 conference, “The Great Unveiling” in Long Beach, CA. USA, 4, Feb 2009. Here is the reference: [https://www.w3.org/2009/Talks/0204-ted-tbl/#\(22\)](https://www.w3.org/2009/Talks/0204-ted-tbl/#(22)).

Reasonable privacy, security, and privilege restrictions may be allowed as governed by other statutes. Open data must be machine-processable with non-proprietary data formats.

3.3 *Linked Open Data*

When the data is linked as well as open, it is termed as Linked Open Data (LOD). LOD comprises of a set of standards and principles for publishing, sharing, and interrelating structured knowledge. LOD makes the open data fully beneficial by providing interoperability and standards. Data and information is put into some context that can create new knowledge to enable powerful applications and services. It facilitates innovation, information management, and integration.

The Linked Open Data (LOD) Cloud depicts publicly available linked datasets. LOD started as a community effort in 2007 with only 12 interlinked datasets. Then, major industry players like BBC, Google, Yahoo, NY Times, Thomson Reuters, Springer Nature, Best Buy, Renault joined. In 2008, the LOD cloud had 45 datasets; in 2009: 95 datasets; in 2010: 203 datasets; in 2017: 1163 datasets; in 2020: 1255 datasets; in 2021: 1301 datasets. One noteworthy example of LOD is DBpedia that is publicly available as the Web of Data and comprises of structured information taken from the Wikipedia Infoboxes. The LOD derives its value from [7]:

- Flexible data integration: Enables the consolidation of previously disparate datasets into a single dataset that is complete, accurate, and up-to-date.
- The network effect: the addition of each new dataset increases the value of already published datasets.
- Better Navigation, Data Management: Because of URIs
- Compatible with existing standards and technologies.
- Better Data Quality: The use and (re)use of any kind of data triggers a growing demand to improve its quality. Errors are corrected through crowd-sourcing and self-service mechanisms.
- Data as a service: Better data usability because of the use of resolvable URIs. Data can be made available in different formats, e.g., XML, CSV, text, JSON, RDF.
- Ease of model updates: The Linked Open data models and vocabularies can be extended, adapted, and updated more easily.
- Cost reduction: Reuse leads to considerable cost reductions.

In 2010, Sir Tim Berner's Lee, the creator of the Linked Data suggested a 5-Star Criteria for the Linked Open Data (Fig. 4).

3.3.1 **LOD Applications**

Search, Exploration, and Q/A (Google Knowledge Graph, Amazon Product Graph, eBay ShopBot, Airbnb End-end Travel Platform); Personal Assistants (like Google

- Tim Berners-Lee's 5-Star Criteria**
- ★ Information is available on the web (whatever format) (under an open licence).
 - ★★ Information is available as machine-readable structured data (e.g., Excel instead of Image scan of a table).
 - ★★★ Non-proprietary formats are used (e.g., CSV instead of Excel; similarly, XML, RDF, JSON).
 - ★★★★ Use Open Standards (RDF and SPARQL) from W₃C. URI identification is used so that people can point at individual data.
 - ★★★★★ Data is linked to other data to provide context.

Fig. 4 Tim Berners-Lee's 5-Star Criteria

Assistant); Big Data Integration (IBM Watson); Movie Recommendations (Netflix); Education (Jungroo Learning); HealthCare; LinkedIn Economic Graph; Uber; Facebook; and many more.

4 Semantic Technologies as the Enabler

The artificial intelligence research started with the advent of expert systems, where the domain knowledge of the problem to be solved is formally modeled as facts and rules and then executed through inference mechanism [1]. The expert systems focused on domain-specific problem solving. Until late 1970s, machine learning had been used as a training program for AI; then around early 1980, there was a split between ML and AI and then AI research moved on using logical, knowledge-based approaches rather than algorithms. In the last two decades, research in artificial intelligence field has been boosted by Semantic Web standards and technologies. The semantic technologies are to be utilized as the kernel technologies to overcome the challenges [5].

4.1 Semantic Web Vision (Web of Machine-Readable Data)

Tim Berners-Lee, James Hendler, and Ora Lassila in their seminal paper “The Semantic Web” published in Scientific American told about a new form of web content that unleashed a revolution of new possibilities. The Semantic Web vision is about extending the existing World Wide Web by adding machine-interpretable metadata to the otherwise existing web content. This resulted in computers making meaningful inter-

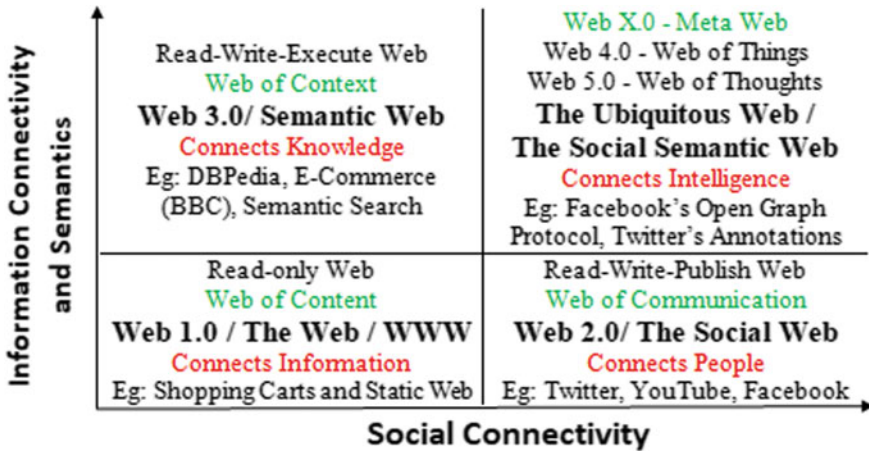


Fig. 5 Semantic Web vision

pretations. What we used to call keywords could now indicate terms whose semantics were defined for the IA through the Semantic Web. Now the IA that is roaming from one page to the other performing various actions will “know” the explicit as well as the implicit terms. Figure 5 shows the advancement of social connectivity and also information connectivity and semantics in the worldwide web through four compartments.

4.2 Semantic Technologies

The concepts, tools, and technologies of the semantic web have been rapidly adopted in data and information management. With the adoption of the Semantic Web vision, the development of a set of standards has been raised. These standards have been established by the international standards body—the World Wide Web Consortium (W3C). In Fig. 6, you can see a time span from 1988 to 2014 for the development of various technologies. Starting from simple web technologies (Unicode, HTTP, HTML) as the foundation technologies. Two important technologies for developing the Semantic Web were already in place: eXtensible Markup Language (XML) and the Resource Description Framework (RDF) [3]. RDF expresses meaning by encoding it in sets of triples, each triple like the subject, predicate, and object. Every term (subject/predicate/object) is identified by a Universal Resource Identifier (URI) enabling anyone to define/identify a term just by defining a URI for it somewhere on the Web. The URIs ensure that all the terms/concepts/keywords used in a document are not mere words/phrases but are tied to unique definitions findable on the Web.

	1988	1990	1996	1997	1998	1999	2000	2001	2004	2005	2006	2008	2009	2010	2013	2014
Foundation Technologies	Unicode	HTTP HTML			URI					IRI						
Serializations (Standard Syntax)			XML				XHTML JSON		RDFa	Microformats				JSON-LD		Turtle
Data Model (Metadata)				RDF								NOSQL revolution				
Ontologies + Inference (Languages)						RDFS	RML	OWL DAML+OIL	SWRL		OWL 1.1	SPARQL 1.0	OWL 2	RIF	SPARQL 1.1	

Fig. 6 Semantic technologies

5 Discussion and Insights

5.1 Artificial Intelligence Systems

There are three types of Artificial Intelligence Systems (Fig. 7):

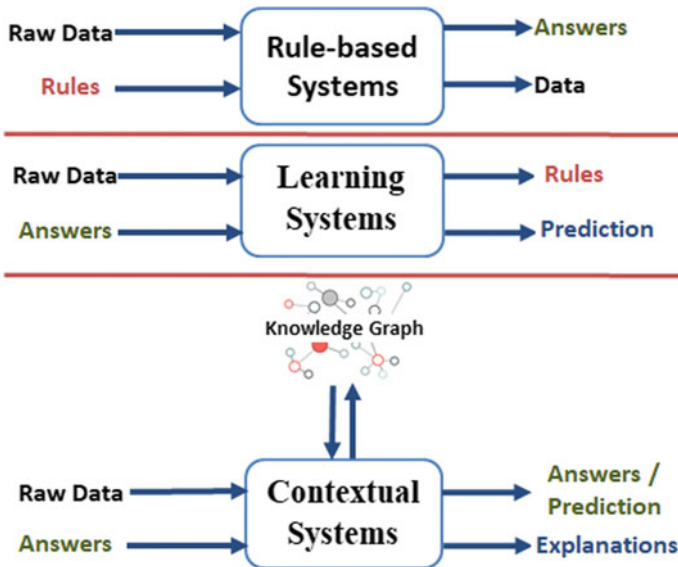


Fig. 7 Artificial intelligence systems

5.1.1 Rule Based Systems

are able to identify and adapt their own set of rules. The defining characteristic of these systems are their ability to identify and utilize a set of relational rules that collectively represent the knowledge captured by the system. It does this by utilizing its learning algorithm and relying on a knowledge base, meaning it does not require human coding or intervention for defining 'if-then' statements. Though exceptionally beneficial, rule-based systems have certain drawbacks associated with them, such as

1. They require deep domain knowledge and manual work.
2. Generating rules for a complex system is quite challenging and time-consuming.
3. It has less learning capacity, as it generates results based on the rules.

Such systems have pre-defined outcomes based on facts and rules (knowledge provided by experts) which cannot learn by themselves. These systems are called as Gold Old-fashioned AI (GOFAI) or Classical AI or Expert System or Symbolic AI.

5.1.2 Learning Systems

have a very ambitious goal in contrast to rule-based systems. Learning system is in principle unlimited in its ability to simulate intelligence. It's said to have adaptive intelligence. The ability to learn causes adaptive intelligence, and adaptive intelligence means that existing knowledge can be changed or discarded, and new knowledge can be acquired. Hence, these systems build the rules on the fly. That is what makes learning systems so different from rule-based testing. A neural network is an instance of a learning system. These systems includes learning techniques from Naive Bayes to n = Neural Networks. Although these systems are adaptive and rapidly improving but they are individually unreliable, requires quite a lot of data to learn and weak in abstract and reasoning. Such systems include Statistical AI/Machine Learning/Representation Learning/Deep Learning Techniques.

5.1.3 Contextual Systems

Symbolic AI and sub-symbolic AI have to go together, so the symbolist approach is nowadays manifested as a knowledge graph that advanced statistics and machine learning can run on top of. Here are two scenarios of the coupling of two approaches.

- If there is some noisy data, we need to gather it using sensors and process it through an Artificial Neural Network to infer the information. This information is then fed to a symbolic algorithm to recommend a possible action.
- Misclassification caused due to some statistical bias or noisy sensor readings should be overlaid with a symbolic constraint system to ensure the enforcement of what is logically obvious.

The power that human beings possess comes from both their ways of learning, i.e., accessing stored knowledge as well as figuring out new and unique problems. The machines get the same power by the combination of the rule based and the learning systems [6]. The rule-based systems follow a top-down model. The knowledge and learning are fed into such a system from the top. The values that such a system acquire during its upbringing through and through generations are the most powerful and effective. The theories and the knowledge base/store developed through generations are thereby passed to the scientists and engineers of successive generations. On the other hand, the learning systems follow a bottom-up model. These systems are provided with lot of training data, that is perceived and patterns are discovered in it.

5.2 Trends of Semantic Intelligence

Semantic Intelligence borrows from many diverse disciplines and the perspective can be traced back to a 2001 Scientific American article [2]. The readers interested in more in-depth reading could pursue the major publication venues in the field.

- since 1987: EKAW International Conference on Knowledge Engineering and Knowledge Management
- since 1995: AAAI Conference on Artificial Intelligence
- since 2001: ISWC International Semantic Web Conference
- since 2004: ESWC European Semantic Web Conference
- since 2006: Ontology Summit
- since 2007: ICSC International Conference on Semantic Computing
- since 2011: JIST: Joint International Semantic Technology Conference
- since 2013: SEMANTiCS
- since 2015: JOWO Joint Ontology Workshops
- since 2021: ISIC International Semantic Intelligence Conference

Here are the major trends of Semantic Intelligence in the above listed publication venues:

- 2003: Semantic Processing using Metadata, Basic Semantic Technologies (DAML+OIL, Ontology)
- 2004–2006: Many Semantic Web applications were developed. Simple and Easy to use Tags (FOAF), wiki, Collaborative development of ontologies, and Web 2.0
- 2007–2010: Linking between data instances, DBPedia, Linked Data, DBPedia was first presented at WWW-2007; First Special Session on Linked Data at ESWC-2008; In 2009 and 2010, ISWC and ESWC majorly discussed Linked Data.
- 2011–2020: Knowledge Graphs; Hybrid Knowledge Engineering, Explainable AI

6 Summary and Conclusions

It was around 1980s when machine learning branched out from artificial intelligence. Since then it is surviving on its own and also evolved into deep learning and neural networks. Now its time for the hybrid model to take off with trends like Hybrid Knowledge Engineering, AI applications in Healthcare, make NLP Apps Smarter, Intelligent Question Answering, From Simple Q/A to Actual Conversation, Fluently processing domain-specific Natural Language, and AI Digital Personas.

References

1. de Azevedo Jacyntho MD, Morais MD (2021) Ontology-based decision-making. In: Web semantics. Elsevier, pp 195–209
2. Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. In: Scientific American 284(5):34–43
3. De Wilde M (2015) From information extraction to knowledge discovery: semantic enrichment of multilingual content with linked open data
4. Fauscette M (2017) What are intelligent applications and real world use cases. <https://learn.g2.com/intelligent-applications-software>
5. Jain S, Murugesan S (2021) Smart connected world: a broader perspective. In: Smart connected world. Springer, pp 3–23
6. SOGETI LABS (2017) Semantic Technologies: The Next Step of AI. <https://labs.sogeti.com/semantic-technologies-next-step-ai/>
7. Linked Open Data Principles Technologies and Examples. <https://www.slideshare.net/OpenDataSupport/linked-open-data-principles-technologies-and-examples>

Sentiment Analysis of Public Health Concerns of Tokyo 2020 Olympics Using LSTM



Ayodeji Olalekan Salau , Temiloluwa Oluwatomisin Omojola, and Wasii Adeyemi Oke 

Abstract In recent years, in this age of social media, the Tokyo Olympic Games 2020 has received massive online sentiments and feedback from individuals, despite the Tokyo Olympic Games 2020's motto of "united by emotions." In this paper, a Recurrent Neural Network model was trained and implemented for the Tokyo 2020 Olympics public health sentiment analysis. A total of ten thousand (10,000) tweets were collected and trained, with tweets classified as positive, negative, or neutral. The RNN model implemented was the LSTM model. The model's performance was evaluated using performance metrics such as accuracy, precision, recall, and F1-score. According to the results, the LSTM model performed effectively, with a final validation accuracy of 88.2% and precision values of 0.86, 0.90, and 0.89 for negative, neutral, and positive predictions, respectively. The results of the LSTM model's high accuracy indicate that the model successfully determined the public health sentiments of the Tokyo 2020 Olympics, which were found to be positive at 40.39%.

Keywords LSTM · Sentiment analysis · Tokyo 2020 Olympics · Public health

1 Introduction

The Tokyo Olympic Games were held from July 23 to August 8, 2021, a year later than planned due to the COVID-19 pandemic caused by SARS-CoV-2. The fifth wave of COVID-19, caused by the new Delta viral variant, began to spread just before the Games. Before and during the event, there was some debate about the Olympic Games' potential contribution to the spread of the epidemic, and editorials were published about the Games' safety [1]. The Tokyo 2020 Olympic Games

A. O. Salau (✉) · T. O. Omojola
Department of Electrical/Electronics and Computer Engineering, Afe Babalola University,
Ado-Ekiti, Nigeria
e-mail: ayodejisalau98@gmail.com; ayodejisalau@abuad.edu.ng

W. A. Oke
Department of Mechanical and Mechatronics Engineering, Afe Babalola University, Ado-Ekiti,
Nigeria

have generated a lot of emotions in recent times, necessitating the need for experts to analyze the effects of such sentiments [2]. Sentiment analysis is a type of text mining that identifies and extracts subjective and objective information from datasets [3]. As a result, researchers can gain a better understanding of the social sentiment surrounding their chosen topic of interest. Sentiment analysis is a tool that classifies online discussions as positive, negative, or neutral using machine learning and Natural Language Processing.

Nowadays, people use social media to share their thoughts and opinions on almost any subject. This makes it simple to collect information from the general public. In Nigeria, the number of social media users increased by 22.2%, compared to a global average increase of 13%. As a result, we have approximately 33 million social media users, with approximately 3.05 million using Twitter.

The summer Olympic Games are the most watched sporting event in history, attracting billions of viewers [4]. The event is one of the most prestigious, and any country would be honored to host the games, as it provides numerous benefits to the participants, participating countries, and the host country. The Olympic Games bring together world-class athletes to compete on an international stage, while also providing host countries with an opportunity to showcase their countries to the rest of the world. The games help to boost tourism in the country. The implication of increased tourism rate is an improved economy due to enhanced trade. It increases exports and allows the host country to expand its trade, which helps to improve the economy. Trade has been found to increase by 30% in countries that have hosted the Olympics, indicating improved trade liberalization and improved trade flows. It accelerates the development of international-standard infrastructure in the host city, such as roads, hotel accommodations, swimming pools, communication facilities, sports stadiums, and so on. The host country also earns millions in ticket sales for various events [5–11].

While the Olympic Games have all these benefits, it also has negative impacts. The bid alone to host an Olympic Game cost millions of dollars and is growing even more expensive; although, the cost varies depending on the host city. Research has shown the cost of hosting is more expensive than the revenue it generates. It leads to environmental damages such as soil damages, erosion and abandoned infrastructure. The games are sometimes referred to as a losers game and somewhat redundant.

The remainder of the paper is structured as follows. Section 2 provides an overview of all related works. Section 3 discusses the Methodology used, including the LSTM model used and the process of training and testing the model. Section 4 presents the findings as well as the conclusions drawn from them. Section 5 concludes the paper.

2 Related Works

In [5], the sentiment analysis of Tokyo 2020 was investigated using Natural Language Processing and discrete mathematical methods to examine several tweets in order to obtain statistical results about the public opinion regarding the event. Python

packages for Natural Language Processing (NLP) were used to determine the polarity and subjectivity of each tweet. Polarity was used to extract positive, negative, or neutral emotions of each tweet from the text data.

The authors make use of API to obtain data (tweets) from Twitter. The authors used a Recurrent Neural Network (RNN) to provide the sentiment analysis on the Tokyo 2020 Olympics. Sequential data is one of the most difficult types of data to work with because the data cannot be assumed to be independent. RNN is helpful in this situation because it uses back-propagation through time to learn rather than a feed-forward pass and instead follows a recurrence relation. The input of the current step is obtained from the output of the previous step. RNN is exceptionally good at modelling units in sequence; which architecture performs better depends on how important it is to semantically understand the whole sequence [3]. This classification algorithm is useful and efficient because it can provide a good testing accuracy.

In [6], the authors obtained data (tweets) from Twitter API between August 6 to August 21 2016 for the Olympics and September 7 to September 17 2016 for the Paralympics. English and Farsi tweets were collected in real time. The data was cleaned (removing extra items like URLs, hashtags, emojis, etc.) and a pre-processing module was applied to extract tweets of the Iranian athletes. A sentimental classification was created using Newbies polynomial, and the classification of tweets was compared using SVM and CRF. The authors compared their method with another method previously used and their results were more accurate (72.67% average precision) and achieved an improvement of 26.62% from the other methods. Also, the average recall criterion of 2.42% was improved by 1.81%.

In [7], the authors show the difference in tone towards the Tokyo 2020 Olympics among major Japanese newspaper articles. The newspapers analyzed are Asahi Shimbun, Mainichi Shimbun, and Yomiuri Shimbun, which are the three top-selling newspapers in Japan. They detect positive and negative tones in the newspaper articles that refer to the Olympics during the period of January 1 to August 9, 2021. The study utilized the list of semantic orientations of words to assign a sentiment score to each article. They calculated the sentiments score by categorizing sentences in each article into each word class utilizing MeCab, an open-source software for morpheme analysis and by calculating the sentiments score of each article using the list of semantic orientations of words. The findings reveal that while conservative newspaper Yomiuri expressed support for the Olympic Games, liberal newspapers Asahi and Mainichi, which are thought to be opposed to holding the Olympics, do not consistently express this support. To the best of the author's knowledge, no research has been done to date on how the general public felt about the Tokyo 2020 Olympics after the event.

3 Methodology

This section presents the method used which includes data acquisition, data pre-processing, model development, and classification of tweets for the sentiment analysis.

3.1 *Extraction of Data*

This is the first step in acquiring the datasets. To gain access to the Tweepy API, a Twitter developer account was created. This enables the ownership of access tokens in the form of OAuth settings such as Consumer Key, Consumer Secret, OAuth Access Token, and OAuth Access Token Secret. These tokens are stored in a config file for safekeeping as anyone can have access to projects being developed if their tokens are exposed. Necessary libraries such as tweepy and pandas were imported on the Jupyter notebook and the OAuth tokens are declared in that same Jupyter notebook. Authorization of the user's identity and tokens is done with tweepy's OAuthHandler after the tokens have been defined.

3.2 *Data Pre-Processing*

Data pre-processing was performed on the data after using the Textblob library to ensure good accuracy for training and testing the model. To perform this task, unwanted characters like alpha-numeric characters and punctuations were identified and removed. All texts including "@" were replaced with empty parentheses, and also all irregular characters were replaced with empty spaces.

During processing, Spacy is first used to tokenize the text, i.e. segment the text into words and punctuation. This was done by applying rules specific to each language. First, the raw text is split (" ") and then, the tokenizer processes the text from left to right. The tokenized sentences usually have different lengths, as such padding is required to make the sentences have equal lengths.

The data is labelled into three classes and a python library tool, Textblob sentiment analyzer which employs NLTK is used to label the tweets. It produces an output which indicates the polarity and subjectivity of the data. Usually, the polarity score lies between the ranges of values -1 to 1 with the negative value (-1) being the most negative word in the dataset and the positive value (1) being the most positive word. The subjectivity score lies between the values of 0 and 1 . Subjectivity shows the personal sentiment in the data, i.e. the text/data will have less of facts and more of personal feelings, opinions or sentiments. Textblob makes use of the polarity generated to label the data as negative, neutral or positive. Table 1 shows the polarity and subjectivity of data used.

Table 1 Polarity and subjectivity of data

0	Polarity	Subjectivity	Sentiment
1752	0.22	0.53	Positive
9743	0.4	0.73	Positive
8232	0.2	0.59	Positive
6847	0	0	Negative
6228	-0.55	0.6	Positive

3.3 Model Development

The model was developed using Keras LSTM, an RNN architecture that primarily solves the memory problem which RNN faces. This model contains 2 hidden layers and a dropout of 0.3. By employing this model, the results from previous cell states are easily recalled to be added to a new input data sequence in a new hidden state. The data needs to be converted to numerical values (vectors) for the machine learning method to perform training and testing; therefore, an embedding layer was introduced into the mode. The embedding layer converts all the input from the padded sequence to a defined size of fixed length vector. The resultant vector obtained is dense containing real values as opposed to having only 0's and 1's. The fixed length of vectors helps us to achieve a better representation of the data with reduced dimensions.

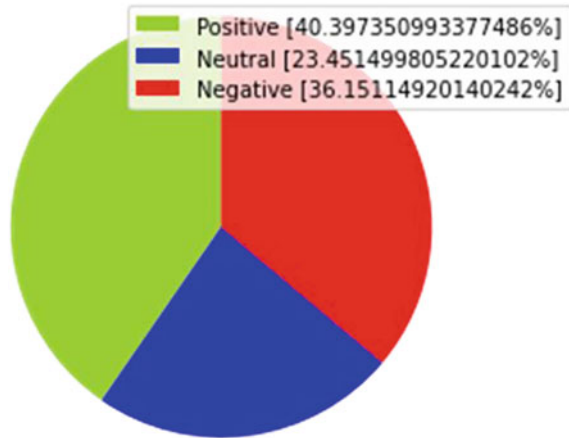
3.4 Model Training and Testing

The LSTM model was trained with 2 hidden layers. The Categorical Cross Entropy loss function which shows the probability between two probability distributions and Adam optimizer is the optimizer function which was implemented. It replaces the optimization algorithm for a stochastic gradient descent for training deep learning. The dataset was divided into the train and test set using the Scikit Learn Python Library's 'train_test_split' function with a ratio of 80% and 20% respectively. The model is trained using an epoch of 70 with a batch size of 64, i.e. the number of data being fed per epoch.

4 Results and Discussion

This section presents the experimental result obtained for the adopted neural network model, which was based on: Pie chart, accuracy, classification report and confusion matrix.

Fig. 1 Pie Chart of Sentiment Analysis based on classification



4.1 Evaluation Based on Model Classification

Figure 1 shows that the public health sentiment was slightly higher on the positive side with a value of 40.39% than negative. According to the data gathered from Twitter, the general public did not perceive the Tokyo 2020 Olympics to have entirely negative effects on the health of the inhabitants of the host city, the athletes, coaches, Olympic workers and officials. Based on the close range of values between the positive and negative classification in Fig. 1, it can be said that the Tokyo 2020 Olympics has some negative effects achieving a negative value of 36.15%.

4.2 Evaluation Based on Model Accuracy

Contrary to the training accuracy, which began at a value of approximately 0.63, the validation accuracy began at a value of approximately 0.78 in the first epoch and gradually dampened to a steady value, as shown in Fig. 2. A final train accuracy of 99% and a validation accuracy of 88.2% were attained after 70 epochs, despite fluctuations in accuracy throughout the training period.

4.3 Evaluation Based on Model Loss

Figure 3 shows that the model's loss also decreases significantly between the first and second epochs as a result of the model's accuracy in Fig. 2 increasing. The model achieved a train and validation loss of 0.0053 and 0.9899, respectively, after 70 epochs.

Fig. 2 LSTM Model Accuracy

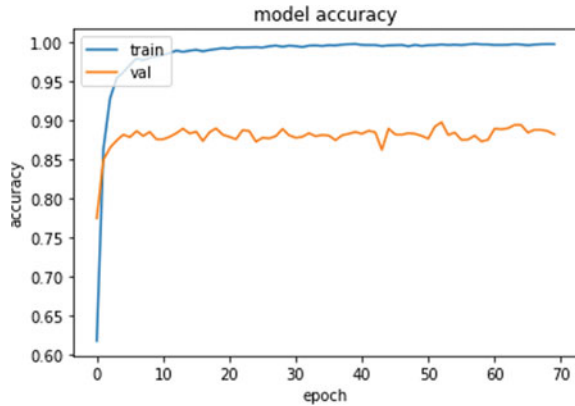
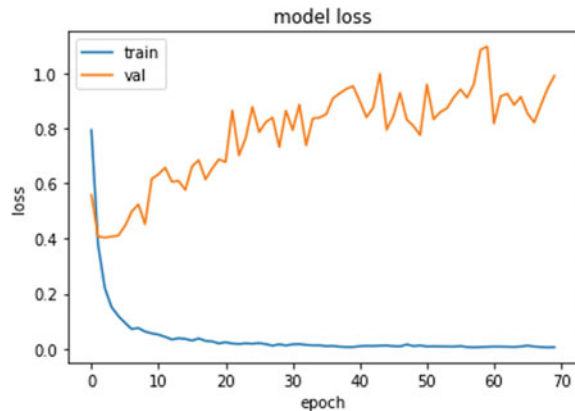


Fig. 3 LSTM Model Loss



4.4 Evaluation Based on Confusion Matrix

From the confusion matrix in Fig. 4, the number of true and false positives can be deduced from both the heat maps as well as the numbers on the plots.

Negative: There were 153 false negatives from a total of roughly 1058 test samples, which were predicted to be neutral and positive. However, due to the insignificance of the number of false negatives as compared to the number of true negatives, the results achieved for the precision, recall, and f1-score are as follows 0.86, 0.96, and 0.90 respectively as presented in the classification report in Table 2.

Neutral: There were 54 false neutrals out of a total of about 518 test samples for neutral, which were predicted to be both negative and positive. However, due to the insignificance of the number of false neutrals as compared to the number of true neutrals, hence, as a result of approximation, the precision, recall and f1-score are 0.90, 0.75, and 0.81 respectively.

Fig. 4 LSTM Model Confusion Matrix

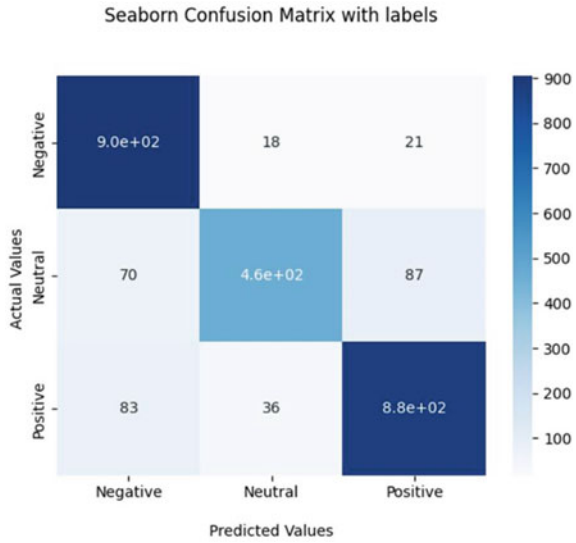


Table 2 Classification report for LSTM model

Labels	Precision	Recall	F1-Score
Negative	0.86	0.90	0.90
Neutral	0.90	0.75	0.81
Positive	0.89	0.88	0.89

Positive: There were 108 false positives out of a total of 991 positive test samples, which were predicted to be neutral and negative. However, due to the insignificance of the number of false positives as compared to the number of true positives, the results of the precision, recall, and f1-score are 0.89, 0.88, and 0.89.

The LSTM model generates appropriate precision, recall, and f1-score for each of the classes based on Table 2. The confusion matrix in Fig. 4 is used to generate the model’s classification results.

To obtain the values for precision, recall, and F1-Score, Eqs. (1–3) are used;

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positives} \tag{1}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

5 Conclusion and Recommendation

In this study, a model using a Recurrent Neural Network (RNN) architecture; a Long-Short Term Memory approach was developed for determining public health sentiment of the Tokyo 2020 Olympic games. After implementing and evaluating the model to obtain the sentiment of people, the obtained results indicate that the LSTM model performed satisfactorily obtaining a validation accuracy of 88.2% after 70 epochs. In addition to the model's accuracy, the confusion matrix clearly demonstrates that the model has good precision and recall values across all classes. As a result, the model clearly demonstrates that it is capable of producing accurate and effective results for sentiment analysis.

References

1. Ribas RM, de Campos PA, de Brito CS, Cavalcanti Dantas CS (2020) 2021 Olympic games Tokyo: safety issues and protection against COVID-19 transmission, *Journal Glob Insect Dis.*, vol. 12, pp 114–115, https://doi.org/10.4103/jgid.jgid_88_20
2. Ilevbare SI, McPherson G, (2022) Understanding COVID-19: A hybrid threat and its impact on sport mega-events. A focus on Japan and the Tokyo 2020 Olympic Games, *Front. Sports Act. Living*, <https://doi.org/10.3389/fspor.2022.720591>
3. Abeje BT, Salau AO, Ehabu HA, Ayalew AM, (2022) Comparative analysis of deep learning models for aspect level amharic news sentiment analysis, In: 2022 International conference on decision aid sciences and applications (DASA), pp 1628–1633, <https://doi.org/10.1109/DASA54658.2022.9765172>
4. Vertalka JJ, Kassens-Noor E, Wilson M (2019) Data on sentiments and emotions of olympic-themed tweets. *Data Brief* 24:103869. <https://doi.org/10.1016/j.dib.2019.103869>
5. Montiel EC (2021 November) Sentiment analysis of tweets about the Tokyo 2020 Olympic games
6. Srivastava R, Upes F (2020) Sentimental analysis of olympics tweets, *Annals of R.S.C.B.*, 24(2), pp 427–435
7. Yin W, Kann K, Yu M, Schütze H (2017) Comparative study of CNN and RNN for natural language processing <http://arxiv.org/abs/1702.01923>
8. Seilsepour A, Ravanmehr R, Sima HR (2019) 2016 Olympic games on twitter: Sentiment, 5(3)
9. Ji X, Chun SA, Wei Z, Geller J (2015) Twitter sentiment classification for measuring public health concerns. *Soc Netw Anal Min*, 5(1) <https://doi.org/10.1007/s13278-015-0253-5>
10. Ji X, Chun SA, Geller J (2013) Monitoring public health concerns using twitter sentiment classifications. *IEEE Int Conf Healthc Inform* 2013:335–344. <https://doi.org/10.1109/ICHI.2013.47>
11. K. Jia, Y. Zhu, Y. Zhang, F. Liu, J. Qi, International Public Opinion Analysis of Four Olympic Games: From 2008 to 2022, *J Saf Sci Resil*. <https://doi.org/10.1016/j.jnlssr.2022.03.002>

Internet of Nano and Bio-Nano Things: A Review



Şeyda Şentürk , İbrahim Kök , and Fatmana Şentürk 

Abstract In recent years, advances in biotechnology, nanotechnology and materials science have led to the development of revolutionizing applications in Internet of Things (IoT). In particular, the interconnection of nanomaterials, nanoimplants, and nanobiosensors with existing IoT networks has inspired the concepts of Internet of Nano Things (IoNT), and Internet of Bio-Nano Things (IoBNT). To date, although there are several survey papers that addressed these concepts separately, there is no current survey covering all studies in IoNT and IoBNT. Therefore, in this paper, we provide the complete overview of the recent studies in these three areas. Furthermore, we emphasize the research challenges, potential applications, and open research areas.

1 Introduction

The number of Internet usage and Internet-based applications in our daily life is increasing day by day [1]. This increase also brings about an exponential increase in number of devices with internet access. In this context, it is estimated that the number of connected devices to the Internet will exceed 75 billion by 2025 [2]. As a natural consequence of these developments, the concept of the Internet of Things (IoT) has become the focus of research and development, especially in the last 15 years [3]. IoT concept represents the connection and communication of all kinds of physical things such as sensors, actuators, personal electronic devices in the real world via the Internet [4]. Today, IoT applications appear in many areas such as smart transportation, real-time monitoring systems, smart cities, smart grids, smart environmental monitoring,

Ş. Şentürk (✉)

Department of Food Engineering, Pamukkale University, Denizli, Turkey

e-mail: seydasenturk@gmail.com

İ. Kök · F. Şentürk

Department of Computer Engineering, Pamukkale University, Denizli, Turkey

e-mail: ikok@pau.edu.tr

F. Şentürk

e-mail: fatmanas@pau.edu.tr

medical and health systems, and smart buildings [5]. On the other hand, thanks to the advancements in materials science and nanotechnology, IoT application areas are expanding further and beating a path for the formation of new concepts like the Internet of Nano Things (IoNT) and the Internet of Bio-Nano Things (IoBNT).

Although these concepts are just emerging today, the history of nanotechnology, which forms the basis of these concepts, goes back to the 1950s. In 1959, Richard Feynman first discussed the possibility of straightly handling materials at the atomic scale and the idea of reproducing everything by miniaturization on nanoscale in his famous talk entitled “There’s Plenty of Room at the Bottom” [11]. In 1974, the term “Nanotechnology” was first coined by Norio Taniguchi to describe dimensional accuracy [12]. Nanotechnology is the study and use of extremely small objects on an atomic and molecular scale. It has the potential to produce many new technological materials and devices for the benefit of people by bringing together many disciplines such as engineering, medicine, biology, physics, and chemistry [6]. Nano things are devices ranging from 1 to 100 nanometers (nm) and can perform tasks such as collecting, creating, computing, processing, and transmitting data at nanoscale [13, 14]. Today, it has revealed promising networking paradigms such as IoNT and IoBNT, which are formed by interconnecting nano-things with existing communication networks via high-speed Internet. These network paradigms have high application potential in different fields such as healthcare, biomedical, military, smart environment, industry, energy, and multimedia [4, 7]. Therefore, there is a need for up-to-date studies that comprehensively explain all these network paradigms. As shown in Table 1, although there are several studies in the literature that individually address IoT, IoNT, and IoBNT, there is currently no review article that encompasses all three subjects together. This suggests a need for a comprehensive examination of these related technologies in a single publication. Based on this motivation, in this paper, we present a survey that fills the existing gaps in the literature. The following is a summary of the survey’s main contributions:

- We present a review paper on IoNT and IoBNT network paradigms, which may have many potential applications in the future.
- We then detail the potential application areas, architectural models, and communication structures of all presented network paradigms.
- We also discuss several IoNT and IoBNT problems, outstanding topics, and future research possibilities.

The remainder of this paper is organized as follows. In Sect. 2, we provide an overview of IoNT, IoBNT, and their applications. In Sect. 3, We also discuss several IoNT and IoBNT problems, outstanding topics, and future research possibilities. Finally, Sect. 4 concludes the paper.

Table 1 Existing survey papers and relation to this paper

Paper	IoT	IoNT	IoBNT	Description
[6]	×	✓	×	This study discusses the usage scenarios, network architecture, communication paths, pros and cons of IoNT and nano-sensors in modern healthcare
[7]	×	✓	×	This paper presents the features, network architecture, application possibilities, key challenges, and future trends of IoNT
[8]	×	✓	×	This paper briefly summarizes the network models of IoNT and the difficulties encountered in its implementation in healthcare
[9]	✓	×	×	Solid waste management was controlled and cost analysis of this system was carried out by using IoT
[10]	×	×	✓	This paper presents the key components, applications, technological challenges, and future research directions of IoBNT
Our work	✓	✓	✓	Our paper covers all nano-network paradigms and potential application areas, architectural models, and communication structures. It also comprehensively addresses current challenges and future research directions

2 Internet of Nano Thing and Internet of Bio-Nano Things

The interconnection of nano-scale devices and machines via the Internet has paved the way for the emergence of new network architectures such as IoNT and IoBNT. These network architectures are based on existing technologies such as IoT, sensor networks, edge, fog, cloud computing and newly developed nano-scale sensor, machine, and smart antenna technologies [15]. In this context, the concept of IoNT was described as “*The interconnection of nanoscale devices with existing communication networks and ultimately the Internet defines a new networking paradigm that is further referred to as the Internet of Nano Things*” in 2010 by Ian F. Akyildiz and Josep Miguel Jornet from the Georgia Institute of Technology [16]. On the other hand, IoBNT is a developed version from IoNT and focuses on information exchange, interaction, and networking within the biochemical field using synthetic biology and nanotechnology tools [4]. The main purpose of IoBNT is to communicate with cells that enable real-time and accurate detection and control of complex biological dynamics occurring in the human body [17]. In other words, while IoNT is based on the integration of nanoscale devices with existing network and communication technologies, IoBNT is additionally based on the behavior and properties of in vitro environments such as molecular communication, Ca²⁺ com-

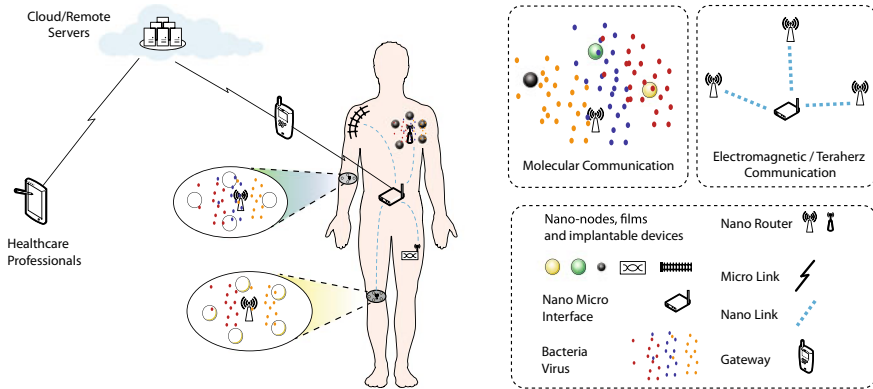


Fig. 1 A typical nano-network architecture for healthcare applications

munication, hormonal communication [18]. The nanodevices used in these network architectures can be created from materials such as biological materials, magnetic fragments, or gold nanoparticles. Here, biological devices are created by reprogramming many biological materials such as cells, viruses, bacteria, bacterial phages, erythrocytes [19].

In recent years, scientific and technological developments in biomaterial fields have enabled the development of smart biomaterials with high bio functions. The smartness level of these biomaterials is divided into four classes as inert, active, responsive, and autonomous. Showing their biofunctionality, these smart materials require a wide range of internal and exterior stimuli. Examples of internal stimuli are proteins, enzymes, molecules, antigens, and ionic factors. Examples of external factors are magnetic field, electric fields, light, temperature, and mechanical stress [20].

In the light of the information given above, nano and bio-nano device-based network architectures have many potential application areas such as health, military, environmental monitoring, multimedia, and entertainment. Therefore, we first present the network architectures and communication models of IoNT and IoBNT in Sect. 2.1. We then provide comprehensive examples of application areas in Sect. 2.2 (Fig. 1).

2.1 Network Architecture and Communication in IoNT and IoBNT

Effective integration and communication between nanodevices and macroscale components are required for the IoNT and IoBNT network paradigms to be fully operational. For this reason, the design of network architectures that include several alternative communication paradigms, such as electromagnetic, acoustic, mechani-

Table 2 Features of communication modes at the nanoscale

Approach	Biocompatibility	Range in body	Energy
Nano wireless	Medium to low	1 cm to 1 m	High
Nano acoustic	Low	1 cm to 10 cm	High
Molecular	High to medium	1 nm to 1 cm	Very low
Nanomechanic	Medium to low	1 mm to 1 nm	Very low
Bacteria-based	High to medium	1 mm to 1 cm	Very low

cal, and molecular communication, comes to the fore [21]. Within these networks, nano-things are expected to interact with each other by exchanging various types of information such as synchronization signals, sensed chemical/physical parameter values, logical operation results, instruction sets, and commands [4].

For the time being, communication in IoNT and IoBNT networks is mainly envisaged as molecular communication and nano-electromagnetic communication [22]. Since these two network paradigms are intertwined, similar technologies can be used in terms of communication. Molecular communication is created by releasing and reacting to certain molecules to transmit or receive information. On the other hand, the transmission and receiving of electromagnetic radio-frequency waves in the terahertz (THz) range constitute nano-electromagnetic communication [23]. However, there are difficulties in using these communication technologies such as coverage, compatibility, energy, and transferred data rate. Although there are strong theoretical knowledge and practical models in the literature on communication today, communication at the nanoscale is quite complex and difficult due to the biocompatibility problems of nanodevices and the signal propagation properties of tissues in the body [21] (Table 2).

IoNT and IoBNT network architectures consist of a series of components that connect the electrical field and the biochemical field with devices at the nano or bio-nano-scale. More specifically, the main components of the IoNT architecture are nano-nodes, nano-routers, nano-micro interface devices, and gateways [13]. IoBNT, on the other hand, may contain additional components that have the function of reading and transferring biochemical domain information. In this context, IoBNT architecture can consist of Bio-cyber interface, gateway, and application-specific server components [19]. All these components are described in Table 3.

2.2 *IoNT and IoBNT Applications*

IoNT and IoBNT network paradigms have added new dimensions to the existing application areas of IoT such as healthcare and communication networking. Therefore, in this section, we present current studies on these two application areas.

Table 3 The components of the IoNT and IoBNT network architectures

Components	Description
Nano-nodes	Data sensing, transmission, and processing are all performed by nano-nodes, which are the simplest nanodevices
Nano-routers	Nano-routers are more advanced devices than nano-nodes in terms of features such as computing and storage. They are responsible for collecting information from nano-nodes and controlling nano-nodes with basic control commands such as on/off, read, and sleep
Nano-micro interface device	Nano-micro interfaces are hybrid devices that can both communicate at the nanoscale and use existing communication paradigms in traditional communication networks. They are in charge of gathering data from nano-routers and transmitting it to micro-scale devices. This component is used in IoNT
Bio-cyber interface	The bio-cyber interface is a hybrid device that converts the biochemical signal received from in-body nano-networks into an electrical signal that can be processed in external networks
Gateways	Gateways are hybrid devices that can be used in both classical and nano-networks at micro and macro scales. These devices allow remote control of the designed IoNT and IoBNT networks via the Internet
Application-specific servers	These devices are responsible for the storage, analysis, and real-time monitoring of information from nano-networks. They can be used in the realization of many applications such as healthcare, medicine, entertainment, or multimedia

In healthcare, it has been shown that these network paradigms can be used in the treatment of circulatory system diseases, infectious diseases, diabetes, thrombosis, and many physical or psychological diseases. Today, due to the Covid-19 pandemic, remote work has become increasingly common in many areas. In this context, studies have been carried out regarding the electronic and remote provision of health services [19]. In this way, it has become possible to transmit data collected from patients connected to the IoNBT network directly and in real time to healthcare professionals. Thanks to the remote collection of patient data, the patient may not need to go to the laboratory for testing. In addition, in case of any infection, the disease can be detected even before the patient shows symptoms and medical support can be given to the patient [17]. Telemedicine applications related to these sample scenarios have been included in our lives with the Covid-19 epidemic. For example, Jarmakiewicz

et al. set forth standards and guidelines for the research and development of nano-sensor networks for some telemedicine applications. They also demonstrated that some applications can be implemented with a nano-sensor network by testing it with a nano-network developed for human circulatory system [24].

Thrombosis is one of the most common causes of mortality in the world, killing one out of every four individuals. In thrombosis, a blood clot forms in a vein, once the blood clot forms, it can stop or delay blood flow, or it can be found in organs. Froud et al. [3] developed a new IoBNT model with a bio-cyber communication interface that allows for better prediction and analysis of blood vessel coagulation. Thanks to this model, the information in the blood vessel is collected and the bio-cyber interface is used to convert the information into electrical equivalent. Moreover, the optical or thermal response has also been used to stimulate the release of certain nanocarrier molecules such as liposomes, nanodevices that can be transported through the bloodstream and predict clots.

IoBNT can also be used in the early diagnosis and mitigation of infectious diseases. For instance, cystic fibrosis disease is a genetic disease that can be seen in organs and systems such as lungs, pancreas, intestines, and sweat glands. It develops in waves and causes the death of patients. In [17], Akyildiz et al. proposed an IoBNT network called PANACEA, which provides an end-to-end solution to infectious diseases. In this network, a submillimeter implantable bio-electronic device that senses communication within body cells is used to determine the level of infection.

Another common disease today is diabetes. Abbasi et al. [25] focused on modeling IoBNT applications that will improve the diagnosis, management, and treatment practices of the insulin-glucose system for this disease. Through an IoBNT network to be developed, insulin and glucose concentration values can be transmitted directly to healthcare providers. In this way, besides monitoring the health status, pump life can be extended by ensuring that the insulin cartridge lasts longer. In addition, thanks to such smart systems, the body is protected from the side effects of excess insulin. On the other hand, preventive health services can be provided by methods such as pre-illness therapy with a holistic approach to physical and psychological diseases. In this context, it is predicted that healthy living conditions can be created by making changes in people's lifestyles with IoBNT [26].

Since IoNT and IoBNT are emerging network paradigms, the developed applications in these networks are also at the beginner level. Moreover, the integration of nanomachines into the human body, their architectural design, and efficient operation (communication, computing, storage, etc.) of these devices within the heterogeneous network structure are among the main challenges. Therefore, there is an increasing number of research efforts aimed at eliminating existing and potential challenges. Al-Turjman proposed an energy-efficient framework focusing on data delivery in nano-networks. With this framework, it is aimed to realize data distribution with energy-sensitive routing protocols by considering the shortest path [14].

It is clear that IoNT has different architectural requirements for different network models and applications. In these networks, communication model design is also another challenge. In [8], Ali et al. investigated the structure of communication models in the IoNT network that was developed for drug delivery and disease

detection. In the study, they evaluated the advantages and disadvantages of these two models by establishing non-additive and single-layer communication models. In [27], Stelzner et al. introduced a new concept of function-centered Nano-network (FCNN) focusing on intra-body communication scenarios. FCNN aims to minimize memory requirements by combining the location and function capabilities of nanomachines. In another study, Canovas-Carrasco et al. [28] emphasized that the development of optimal transmission policies to be used in nano-networks, reducing implementation costs and in vivo monitoring of nano-sensors depending on the generation of smart policies. For this reason, the authors proposed the Markov decision process model, which enables the derivation of smart policies.

From the network architecture perspective, Galal and Hesselbach [29] presented a multi-layered architectural model in nano networking that combines software-defined networking (SDN), network function virtualization (NFV), and IoT technologies. In the study, the authors proposed a number of functionalities and usage scenarios that could be implemented for nanodevices. In addition, significant challenges and gaps in implementing the proposed functions with nanotechnology are discussed. In [6], the usage possibilities, architecture, communication models, advantages, and challenges of nanotechnology through nano biosensors and IoNT in modern health care were determined. It has been emphasized that the concept of placing nanoscale devices in the human body has the potential to be used in all health and medical applications that exist today (Table 4).

3 Challenges, Open Issues, and Future Research Directions

IoNT and IoBNT networks have wide application potential to improve the health and quality of life of living things. However, many challenges are encountered in real-life applications of these networks. These challenges also present the topic of open research issues and future research directions.

- **Nano-network architecture:** Traditional IoT networks are usually built with three, five, and seven layers. In nano-networks, it is very difficult to establish a layered network structure in terms of device size, computation, storage capacity, and communication types.
- **Nano communication and Standards:** There are short and limited (from 1 nm to 1 m) in vivo communication modes in nano-networks. There is also a high level of biological noise in the environment. These communication difficulties can cause data loss, delay, and network congestion in nano communication. In addition, since these networks differ from conventional networks in many aspects such as operating environment and communication mode, the OSI/TCP model is not fully suitable for these networks. In this context, the foundation of new communication standards is quite necessary [6, 10].
- **Bio-cyber Interface:** As mentioned earlier, different bio-cyber interfaces have been proposed for each considered network. However, these interfaces are not yet at the

Table 4 Summary of some selected papers in the area of IoNT and IoBNT

Paper	Scope	Domain	Technology	Obtained/used datasets	Contribution of the work
[24]	IoNT	Healthcare	Nanosensor Radio channels	Simulation Data	The telemedicine applications in the literature were examined and the points that could be improved were determined. It has been stated that it would be beneficial to support these applications by using a nano-sensor network that has the potential to be applied in some scenarios and that works in the human circulatory system
[27]	IoNT	Biomedical (Human Body)	Sensor Actuator	None	They tried to minimize memory requirements by combining the data collected about the location and function of the respective nano-devices, using the Function-centered Nano-network (FCNN) architecture.
[28]	IoNT Nano Networks	Healthcare (Human Cardiovascular System)	Nanodevices Nano-routers	Cardiovascular system	A Markov-based decision process model is proposed to generate optimal transmission policies to be used by nano-nodes. In addition, a series of simulations were executed.
[29]	Nano Networks	Healthcare	Nanodevices Nano-antennas	Nano drugs Wearable devices	It is represented an approach which is combining SDN, NFV, and IoT. Also, a composite architecture model of nano-network communication is proposed.
[30]	IoNT	Human Body	RF radio channels	Simulation Data	The paper defines challenges and opportunities of IoNT systems. Also it is constructed simulation area for human body area network technology using nano-devices

practical and clinically desirable level. For these reasons generalized and realistic biointerface deployments are needed [10, 31].

- **Data Management and Analysis:** It is predicted that there will be a rapid increase in the number of new devices connected to the internet together with nano-networks. Currently, data method and analysis are a serious challenge. This challenge will become a hot topic with nano-networks. In addition, the development of smart data analysis algorithm models that can work on nano-scale devices is open to research [32].
- **Smart Biomaterials:** Although there are many types of biocompatible in-capsular sensors (camera, pH, temperature, gas, ultrasound imaging, physisorption, surface acoustic wave, etc.), the field of ingestible sensors and smart biomaterials is still in its absolute infancy [20, 33].

- **Security and Privacy:** The new nano-network paradigms mostly include applications that have not yet established standards and frameworks in the medical, biological, chemical, or personal fields. Therefore, these networks are now exposed to a wide range of threats, including multidimensional attack vectors [34]. At this point, ensuring the confidentiality, integrity, and availability of the data stored and circulating on the network is more critical than other application areas. The weakness that will arise in this regard may harm individuals and societies by laying the groundwork for negative situations such as data manipulation, theft, espionage and even bioterrorism [35]. Therefore, trust management systems, user access control systems, lightweight data encryption, and compression models are needed.
- **Other challenges:** Besides the current challenges, there are many uncovered research topics such as content management, mobility management, service aggregation and discovery, energy conservation, energy harvesting, power transfer, nanodevice addressing, integration with next-generation technologies (SDN, NFV, Blockchain, NFTs, Metaverse and etc.) [10, 32].

4 Conclusion

This paper provides comprehensive literature to contribute to a holistic understanding of emerging nano-networks. For this purpose, we present all of the promising network paradigms such as IoNT, and IoBNT that have emerged based on nanoscale devices and materials in connection with the “All-thing connected” vision. We give comprehensive explanations of communication, network architecture, and application domains of these network paradigms. Moreover, we comprehensively provide the challenges of these networks, open issues, and future research guidelines. In this work, we comprehensively present all emerging nano-network paradigms in a single survey paper, and we hope that the paper will shed light on future works.

Acknowledgements Seyda Şentürk is supported by the Council of Higher Education (CoHE) under the special 100/2000 scholarship program.

References

1. Miraz MH, Ali M, Excell PS, Picking R (2018) Internet of nano-things, things and everything: future growth trends. *Future Internet* 10(8):68
2. Nawaratne R, Alahakoon D, De Silva D, Chhetri P, Chilamkurti N (2018) Self-evolving intelligent algorithms for facilitating data interoperability in IOT environments. *Future Gener Comput Syst* 86:421–432
3. Fouad H, Hassanein AS, Soliman AM, Al-Feel H (2020) Analyzing patient health information based on IOT sensor with AI for improving patient assistance in the future direction. *Measurement* 159:107,757
4. Akyildiz IF, Pierobon M, Balasubramaniam S, Koucheryavy Y (2015) The internet of bio-nano things. *IEEE Commun Mag* 53(3):32–40

5. Kök İ, Özdemir S (2020) Deepmdp: a novel deep-learning-based missing data prediction protocol for IOT. *IEEE Internet Things J* 8(1):232–243
6. Pramanik PKD, Solanki A, Debnath A, Nayyar A, El-Sappagh S, Kwak KS (2020) Advancing modern healthcare with nanotechnology, nanobiosensors, and internet of nano things: taxonomies, applications, architecture, and challenges. *IEEE Access* 8:65230–65266
7. Cruz Alvarado MA, Bazán P (2019) Understanding the internet of nano things: overview, trends, and challenges. *E-Ciencias de la Información* 9(1):152–182
8. Ali NA, Aleyadeh W, AbuElkhair M (2016) Internet of nano-things network models and medical applications. In: 2016 international wireless communications and mobile computing conference (IWCMC). IEEE, pp 211–215
9. Velvizhi G, Shanthakumar S, Das B, Pugazhendhi A, Priya TS, Ashok B, Nanthagopal K, Vignesh R, Karthick C (2020) Biodegradable and non-biodegradable fraction of municipal solid waste for multifaceted applications through a closed loop integrated refinery platform: paving a path towards circular economy. *Sci Total Environ* 731:138,049
10. Kusu M, Unluturk BD (2021) Internet of bio-nano things: a review of applications, enabling technologies and key challenges. arXiv preprint [arXiv:2112.09249](https://arxiv.org/abs/2112.09249)
11. Feynman RP (1992) There's plenty of room at the bottom [data storage]. *J Microelectromech Syst* 1(1):60–66. <https://doi.org/10.1109/84.128057>
12. Hulla J, Sahu S, Hayes A (2015) Nanotechnology: history and future. *Hum Exp Toxicol* 34(12):1318–1321
13. Nayyar A, Puri V, Le DN (2017) Internet of nano things (iont): next evolutionary step in nanotechnology. *Nanosci Nanotechnol* 7(1):4–8
14. Al-Turjman F (2020) A cognitive routing protocol for bio-inspired networking in the internet of nano-things (iont). *Mob Netw Appl* 25(5):1929–1943
15. Atlam HF, Walters RJ, Wills GB (2018) Internet of nano things: Security issues and applications. In: Proceedings of the 2018 2nd international conference on cloud and big data computing, pp 71–77
16. Akyildiz IF, Jornet JM (2010) The internet of nano-things. *IEEE Wireless Commun* 17(6):58–63
17. Akyildiz IF, Ghovanloo M, Guler U, Ozkaya-Ahmadov T, Sarioglu AF, Unluturk BD (2020) Panacea: an internet of bio-nanotechnology application for early detection and mitigation of infectious diseases. *IEEE Access* 8:140,512–140,523
18. Bi D, Almpanis A, Noel A, Deng Y, Schober R (2021) A survey of molecular communication in cell biology: establishing a new hierarchy for interdisciplinary applications. *IEEE Commun Surv Tutor* 23(3):1494–1545
19. Zafar S, Nazir M, Bakhshi T, Khattak HA, Khan S, Bilal M, Choo KKR, Kwak KS, Sabah A (2021) A systematic review of bio-cyber interface technologies and security issues for internet of bio-nano things. *IEEE Access*
20. Montoya C, Du Y, Gianforcaro AL, Orrego S, Yang M, Lelkes PI (2021) On the road to smart biomaterials for bone research: definitions, concepts, advances, and outlook. *Bone Res* 9(1):1–16
21. Agoulmine N, Kim K, Kim S, Rim T, Lee JS, Meyyappan M (2012) Enabling communication and cooperation in bio-nanosensor networks: toward innovative healthcare solutions. *IEEE Wireless Commun* 19(5):42–51
22. Zafar S, Nazir M, Sabah A, Jurcut AD (2021) Securing bio-cyber interface for the internet of bio-nano things using particle swarm optimization and artificial neural networks based parameter profiling. *Comput Biol Med* 136:104,707
23. Ali NA, Abu-Elkheir M (2015) Internet of nano-things healthcare applications: Requirements, opportunities, and challenges. In: 2015 IEEE 11th international conference on wireless and mobile computing, networking and communications (WiMob). IEEE, pp 9–14
24. Jarmakiewicz J, Parobczak K, Maślanka K (2016) On the internet of nano things in healthcare network. In: 2016 international conference on military communications and information systems (ICMCIS). IEEE, pp 1–6

25. Abbasi NA, Akan OB (2017) An information theoretical analysis of human insulin-glucose system toward the internet of bio-nano things. *IEEE Trans Nanobiosci* 16(8):783–791
26. Sarker SH (2019) Holistic health improvement using the internet of bio-nano things based treatment
27. Stelzner M, Dressler F, Fischer S (2017) Function centric nano-networking: addressing nano machines in a medical application scenario. *Nano Commun Netw* 14:29–39
28. Canovas-Carrasco S, Sandoval RM, Garcia-Sanchez AJ, Garcia-Haro J (2019) Optimal transmission policy derivation for IoNT flow-guided nano-sensor networks. *IEEE Internet Things J* 6(2):2288–2298
29. Galal A, Hesselbach X (2018) Nano-networks communication architecture: modeling and functions. *Nano Commun Netw* 17:45–62
30. Dressler F, Fischer S (2015) Connecting in-body nano communication with body area networks: challenges and opportunities of the internet of nano things. *Nano Commun Netw* 6(2):29–38
31. Bakhshi T, Shahid S (2019) Securing internet of bio-nano things: MI-enabled parameter profiling of bio-cyber interfaces. In: 2019 22nd international multitopic conference (INMIC). IEEE, pp 1–8
32. Balasubramaniam S, Kangasharju J (2012) Realizing the internet of nano things: challenges, solutions, and applications. *Computer* 46(2):62–68
33. Kalantar-Zadeh K, Ha N, Ou JZ, Berean KJ (2017) Ingestible sensors. *ACS Sens* 2(4):468–483
34. Ali O, Ishak MK, Bhatti MKL (2021) Emerging IoT domains, current standings and open research challenges: a review. *Peer J Comput Sci* 7:e659
35. Giaretta A, Balasubramaniam S, Conti M (2015) Security vulnerabilities and countermeasures for target localization in bio-nanotechnology communication networks. *IEEE Trans Inform Forensics Secur* 11(4):665–676

iTelos—Case Studies in Building Domain-Specific Knowledge Graphs



Simone Bocca, Mauro Dragoni, and Fausto Giunchiglia

Abstract The construction of a domain-specific knowledge graphs is a complex task requiring the synergistic work of domain experts, knowledge engineers, and data scientists. The goal of the iTelos methodology is to support this task when carried out by working groups with competence in computer science and in the targeted domain, but with little know-how in the software and knowledge engineering development process. The concrete usefulness of iTelos is presented by discussing six concrete case studies where it has been applied and evaluated. The main features of iTelos validated in this paper are (i) how to create an explicit and clear purpose for the construction of the domain-specific Knowledge Graph; (ii) how to enhance of the quality of the Knowledge Graph produced by defining an iterative validation of intermediate results; (iii) how to organize the project activities based on the specific roles involved; and (iv) how to make the methodology as automated and iterative as possible, resulting in a more cost-efficient and time-efficient methodology. The evaluation is both qualitative and quantitative. The results clearly show that iTelos is an important step in the right direction, still highlighting the need for refinement and improvement of the methodology.

1 Introduction

Semantic heterogeneity is the phenomenon that arises any time there is a need of integrating two or more data sources, e.g., databases, representing, possibly in part, the same real-world phenomenon [7, 24]. This phenomenon manifests itself in the fact that there is usually a many-to-many relation among the models of the world

S. Bocca (✉) · F. Giunchiglia
University of Trento, Trento, Italy
e-mail: simone.bocca@unitn.it

F. Giunchiglia
e-mail: fausto.giunchiglia@unitn.it

M. Dragoni
Fondazione Bruno Kessler, Trento, Italy
e-mail: dragoni@fbk.eu

adopted by any two independently developed databases. This challenge has been studied extensively in the literature. Two are the main problems which have been tackled. The first is the problem of *schema-level* heterogeneity which has been dealt via the use of *ontologies* [8, 17] as the means for agreements on a fixed language and schema to be shared across applications; see, e.g., [2, 3]. The second, more recent, is the problem of *data-level* heterogeneity, where the idea is to exploit the intrinsic representational flexibility and extensibility of *Knowledge Graphs (KGs)* [4, 9, 14, 18, 19].

However, despite the results developed so far, dealing with the problem of semantic heterogeneity is still a very costly and labor-intensive task, this being the case mainly because of the combinatorial explosion of possibilities that arise by combining the differences at the schema level with those which arise at the data level [6, 16]. Thus, we may be in a situation where, in two datasets, the same entity type is described using different, and possibly incompatible, properties and where, once one has aligned the two schemata, it is still left with the data-level heterogeneity, a problem which can manifest itself with different data types for the same property and very often also with different values [5, 25, 26]. There are two crucial observations on which the work described in this paper is based. First, this problem is very critical in the case of *domain-specific data integration*: independently of the fact that one tries to maximize the reuse of previous schemata and datasets, the necessity of dealing with even a single new data set, as it is always the case when one develops a new application, requires redoing the work from scratch, with little possibility of reuse. Second, ultimately, the cost lies in the data-level integration whose size is always orders of magnitude bigger than that of the schema level. The clear consequence is that the schema-level knowledge integration, which happens before the data-level integration, being a prerequisite for this, *should not be done independently of the actual data* and it should rather be done in a way to minimize the cost of data integration. As discussed in detail in [12], in settings like this, the adoption of data-independent foundational ontology engineering methodologies is not the most effective and efficient strategy,

iTelos is a data integration methodology, whose theoretical foundations and main steps were first introduced in [13]. A first account of iTelos can be found in [11], but see also [15]. Given a certain purpose informally specified, it takes into input a set of pre-existing datasets and ontologies and it produces in output a domain-specific KG that satisfies the requirements specified by the purpose. iTelos is based on three working hypotheses. The first is to stratify the data integration process into problems which can be solved largely independently. Thus we have:

- a *knowledge* integration problem, where knowledge (e.g., the dataset schema and possibly a set of reference ontologies to be reused) will, at the end of the process, be represented as a graph where nodes are entity types and links are properties; and
- a *data* integration problem, where the input datasets will, at the end of the process, be represented as a graph of entities populating the knowledge level of the KG.

This design choice allows us to avoid the combinatorial explosion deriving from the interaction of the two different types of diversity. Notice how the complexity of the data integration problem reduces to the sum of the complexities of the two layers. The second hypothesis is that the techniques developed for each layer can be composed with the ones developed in the other layer irrespective of how heterogeneity appears in the other. The third hypothesis, which is a direct consequence of the second, is that, within each layer, it is possible to exploit the large body of work which can be found in the literature. Finally, another iTelos main design choice is that the data and knowledge layers are dealt in parallel, through a sequence of steps where the integration at both levels is progressively refined *in parallel*, a those taken in the other level.

The main goal of this paper is to briefly introduce the iTelos methodology and to show how it has been effectively exploited in six case studies, each developed over a period of fourteen weeks by a group of up to four people with mixed background, expertise, and skills.

2 Methodology

In this section, we introduce iTelos by describing each step, the outcome expected for each step, and the roles involved. The description is quite synthetic, the reader is urged to consult [11] for more details. Relevant material can also be found on the Web site of the iTelos class.¹

The iTelos methodology has four main objectives: (i) to model the initial purpose for the construction of the domain-specific KG; (ii) to enhance the quality of the KG; (iii) to organize the development activities with reference to the specific project roles; and, (iv) to make the methodology as automated and iterative as possible, resulting in a more cost-efficient and time-efficient methodology. It is organized into five main development phases, namely *Scope Definition*, *Inception*, *Informal Modeling*, *Formal Modeling*, and *Data Integration*.

Figure 1 shows a top-level view of the iTelos process. The iTelos process takes in input a set of data and knowledge repositories and an informal specification of the problem to be solved and it produces in output a KG satisfying the input specification and generated by integrating a number of datasets to be found in the input repositories. In Fig. 1, by SKG (see formal modeling phase) we mean the Schema of the KG and by DKG (see Data integration phase) we mean the actual (Data) KG produced as a result of the overall process.

We have the following highlights. First, the five phases are connected via specific evaluation activities that are used to evaluate intermediate and final outputs. Second, within each phase, two main levels have been identified, namely the *Schema* and *Data* level, the first aiming at generating the KG schema, the latter aiming at generating the actual KG. Third, notice how these phases might be executed any number of

¹ <https://unitn-kdi-2021.github.io/unitn-kdi-2021-website/>.

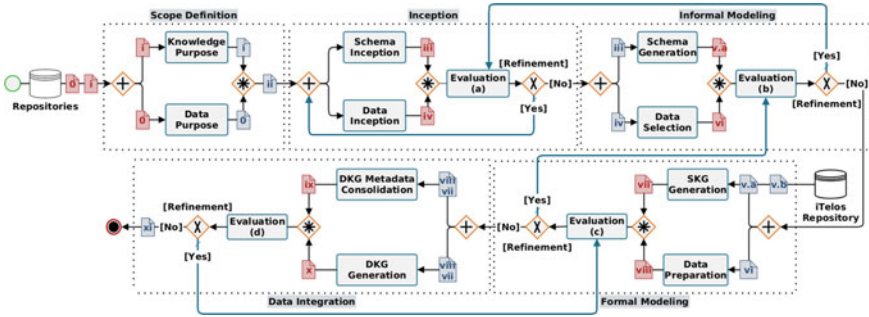


Fig. 1 Top level view of the iTelos methodology

times. An iteration in the iTelos methodology can happen for two reasons. The first reason is because of stratification and sequentialization of the work within each phase, with a planned production of intermediate results for that phase. The second reason is the failure of the evaluation of what was initially considered the final output of that phase. This process can iterate back all the way till the scope definition. As such, if, for example, the outcome of the *Formal Modeling* phase does not satisfy the requirements, it might be necessary to return to the *Inception* phase because of the impossibility of solving in the *Informal Modeling* phase the problem originally detected.

iTelos foresees three main roles involved throughout the entire KG creation process:

- *Domain Expert*: she coordinates the project. She is interested primarily by the output of the project as envisaged to be the main stakeholder interested in the KG to be constructed.
- *Knowledge Engineer*: she is assigned the development and execution of the schema-level activities. She is mainly interested in those activities in which a knowledge of ontology, schema, and knowledge modeling is required.
- *Data Scientist*: she is assigned the development and execution of the data-level activities. She is mainly interested in those activities where a knowledge of ETL techniques (extract, transform, load)² and data science techniques is required.

For each main phase, we discuss below the top-level view, plus the activities at the schema level, at the data level, and the evaluation (if present). The methodology includes also guidelines about the deliverables to be produced in each phase, the tools to use, and illustrative examples.

² <https://it.wikipedia.org/wiki/Extract,Transform,Load>.

2.1 Scope Definition

The *Scope Definition* is the first phase of the iTelos methodology. It aims to define exhaustively the project’s purpose, and more in detail it has to answer the following questions: (i) Why the iTelos methodology has to be adopted? (ii) Which is the problem’s context and how it is defined? (iii) Which problem the methodology will solve in the context? Keeping the focus on the questions above, the first phase aims also to identify and localize the data needed to solve the problem. The main role interested in both macro-activities is the *Domain Expert* which describes the problem both at *Schema* and *Data* level. The *Scope Definition* phase contains two macro-activities: *Knowledge Purpose* (Fig. 2) and *Data Purpose* (Fig. 3).

Concerning the *Knowledge Purpose*, in the *Problem Context* sub-activity the Domain Expert identifies and defines the context in which the KG will work in. The context definition is crucial in order to identify the correct datasets and how to use them. The Domain Expert has to define the fundamental aspect of the problem’s context, above all the space and time requirements. So, for instance, the requirements are usually organized in the following three dimensions: (i) Geographical aspects are the geographical scope in which the problem tries to solve a certain set of tasks, if specific to a certain location (i.e., a Transportation solution in a given city); (ii) Temporal aspects are the temporal scope in which the problem solves a certain set of tasks such as the future, past or present data being given to study and execute a certain task (i.e., a Museum requiring an easy-to-use paintings or monuments information visualization product); and (iii) aspects specific to the domain of application. The definition of these aspects enables the Domain Expert to have the fundamental elements for defining the problem as well as the objects used to solve it. In the *Problem Why* sub-activity

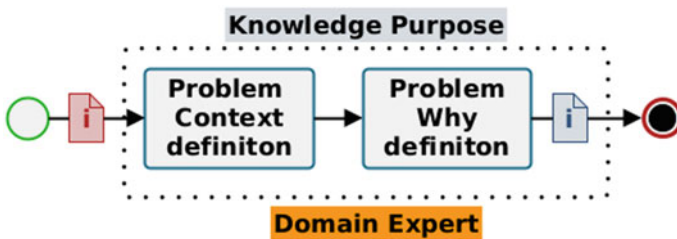
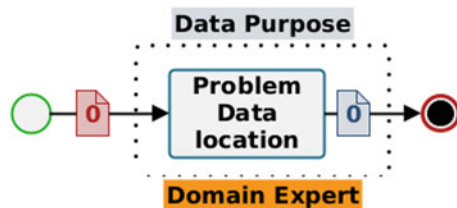


Fig. 2 Knowledge purpose diagram

Fig. 3 Data purpose diagram



the Domain Expert has to integrate the documentation, started in the previous step, with the information regards why the problem has to be solved. More in detail, the main aspects to define are: the purpose for creating the KG, the personas, and the scenarios relative to the problem.

In the *Data Purpose*, the main activity is *Problem Data location*. This sub-activity describes how the project’s purpose definition influences the identification of the datasets used for instancing the project. Here, the Domain Expert needs to produce a list of main data sources to be considered to collect data for the project, e.g., open data repositories, data sets, private databases, and web pages to gather data.

2.2 Inception

The second phase, called *Inception*, starting from the documentation coming from the previous phase, aims at defining the *Competency Questions*, that in the end of this phase will become *Competency Queries*. Competency queries enable a more precise definition of the data that, ultimately, the KG needs to be able to provide.

The Knowledge Engineer and the Data Scientist are respectively in charge of the *Schema Inception* and *Data Inception* activities (Figs.4 and 5). Starting from this phase the work is scheduled following an iterative process, as from Fig. 1.

The objective of the *Schema Inception* sub-phase is to describe the first steps needed in the definition of the structure of the data schema that the Knowledge Engineer has to define. The main sub-activities being executed are (i) the definition of the Competency Questions, (ii) the selection of the data objects; and, (iii) the definition of the generalized queries. Concerning the definition of the competency questions,

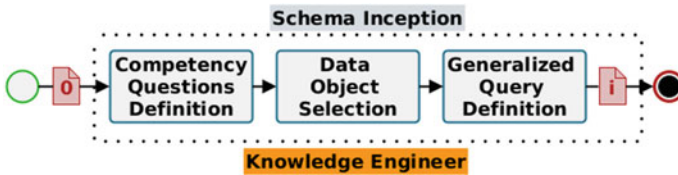


Fig. 4 Schema inception diagram

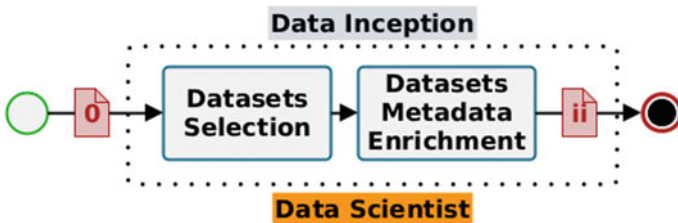


Fig. 5 Data inception diagram

the Knowledge Engineer starts from the personas as well as the scenarios definitions coming from the previous phase, and proceeds to list (i.e., interviewing specific actors in specific scenarios) all the possible questions and relative answers, relevant to solve the problem. The Knowledge Engineer categorizes the Competency Questions collected by following the different data typologies. *Core*: they describe the most important entities and properties relevant to the problem space. These entities are the subject of the project, the most important in the solution achievement. *Common*: they describe the common entities and properties relevant to the problem space (such as time and location). These entities are the most used in order to define common aspects of the world in which the data live. *Contextual*: they describe the extensions of the entities and properties relevant to providing specific details to the problem space. These entities are used to describe domain-specific aspects of the data. The *Data Object Selection* starts using the Competency Questions previously defined. The scope of this internal step is to identify and list, in a preliminary and general way, the main data object involved in the questions defined before. These general objects are the first version of what will be called *etypes*.

The last sub-activity of the schema level, in this phase, aims at defining more precisely all kinds of queries that can be useful in the solution achievement. To obtain this result, the Knowledge Engineer uses the refined defined Competency Questions together with the data objects defined in the previous step, into a list of queries in a more precise format (i.e., SQL-like language). The output of this sub-activity will be a document defining a semi-structured version of the queries needed to support, merged with a first definition of the kinds of objects that have to be handled within the project.

Concerning the *Data Inception*, the two sub-activities executed are the selection of the datasets, and their enrichment with metadata, specific to each dataset selected. The Data Scientist is in charge of this task which will achieve by exploiting her skills in ETL techniques. Within the datasets selection the Data Scientist has to analyze all the data sources listed in the purpose documentation. The objective of this step is to identify, within those sources, the single datasets needed to solve the project's problem. Finally, within the datasets' metadata enrichment task, the Data Scientist will enrich the selected datasets with record-level metadata. Examples of metadata are provenance and timestamp.

2.3 Informal Modeling

The *Informal Modeling* is the third phase of the iTelos methodology. This phase deals with the definition of the informal model in terms of the main entity types related to the problem and the filtering of the preliminary datasets produced in the *Inception* phase. The Knowledge Engineer, during this phase, needs to proceed with the *Schema Generation* while the Data Scientist needs to produce the final *Data Selection*. See Figs. 6 and 7.

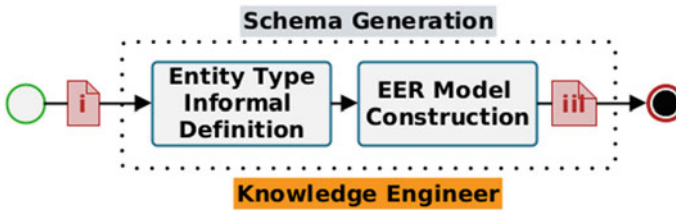


Fig. 6 Schema generation diagram

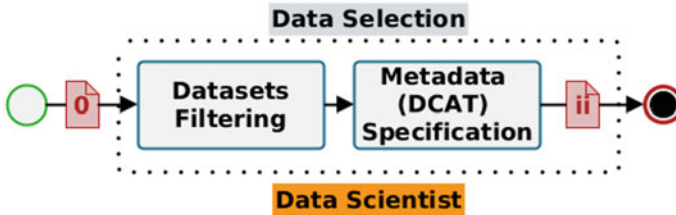


Fig. 7 Data selection diagram

Within the *Schema Generation* activity, the Knowledge Engineer generates the informal schema model of the problem, starting from the Competency Queries and data objects definitions coming from the previous phase [13]. This model includes the definitions of all the entity types and relative attributes, needed to achieve the project purpose, as well as the relations among those kinds of entity types. In the *Entity Type Informal Definition*, the Knowledge Engineer needs to define all the possible kinds of entities to be used in the problem and to find the properties that describe each entity. iTelos categorizes three kinds of Entity types and properties, namely: *Common*, *Core*, and *Contextual*, which are generated based on the organization of competency queries and data generated during Inception (see above). Then, in the *Extended ER (EER) Model Construction* the Knowledge Engineer will produce a detailed first informal Entity–Relationship model based on the etypes defined beforehand and following the relations defined within their properties. This model is constructed using an EER Model. This makes it easier for the Knowledge Engineer to edit and transform the model as needed and by making it easier to visualize the entire problem through a relationship model.

The *Data Selection* activity is performed by the Data Scientist and it aims at providing the final selection of the datasets needed in the project. Together with the data, also the metadata are specified better, eventually collecting new information, and standardized using DCAT standard.³ Within the *Datasets Filtering* sub-activity, the preliminary datasets extracted in the Inception phase are filtered through various ETL procedures performed by the Data Scientist. The objective is to finalize the dataset selection in order to have all the datasets needed for the project, to reach a data

³ <https://www.w3.org/TR/vocab-dcat-2/>.

status where no other information has to be collected anymore (at most manipulated). While, during the *Metadata Specification*, the Data Scientist is tasked with improving the metadata collection, adding new information regarding the operations performed on the data in the previous sub-activity. Moreover, a dedicated standard is adopted to categorize all the metadata collected, which will be one of the main outputs of the project.

2.4 Formal Modeling

Formal Modeling is the fourth phase of the iTelos methodology. This phase is in charge of the construction of the final schema of the KG. This schema is annotated and composed with a set of alinguistic concepts and language terms (i.e., words). Giunchiglia et al. [15] provides a precise description of how SKGs, which we also call L4 schemas, are defined using a language agnostic terminology, that we call L1 concepts. L1 concepts support the generation of multilingual KGs, being annotated by the multiple words, which we call L2 language terms, that, in different languages, linguistically denote them. The key intuition is that an L4 schema is a formal object where all etype and property names (i.e., L2 language terms) are given univocally defined meanings by associating them with WordNet-like senses (i.e., L1 concepts) [10, 22]. Figures 8 and 9 present the details of the two macro-activities performed in this phase.

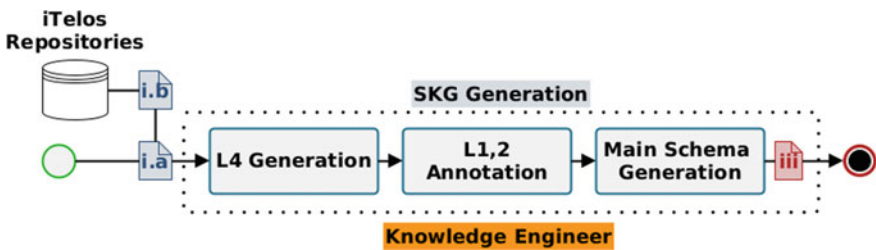


Fig. 8 SKG generation diagram

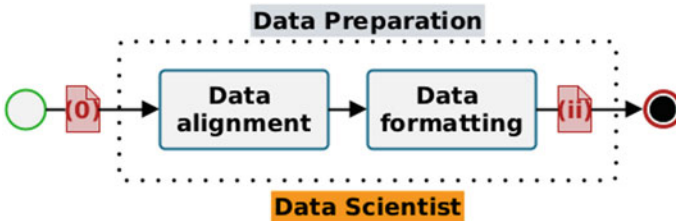


Fig. 9 Data preparation diagram

During the *SKG Generation*, three sub-activities have to be performed: (i) the L4 schema Generation; (ii) the L1, L2 Annotation; and, (iii) the Main Schema Generation. During the *L4 Generation* sub-activity, the Knowledge Engineer aims at obtaining a new formal version of the knowledge schema (L4) starting from a specific type of reference ontology, called Teleology [12], selected from the iTelos repository, and the informal schema produced as an output of the previous phase. Then, during the *L1-2 Annotation* sub-activity, the Knowledge Engineer has to identify within the Teleology those lexicon-semantic elements of L1 and L2 that can be used to represent etypes and etype properties for the data which have to be integrated. Finally, during the *Main Schema Generation* sub-activity, the two schemata mentioned above are merged for generating the final SKG.

During the *Data Preparation* activity the Data Scientist is in charge of performing the *Data alignment* and *Data formatting* sub-activities. The Data Preparation macro-activity aims at handling the data identified and extracted during the previous phases (*Inception* and *Informal modeling*). The *Data alignment* sub-activity compares the data extracted from the *Informal Modeling* phase, with the informal definition of the etypes, trying to understand if the data are correctly shaped to be represented by those etypes. In case differences appear between the data form and the etypes structure, this activity will reorganize the data with the objective of reducing as much as possible the gap between the data layer and knowledge layer. Then, the *Data formatting* sub-activity has the same objective as the previous activity, but focused on the values of the data instead of their structure. It checks and models the data values to ensure that those values respect the data types for the data that they represent. The information about the correct data types to adopt comes from the schema defined for the respective data.

The output of this phase is a formal model which takes the format of an ontology, formalized in Description Logics [1], more precisely in the OWL language,⁴ and represented in RDF⁵ There are three levels of criteria according to which the ontology is evaluated against: schema, linguistic, and metadata levels. The schema-level evaluation checks the logical perspective of the conceptual modeling, which covers the following three dimensions: (i) Consistency (i.e., to verify the logical satisfaction of the knowledge system); (ii) Accuracy (i.e., to verify the correctness of the modeling); and, (iii) Completeness (i.e., to verify the model as an organic whole). At the linguistic Level, the evaluation aims at the usage perspective of the model, which covers the compliance and understandability dimensions. Finally, at the metadata Level, the evaluation focuses on the data which need to keep track of all of the information about the process of schema construction. Examples of metadata are: the *creator*, *version*, *construction date*, *purpose*.

⁴ <https://www.w3.org/OWL/>.

⁵ <https://www.w3.org/RDF/>.

2.5 Data Integration

The *Data Integration* is the final phase of the iTelos methodology. This phase is concerned with the construction of the KG and the relative Codebook for the metadata collection as well as the relative deliverables. In iTelos by Codebook, we mean a document, typically in textual form, to be used by future developers, which contains a detailed description of the etypes used and their properties, element by element. In this phase, the Knowledge Engineer is interested in the Schema-level macro-activities of the Data Knowledge Graph, while the Data Scientist is interested in the Data-level macro-activities. Figures 10 and 11 describe the articulation of these two sub-phases. Within the *Project Metadata Consolidation* the main activities executed by the Knowledge Engineer are the *DKG metadata collection*, the *Codebook Documentation*, and the realization of the *Project report and Slides*. The *DKG metadata collection* sub-activity takes in input the formal SKG and the data produced, both defined in the previous phase, with the scope to identify and collect useful metadata about schema elements and data extracted. In the *Codebook Documentation*, the Knowledge Engineer concentrates on collecting together and describing all the metadata identified in the previous sub-activity into the Codebook. In the last sub-activity, the *Project report and Slides*, the Knowledge Engineer has to finalize the documentation which has to be produced as part of the final output of the whole methodology.

Finally, within the *DKG Generation* activity, the sub-activities being executed are the *Data Mapping* and the *EML data import*. In this context, by *EML*, we mean a

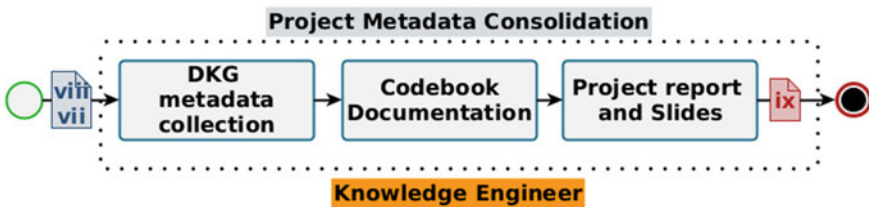


Fig. 10 DKG metadata consolidation diagram

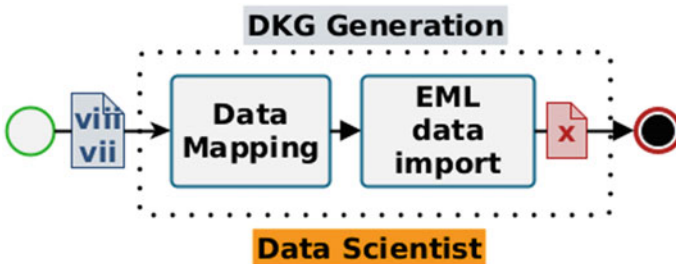


Fig. 11 DKG generation diagram

specific language used to represent both SKGs and DKGs which is at the core of the iTelos methodology [15]. In this phase, all the datasets are merged (i.e., suitably imported) in a single DKG. During this final part of the process, the Data Scientist maps the “well formed” entity data received as input from the previous phase with the SKG previously defined, therefore, generating the DKG. This is performed via a data mapping tool, called *KarmaLinker* [15], which automates the mapping activities related to both etypes and etype properties, also allowing some final minor cleaning operations on the input data.

3 Case Studies

The iTelos methodology described in the previous section has been applied and validated on the six case studies described below. Due to space reasons, we provide their descriptions only. The interested reader may refer to our GitHub⁶ repository for downloading the created knowledge graphs and the produced documentation.

Geospatial: A large part of the information we deal with on a daily basis has some kind of geographic dimension. In our private life, we might be looking for different mountains nearby our homes for a memorable hiking or trekking experience. In our professional life, we may be interested in studying consumer patterns in Western Europe, perhaps characterized by the Human Development Index of the country the customer is a resident of. Often the information required to answer our queries is available, but dispersed among a multiplicity of geospatial information sources. Geospatial data or geographic information is the data that identifies a geographic location of natural or constructed features and boundaries on the Earth (e.g. oceans, countries, rivers, mountains, landforms, administrative divisions, etc.). The aim of this case study is to make information seeking easier by allowing exploration, editing, and interlinking of heterogeneous information sources with a spatio-temporal dimension, by developing a modularized knowledge graph using the iTelos methodology.

Transportation: Transportation involves the particular movement of humans, goods or other things from a point A (a place in space) to a point B [i.e. to and from different locations]. Modes of transport include air, land (rail and road), water, cable, and pipeline. The field can be divided into infrastructure, vehicles, and operations. Transportation infrastructure consists of the fixed installations, including roads, railways, airways, etc., and terminals such as airports, railway stations, trucking terminals, etc. Vehicles traveling on these networks may include automobiles, trains, trucks, aircrafts, etc. The aim of this case study is to create a modularized knowledge graph on transportation for data integration, using the iTelos Methodology.

Covid-19: This case study aims to collect and manage medical data. The medical data domain is composed by a large number of aspects involving persons, structures, medicines, treatments, and not only those. For this case study, the focus selected

⁶ <https://github.com/UNITN-KDI-2020>.

is the current pandemic situation of the Covid-19. A huge amount of data have been generated due to the virus, regarding infected people, symptoms, medicines and therapy adopted and so on. Due to the fact that the each country generated its personal statistics on the pandemic, starting from several different data sources, the representations of the data on this topic are defined and structured in several different ways.

Health Data Integration (FHIR-based): This case study aims to collect and manage medical data. The medical data domain is composed of a large number of aspects involving persons, structures, medicines, treatments, and not only those. For this case study, the focus selected is the collection of data relative to persons, or patients. By retrieving medical data on different aspects regarding patients, it is possible to generate a set of Electronic Health Record (EHR) describing the medical situation of a set of patients. In the European context, the case study has to consider the differences between the different countries' health systems, involving different data standards and format adopted to represent the data.

Tourism Facilities: Facilities are geospatial infrastructures that are crucial to the efficient and effective delivery of support services for the organizational ecosystems that they serve, with the ultimate purpose of improving the quality of life of people and the productivity of the core business. Facilities can be of diverse nature considering the services they provide, e.g., Healthcare Facilities (Hospitals, Residential Care Homes, etc.), Educational Facilities (different arms of universities), Tourism Facilities (Holiday Cabins, Amusement Parks, Safaris, etc.). The aim of this case study is to create a modularized knowledge graph on tourist facilities for data integration, with a particular focus on its interoperability with the Geospatial domain by using the iTelos Methodology.

Tourist Events: Smartphones are powerful data-centric gadgetry facilitators to understand their user's personal life patterns. In case of a tourist, for example, it can help to pinpoint where a tourist goes and can possibly go, possible amusement activities, and to exploit this knowledge to support the tourist's itinerary. Personal tourist events are focused along four macro cardinals: (i) Spatial (location patterns); (ii) Temporal (time patterns); (iii) Personal (behavioral patterns); and (iv) Social (interaction patterns). The aim of this case study is to create a modularized knowledge graph for integrating data on personal tourist events by using the iTelos Methodology.

4 Evaluation

The users engaged in the case studies were students of the Knowledge and Data Integration course at the University of Trento. They were organized into teams and then assigned to a specific case study. Each case study consisted of a real-world scenario aiming at building a knowledge graph to be used within a concrete application. Each group was supported by a Domain Expert and, depending on the skills of each member, the roles of project manager, knowledge engineers, and data scientists were assigned

Table 1 Evaluation's subjects in KDI class—2018, 2019, 2020

	2018	2019	2020	Tot
# Students	29	20	26	75
# Project teams	14	4	6	24
% Male	69%	75%	95%	59
% Female	31%	25%	5%	16
% Undergraduate	20.7%	20%	25%	16.5
% Post graduate	79.3%	80%	75%	59

The choice of this setup is twofold. First, we were able to observe the extent to which the iTelos methodology was appropriate for driving a group of skilled computer scientists having limited background in knowledge engineering through the entire modeling and population process. This is an important aspect if we consider that the iTelos methodology aims at supporting the creation of high-quality knowledge graphs by people with little or no expertise in knowledge modeling. Second, the context of the course allows to define a proper schedule for applying the entire methodology and for observing the timing aspects related to the execution of each phase. Notice that preliminary versions of the iTelos methodology were adopted and validated within the context of European projects. However, we noticed that, because of the misalignment between the project work plan (defined without taking into account the constraints posed by iTelos) and the methodology timing and process requirements, this was not the proper setting for a thorough validation of the entire process (though we were able to validate various sub-processes).

Below, we report the evaluation of the methodology related over a period of three years. Throughout this period, the methodology has been refined after each course cycle by addressing the feedback provided by the students and the experts. Table 1 shows the demographic distribution of the users engaged for validating the iTelos methodology from both the quantitative and qualitative perspectives.

We can observe that, in the first (2018) application of the methodology, the groups were smaller compared to the following years. This was in fact a consequence of the first year evaluation. Thanks to this important feedback, we decided to reorganize the group composition with the setting that we implemented in both 2019 and 2020. The summary of the quantitative evaluation collected over the last three years is reported in Fig. 12. We show three heat maps containing the distribution of the answers to the four quantitative questions submitted to users. All questions have foreseen a scaled rating within the range [1, 7], where 1 represents the most positive score and 7 the most negative one. Our aim was to observe if, throughout the entire three years period, the overall evaluation of the methodology revealed an increasing appreciation from the users. This point is validated by observing the heat maps related to the years 2018 and 2019. Indeed, the reader can appreciate the scores shifted to the left part of the heat map. Differently, from the evaluation collected in 2020, we observed a worsening of the scores (i.e. shift to right). By combining the quanti-

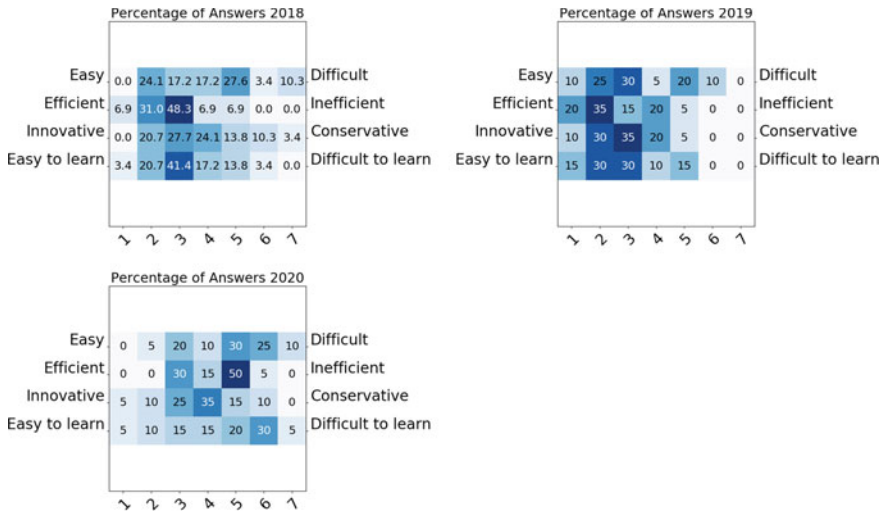


Fig. 12 Usability of the methodology: a 2018, b 2019, c 2020

tative evaluation with the analysis of the qualitative one (described in Sect. 5), we concluded that during the first two years, we applied the iTelos methodology within an offline context where the face-to-face interaction was predominant. Instead in the last year, given the pandemic, we applied entirely the iTelos within an online context for the first time. This was an important test bed since one of the long-term objectives of iTelos is to support the distributed collaboration between people in creating and deploying high-quality knowledge graphs. Hence, on the one hand, we observed how the refinement process of the methodology, we setup for finding the best trade-off between effectiveness and efficiency in building knowledge graph, is proper. On the other hand, the deployment of the methodology within a complete online context triggered the need for further refinements to increase its overall usability.

5 Lessons Learned and Discussion

On top of the quantitative evaluation provided in Sect. 4, we also asked the users to answer a questionnaire containing a set of qualitative queries with the aim of collecting more fine-grained feedback about the strengths and weaknesses of the methodology. Below, we report the main lessons we learned from users’ feedback and we provide some insights about the future directions that will support us in evolving the proposed methodology.

Weaknesses. The qualitative analysis of users’ feedback highlighted three main weaknesses, as follows.

First—cost of early mistakes. The users noticed that the effects of wrong assumptions during the first steps of the methodology propagate pervasively in the next steps. The only very costly solution is to roll back at the step where wrong assumptions were actually done. Indeed, it would be quite difficult to change the model once a team has established fundamental concepts which reveal to be wrong. A possible mitigation would be the increment of the middle-evaluations step. However, this would reduce the efficiency of the methodology. This aspect will be considered within the next refinement iteration of the methodology.

Second—usefulness of an upper ontology [8, 17]. It was observed that a phase dedicated to the detection and, if it is required, the definition of an upper ontology should be introduced. During the *Informal Modeling* phase, the users need to reason about and generate the model to be used to generate the target KG. This point emerged during some case studies where it was noticed that the use of an upper ontology could speed up substantially the KG construction process.

Third—unbalanced work between schema and data layers. The iTelos methodology has been thought for supporting a team covering the roles described in Sect. 2. The coordination of such a working team includes proper synchronization and the equal distribution of the work. Hence, the skills of the project manager are crucial for the good success of the case studies. In some of them, this did not occur due to a (limited) misalignment between the knowledge engineers and the data scientists during the process. Such a misalignment was partially caused by the unbalanced work between the two roles in some of the phases, in particular, during the *Informal Modeling* one. Finally, there was the risk that such a misalignment led to a not complete understanding, by each team member, of the work done by the members covering the opposite role. This aspect is definitely interesting since it highlights the need of improving both the collaborative and monitoring aspects of the methodology.

Strengths. The qualitative analysis revealed also some important strengths. We report below the four which we consider most relevant.

First—easy step by step structure. The most prominent strength of the methodology is related to the clear and well-organized process that users have to follow for building the final knowledge graph. About this point, a substantial amount of feedback has been provided. Examples of that are the clear separation of the various steps of the methodology, the fact that the step by step procedure eases the path to follow during the entire process, and the revision aspect introduced in each phase allowing the capability of fixing and refining the intermediate outputs at any time.

Second—intermediate evaluation process. In this paper, we did not dedicate too much space to the iterative and evaluation aspects since we focused more on describing the main tasks of each phase. The iterative approach and the evaluation task that have to be performed within each phase have been particularly appreciated by the users. Indeed, from the collected feedback, we observed how the methodology incremental approach has been perceived as a way for improving the overall outcome of the case study. At the same time, such an approach has been considered an effective support for the refinement activities of the created artifact. We intend to dedicate further effort in the future for better detailing and testing of all the evaluation activities introduced in the methodology. The goal will be to find the best trade-off between an efficient

knowledge graph construction and an intensive roll back at the beginning of the process for fixing the output of wrong decisions (as reported in the first weakness). *Third—use of an effective set of tools.* One of the key aspects of iTelos is the set of tools used. They are both external third party tools, and tools developed internally for the methodology. Let us describe them in brief. During the scope definition phase iTelos assumes the usage of tools, such as *Overleaf*⁷, facilitating the documentation production. The inception phase requires the usage of any kind of spreadsheet editor to clearly define the *Competency Questions* and *Queries*. Moreover, the data selection activity, executed in this phase, may require scraping and data management tools to collect and compose all the necessary datasets. Due to that, for the inception phase, in the case studies described in Sect. 3, a Python *Notebook Jupyter*⁸ environment, including several data manipulation libraries (i.e., *Pandas*⁹), has been used. In the informal modeling phase, iTelos users have access to a Data Integration (DI) platform which, through a dedicated API layer, offers access to a Knowledge Base (KB) used to define and maintain the Etypes and properties composing the SKG. Moreover, in this phase, a diagram editor (i.e., *yEd*¹⁰) is required to produce the *EER Model*. The formal modeling phase requires an ontology editor (i.e., *Protégè* [23]), in order to formally define the SKG, following the OWL language, within an RDF file. The SKG formalized is then imported in the KB through *KarmaLinker* (see above), an extension of the *Karma* data integration tool [20, 21]. The iTelos users have shown substantial appreciation of the fact the exposure to such a substantial set of state-of-the-art tools and technologies. On the one hand, this could be perceived as a barrier due to the necessity of dedicating time and effort to learning such tools. On the other hand, instead, this aspect seems to be appreciated since it enables the growth of the users' technological skills.

Fourth—focus on purpose. Last, but not least, the strong point highlighted within the qualitative evaluation is how the iTelos methodology is particularly suited for working on real-world case studies. The main key enabling this factor is the fact that iTelos supports, since the early steps, the focus on the concrete problems which might arise at any moment during development. The purpose, in particular, seems to be the key enabler for effective and efficient construction of a KG.

6 Conclusion

In this paper, we have presented an evaluation of iTelos: a step by step methodology for building KGs in a collaborative way. We have briefly described the structure of the methodology, the case studies in which it has been applied, and the quantitative and qualitative evaluations we collected from users. The results obtained have

⁷ <https://it.overleaf.com/>.

⁸ <https://jupyter.org>.

⁹ <https://pandas.pydata.org>.

¹⁰ <https://www.yworks.com/products/yed>.

demonstrated the overall effectiveness and efficiency of iTelos in supporting teams with relatively limited skills in the design and deployment of a working KG. They have also highlighted some important refinements, in particular, about how to deploy iTelos within a complete online setting.

Acknowledgements The research conducted by Fausto Giunchiglia and Simone Bocca has received funding from the “*DELPhi - Discovering Life Patterns*” project funded by the MIUR (PRIN) 2017. The authors thank Mayukh Magchi and Alessio Zamboni for their support and contributions in the generation and teaching of the KDI material.

References

1. Baader F, Horrocks I, Sattler U (2004) Description logics. In: Handbook on ontologies. Springer, pp 3–28
2. Bella G, Elliot L, Das S, Pavis S, Turra E, Robertson D, Giunchiglia F (2020) Cross-border medical research using multi-layered and distributed knowledge. KI-Kunstliche Intelligenz
3. Bella G, Giunchiglia F, McNeill F (2017) Language and domain aware lightweight ontology matching. J Web Semant 43:1–17
4. Bonatti PA, Decker S, Polleres A, Presutti V (2019) Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371). In: Dagstuhl Reports, vol 8. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik
5. Bouquet P, Giunchiglia F (1995) Reasoning about theory adequacy. A new solution to the qualification problem. Fundamenta Informaticae 23(2, 3, 4):247–262
6. Chen C, Golshan B, Halevy AY, Tan WC, Doan A (2018) Biggorilla: an open-source ecosystem for data preparation and integration. IEEE Data Eng Bull 41(2):10–22
7. Euzenat J, Shvaiko P et al (2007) Ontology matching, vol 18. Springer
8. Gangemi A, Guarino N, Masolo C, Oltramari A, Schneider L (2002) Sweetening ontologies with dolce. In: International conference on knowledge engineering and knowledge management. Springer, pp 166–181
9. Giunchiglia F, Dutta, Maltese V (2014) From knowledge organization to knowledge representation. Knowl Organ 41(1)
10. Giunchiglia F, Batsuren K, Bella G (2017) Understanding and exploiting language diversity. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence (IJCAI-17), pp 4009–4017
11. Giunchiglia F, Bocca S, Fumagalli M, Bagchi M, Zamboni A (2021) iTelos—purpose driven knowledge graph generation. arXiv preprint [arXiv:2105.09418](https://arxiv.org/abs/2105.09418)
12. Giunchiglia F, Fumagalli M (2017) Teleologies: objects, actions and functions. In: ER- international conference on conceptual modeling. ICCM, pp 520–534
13. Giunchiglia F, Fumagalli M (2020) Entity type recognition—dealing with the diversity of knowledge. In: Knowledge representation conference (KR). Rhodes, Greece
14. Giunchiglia F, Shi D (2021) Property-based entity type graph matching. In: 16th international workshop on ontology matching, co-located with ISWC, vol 3063. CEUR-WS, pp 25–36, also: arXiv preprint [arXiv:2109.09140](https://arxiv.org/abs/2109.09140)
15. Giunchiglia F, Zamboni A, Bagchi M, Bocca S (2021) Stratified data integration. arXiv preprint [arXiv:2105.09432](https://arxiv.org/abs/2105.09432)
16. Golshan B, Halevy A, Mihaila G, Tan WC (2017) Data integration: after the teenage years. In: Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI symposium on principles of database systems, pp 101–106
17. Guarino N, Welyt C (2002) Evaluating ontological decisions with ontoclean. Commun ACM 45(2):61–65

18. Kejriwal M (2019) Domain-specific knowledge graph construction. Springer
19. Kejriwal M, Knoblock CA, Szekely P (2021) Knowledge graphs: fundamentals, techniques, and applications. MIT Press
20. Knoblock C, Szekely P (2015) Exploiting semantics for big data integration. *AI Mag* 36(1)
21. Knoblock C, Szekely P, Ambite J, Goel A, Gupta S, Lerman K, Muslea M, Mallick MTP (2012) Semi-automatically mapping structured sources into the semantic web. *Extended semantic web conference*, pp 375–390
22. Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ (1990) Introduction to wordnet: an on-line lexical database. *Int J Lexicogr* 3(4):235–244
23. Noy NF, Crubézy M, Fergerson RW, Knublauch H, Tu SW, Vendetti J, Musen MA (2003) Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. In: *AMIA... annual symposium proceedings. AMIA Symposium*, pp 953–953
24. Shvaiko P, Euzenat J (2011) Ontology matching: state of the art and future challenges. *IEEE Trans Knowl Data Eng* 25(1):158–176
25. Sleeman J, Finin T (2013) Type prediction for efficient coreference resolution in heterogeneous semantic graphs. In: *2013 IEEE seventh international conference on semantic computing. IEEE*, pp 78–85
26. Sleeman J, Finin T, Joshi A (2015) Entity type recognition for heterogeneous semantic graphs. *AI Mag* 36(1):75–86

Comparative Study of Image Encryption and Image Steganography Using Cryptographic Algorithms and Image Evaluation Metrics



Surya Teja Chavali, Charan Tej Kandavalli, T. M. Sugash, and G. Prakash

Abstract In this paper, Steganography is carried out by converting the message hidden into ciphertext using renowned cryptographic algorithms like Advanced Encryption Standard (AES) and Triple Data Encryption Algorithm (TDEA or Triple DEA). The embedding technique produces a Stego image. Considering this as the first level of encryption, we then encrypted the whole stego image using AES and TDEA algorithms, giving us the second layer of encryption. The encrypted stego image will be returned in enc format, unreadable without the proper key. We have good resistance to brute-force, statistical, and differential attacks due to the experimental results of two tiers of protection. In addition to the encryption and decryption process, this paper also talks about Image evaluation metrics like PSNR value and SSIM value, performed on stego images, which use ciphertext from AES and TDEA algorithms that provides the statistical information.

Keywords Image Encryption & Decryption · Image Steganography · Hashing · SHA—256 · MD5 · AES · Triple-DES · PSNR · And SSIM

1 Introduction

Our transactions, communications, personal information, and private data are all in desperate need of security. As cyber-attacks become more common, we must protect our data using well-established and secure cryptographic methods or approaches. Cryptography studies secure communications mechanisms that only the sender and authorized receiver may use or access the data using a key. Steganography conceals data in a non-suspicious manner. The method of hiding data inside an image file is called Image Steganography. The cover picture is chosen for this purpose, while the

S. T. Chavali · C. T. Kandavalli · T. M. Sugash
Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, Karnataka, India

G. Prakash (✉)
School of Computer Science and Engineering, Vellore Institute of Technology, Tamil Nadu 632014 Vellore, India
e-mail: g.prakash@vit.ac.in

stego image is acquired following steganography. In this paper, we employed well-known cryptographic techniques such as the Advanced Encryption Standard (AES) and the Triple Data Encryption Algorithm (TDEA or Triple-DES) for encrypting the message that needs to be hidden in the cover image.

We considered a text message a private key, applied AES encryption for that message, and used TDEA encryption for the exact plain text. This image steganography is done by embedding the ciphertext by systematically altering the pixel values so that the actual data of the cover image is not much disturbed and not observable to a person other than the sender and authorized receiver. Then, we performed some Image evaluation metrics on images with stego images with ciphertext from AES and Stego images with ciphertext from TDEA. We then compared and analyzed the performance of these two kinds of images. Image encryption may be described as the act of encoding a secret image with the assistance of an encryption algorithm so that unauthorized users cannot access it. Once we had our Stego Image ready, we utilized the same AES TDEA encryption techniques to encrypt the entire image using a private key as the second level of encryption. We return the encrypted picture in “enc” format when Image Encryption is completed, which cannot be accessed without the actual key used in the encryption procedure. With this, our second and last layer of encryption is now complete. The comparison and analysis part discusses the Image evaluation metrics carried out on stego images. We considered metrics like Peak Signal—To Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) to give us the statistical information between the cover (original) and stego images when encoded with different ciphertexts from different algorithms used.

2 Literature Survey

Different secret text embedding techniques along with DES encryption have been considered previously, such as in [14]. The authors have used K—means clustering to group all the 3-pixel values in the image (R, G, B) and then store the secret text in partitions using LSB and MSB. Unlike this, we have found and characterized a research gap for this application. As a reason, we have used a unique approach, we have used ASCII to bit conversions, for carrying out the process of steganography.

In [16], the authors have considered a fuzzy logic system, to make wise approximations, as to where and how the textual data is placed and used neural networks to hide the data in the embedding process. This shows that the scope of steganography is extended to the field of Deep Learning too.

Several comparative techniques have been considered for understanding image distortion before. In [6], Ching-Chiuan Lin has proven that grayscale images obtained after the decryption process are almost the same as the cover images considered for steganography. We have aimed at dealing with RGB images without compromising on any channel and on any pixel value.

Some of the previous works in the domain, have highlighted the similarity of images considering only a single image format. For example, in [7], the authors

have explicitly discussed that the mean percentage of correct classification is approximately 95.3%.

In [18], the authors have carried out the evaluation metrics part using MSE and PSNR, showing LSB Substitution with some amount of encryption is the better one compared to PVD. Whereas, we have mentioned SSIM which is more robust and accurate than MSE in comparing the pixel intensity values of two images (discussed further in Sect. 5).

3 Preliminaries

3.1 Triple-DES

After 1990 users started using Triple-DES. The reason for using Triple-DES against DES was because of an exhaustive key search that was performed successfully on DES. In those days, users were not ready to switch the cryptographic algorithm because of the cost and time of replacing it; instead, they built a Triple-DES variant of the original DES [11–13].

The Normal key size of the DES algorithm is 56, but in Triple-DES, the length of the key size is increased by three times, i.e., the key size of Triple-DES is 168. This increased key size provides more protection against exhaustive searches. 64 bits of data constitute each block, which is undergone a 3—time encryption using the DES encryption technique.

The architecture shown in Fig. 1 as above first uses a single-DES to encrypt the plaintext block using key K1. The result of the previous step is decrypted with key K2, and finally, again, the plaintext block is encrypted with key K3 to produce ciphertext. The ciphertext is then decrypted by reversing the encryption algorithm, i.e., K3 is used to decrypt, then K2 is used to encrypt, and lastly, K1 is used to decrypt. Although triple-DES systems are significantly safer and more protected than single-DES, they are much slower to operate.

3.2 Advanced Encryption System (AES)

Advanced Encryption Standard (AES) is the most common and commonly used symmetric encryption technique. It is much faster than triple-DES. The DES key size was too tiny and needed to be replaced. When the processing power increased, it was assumed to be susceptible to an exhaustive key search attack. To overcome this weakness, Triple-DES was used; nevertheless, it was sluggish.

Because of these reasons, users switched to AES, and they found it is more efficient than early cryptographic algorithms. AES uses an iterative approach rather than a Feistel cipher approach. The number of iterations in AES is governed by the key

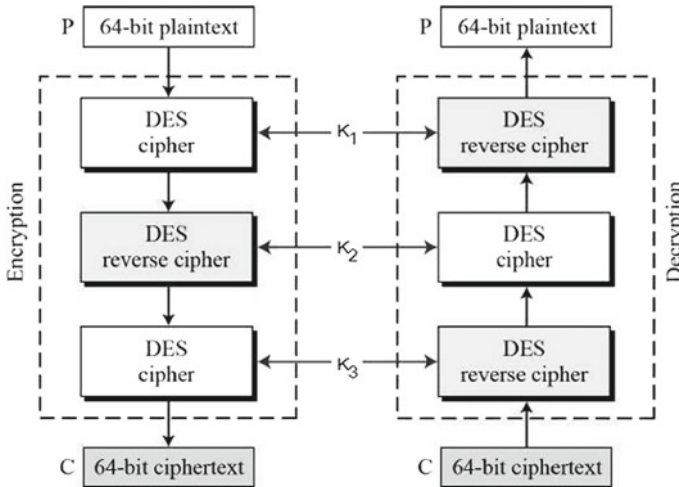


Fig. 1 Structural working of the Triple-DES algorithm

length and is adjustable. AES uses ten rounds for 128-bit keys, 12 rounds for 192-bit keys, and 14 rounds for 256-bit keys. The different 128-bit key round key is used in each round, calculated from the native AES key. AES does all operations in bytes instead of bits. Moreover, it is built on a ‘substitution–permutation network.’

Each round comprises four operations in the encryption of plaintext in the AES algorithm which is shown above in Fig. 2. They are functioning as follows.”

- **Byte Substitution**—By using S-Box provided in the design, the 16 input bytes are substituted.
- **Shift rows**—Each of the matrix’s four rows is shifted to the left in Shift rows.
- **Mix Columns**—Each four-byte column is altered using a specific mathematical algorithm. This function accepts 32 bits from one column as input and replaces them with four new bytes.
- **Add Round Key**—The 16 bytes of the matrix are considered 128 bits, and they are XOR-ed with the 128 bits of the round key. The 128 bits are translated into 16 bytes in regular rounds, and we repeat the process. However, the ciphertext is produced at the last round after adding the round key.

The decryption of ciphertext in the AES algorithm is similar to encryption, but the difference is that the operations in each block are reversed. As with DES, AES [8] security is only assured if adequately implemented and standard key management is employed.

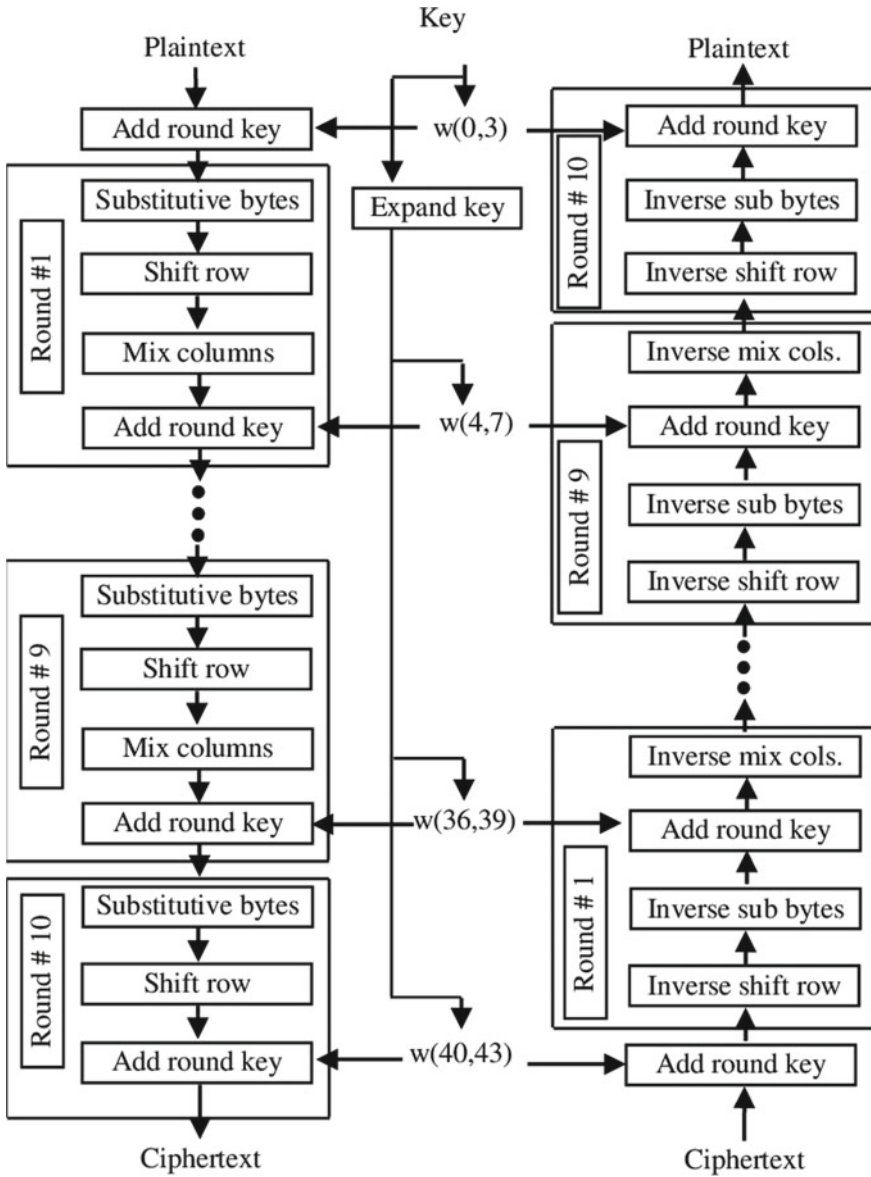


Fig. 2 Structural Working of the AES encryption technique

3.3 Steganography

The technique of concealing a message in a non-secret object without changing its characteristic properties is defined as the art of steganography. Its goal is to hide and defraud the message. It is a sort of covert communication in which messages are concealed through the use of any medium. The key contrast between cryptography and steganography is that cryptography renders the data illegible or hides the meaning, and steganography hides the data's presence. A message can be concealed in several different ways. Some bytes in a file or image are redundant, and they can be substituted with a message without affecting the original message. Different types of steganography exist, Image Steganography, Audio Steganography, Video Steganography, [16, 20], etc.

In our project, we used steganography in digital images. Because digital images arrive in various formats, the algorithms used to process them differ substantially. The approach used in the project is the Least-Significant-bit [9] and [15] techniques. The attacker searches the file for the least essential data and replaces it with the hidden message, usually damaging the code. There are many such techniques, such as Palette based technique in which, using digital photographs as malware carriers, the attackers encrypt the message and then hide it in a broad palette of the cover image. Techniques like Secure Cover Selection in which the carrier image's blocks must be compared to specified malware blocks can also be used. Nevertheless, the advantages of Least-Significant-Bit (LSB) steganographic data embedding are that it is simple to comprehend, simple to execute, and produces stego images [1, 2] with hidden data.

4 Implementation

4.1 Image Steganography

Steganography incorporated within images, in one way, is thought of implementing using a clustering technique, where the pixels are clustered into three regions (R, G, B). Our execution is governed using a novel conversion technique, unlike what is used in [14]. In implementing our project, we used Python as the programming language. In Python, the 'Cryptodomex' module calls certain essential functions for the project like AES, DES, SHA256, etc. Image Encryption and Steganography have been implemented in various forms to analyze different results.

In order to understand steganography, let us consider a message to be hidden inside the image—"Hello." First, we need to convert each character to its respective ASCII value as shown in Fig. 3. Unlike the approach mentioned in [19], where the ASCII of all pixels is converted to a matrix for further encoding, the ASCII value in our case is converted to its respective 8-bit binary number. Each pixel contains three values corresponding to Red, Green, and Blue, respectively. From the image, we take

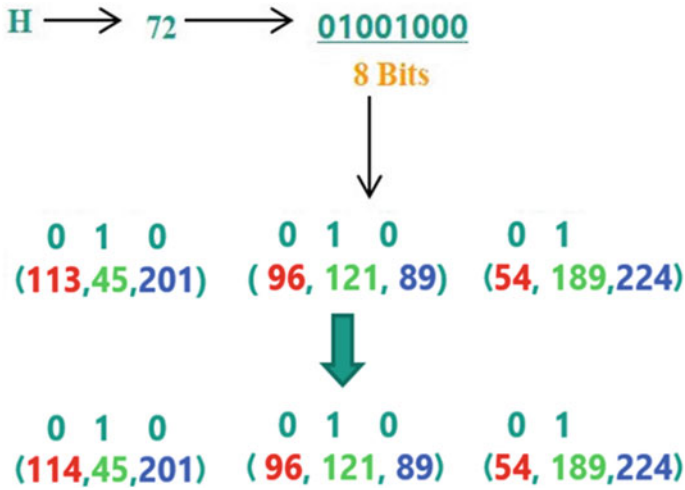


Fig. 3 Encoding the given secret text into the original image to generate a stego image

out three pixels which give nine values. These nine values are modified according to 8-bit binary values by following specific conditions.

If the bit is 0—change its respective pixel intensity value as even and if the bit is 1—change its respective pixel intensity value as odd. If the last pixel intensity value is even, it conveys the message continues, or else if it is odd, it tells that the message is terminating.

To carry out the above process, the user has to enter a password in the encryption part, which is nothing but the key in this process. This particular key is hashed using the SHA—256 algorithm. We adopted the Cipher Block Chaining mode in the AES algorithm to perform encryption in the later stages. An initialization vector (IV) of a certain length is utilized in cipher block chaining. By combining this with a single encryption key, individuals and organizations may securely encrypt and decrypt huge quantities of plaintext. Cipher block chaining is a method for encrypting and decrypting significant plaintext inputs that involve building a cryptographic chain in which each ciphertext [5] block is dependent on the previous one.

To begin a cipher blockchain, XOR is considered the first block of the many, with an IV, which is a one-of-a-kind conversion function with a fixed length, to generate a pseudorandom or random output. The ciphertext block, an encrypted text format that can be decrypted with the correct key, is created by encrypting the XOR output with a cipher key. CBC decryption works in a similar but separate manner. The process does not start with the last ciphertext block, as it does with similar decoding algorithms. In reality, because all inputs are present, it might all happen simultaneously. After combining the second ciphertext block with the cipher key, the output is XOR-ed with the initial ciphertext block to generate the second plaintext block. The IV is substituted with the preceding ciphertext block during the decoding process.

Padding is a critical method to be applied during encryption and decryption. Padding is a cryptographic phrase that refers to various processes that all entail appending data to the start, middle, or conclusion of communication before encryption. This step ensures that our secret message is encrypted and appropriately hidden inside the image in the manner mentioned in the above explanations.

We should always consider all of our design and implementation corner cases. While the receiver may be confident that the data is secure enough while the transmission is taking place, attackers who are clever enough might still be able to retrieve the file and decode the text inside it. Therefore, we have considered Image Encryption a viable step, which is discussed in the latter.

In the above procedure, the algorithms AES and T—DES are resisting the brute-force approach as the key size is high, i.e., 256 and 192-bit keys for AES and T—DES respectively. The length of the encryption key governs whether or not a brute-force attack is practical, with longer keys being significantly more difficult to breach than smaller ones. Applying brute-force methods, AES 256 is essentially impenetrable. Whereas a 56-bit DES key can be broken within a day, given current computing capability, breaking AES will take billions of years. At the time of writing this paper, the fastest supercomputer is the Fugaku supercomputer, which can computer 442 Peta Flops (Flops—floating-point operations per second), unlike the Intel Core i9 which can perform only close to a100 Giga Flops. Even though, this fastest supercomputer itself takes (~)30 trillion trillion trillion years (which is still impossible). It would be unwise for hackers to launch such an attack on a normal PC. However, no encryption system is completely secure.

4.2 *Image Encryption*

Image Encryption is one of the techniques used in digital image protection, and its primary premise is to encrypt the digital content included in a digital image. As a result, the appearance and the original digital image are completely independent encrypted images, preventing direct viewing of the digital image's content. We are precisely replicating this process in Python. Let us now discuss the structure of this part.

At first, the user is requested to enter a password (key for the AES algorithm), which is bound to the image file. The key is then hashed using the SHA—256 algorithm. Key hashing is necessary to provide a more secure and customizable technique for obtaining data. The key is then passed into the digest function. A hash function computes a message digest function, a finite size numeric encapsulation of the contents of a message. Encrypting a message digest can be used to establish a digital signature. Next, the image file is directly passed into the encryption function, which processes the entire data—each pixel value gets encrypted, and the file is transformed into another image. However, directly sending this image to the receiver may not be the most feasible step here, as the attacker might still trap and disrupt the image, even though he cannot retrieve the hidden data.

Therefore, we have converted the encrypted image file to a generic encoded file with an extension '.enc' (known as an enc file), which protects the file against illegal access or aids in configuring the file for a specific purpose of Internet communication. The enc file generated for our stego image [3] contains information that an average human cannot understand (as shown in Fig. 4). You can't open or install ENC files like you can decrypt APK files since they're encrypted using powerful cryptographic algorithms.

This particular file is then sent to the receiver for further proceedings. The above-mentioned is also performed using the Triple-DES algorithm with MD5 as the hashing technique for the key. We have also considered a method where the hidden test is not encrypted using an algorithm and is directly stored in the image. These procedures will help us analyze the results to conclude what a viable technique to consider in all three of the methods mentioned above is.

5 Results and Analysis

The final output obtained after encrypting the image, the receiver gets a file that looks like the below image.

Figure 4. clearly portrays that no unauthorized person (attacker/hacker) will understand this file format, as it consists of characters and symbols that are meaningless.

The proposed approaches—visibility, security, durability, and capacity are all measured using evaluation measures. Since we are comparing the same image before and after steganography, PSNR and SSIM are the best suitable, unlike other evaluation methods. [18]. The metrics most often used are discussed below and the functionality they measure.

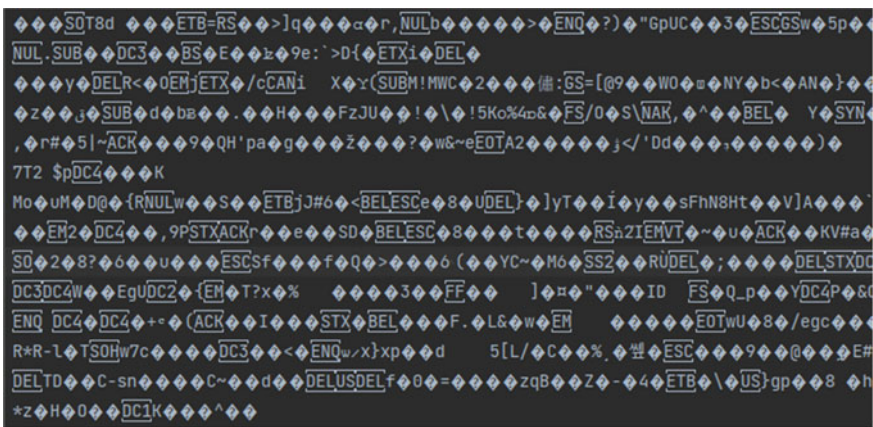


Fig. 4 The generic encoded file (.enc), seen at the receiver side after the final step of encryption

First is the ‘Peak Noise to Signal Ratio,’ which gives us the suggested steganography method’s quality, robustness, and transparency are determined using the Peak Signal to Noise Ratio (PSNR) [6, 6]. PSNR is the ratio of the cover image’s maximum quality representation to the stego image’s maximum quality representation. It is formulated as below:

$$PSNR = 10 \log_{10}((L - 1)^2 / MSE)$$

The subsequent evaluation to be considered is the Structural similarity index (SSIM) between the original image and the stego image. It is used to measure the similarity between two given images. It is a complete reference measurement that takes two images—the cover image (original image) and a processed image—from the same image capture. Typically, the image is compressed after it has been processed. It may be obtained, for example, by storing image data as a JPEG (of whatever quality level) and then reading it back in. It is a value between 0 and 1, where 0 indicates the similarity between 2 images is nil, and 1 shows where the images are precisely the same. This metric for determining quality is dependent on the computation of three primary factors: Brightness (luminance), contrast, and the structural or correlation term. This index is the result of multiplying these three factors together. It is calculated as the below formula:

$$SSIM(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma,$$

where

- *l*—indicates the luminance between two images (used to differentiate the brightness of the two images),
- *c*—indicates the contrast between two images (used to differentiate the brightest and darkest region of two images),
- *s*—displays the structure (used to assess the local luminance pattern between two photos to identify the similarity and dissimilarity of the images),
- α, β, γ —are the positive constants.

SSIM and FSIM are normalized from a representation standpoint, while MSE and PSNR are not. As a result, SSIM and FSIM can be dealt with in a more straightforward manner than MSE and PSNR. PSNR and MSE measure the absolute errors, but SSIM and FSIM are based on perception and saliency. When the noise level rises, the recovery quality of the output image suffers as well. SSIM error map shows the area which is more affected by noise.

Hence it is easy to reconstruct the distorted image. As a result, we may conclude that, from a human visual standpoint, SSIM and FSIM are superior to MSE and PSNR measurements. Figure 5. shows the comparison between the PSNR values for all the 20 sample images when performed steganography using AES and Triple-DES. We can infer that Triple-DES has slightly higher PSNR values than the images where the text is encrypted with AES.

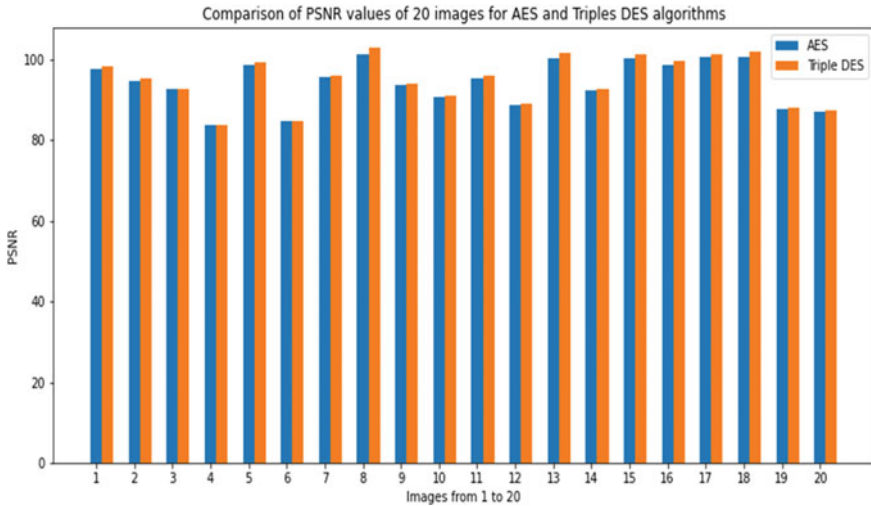


Fig. 5 The comparative depiction of PSNR values for 20 images, when the encryption was done using AES and TDEA algorithms

Figure 6 tells us that the SSIM values for all 20 images are above 0.9 (exactly the same up till the 6th decimal point) for both the encryption techniques, which indicates the similarity between the original and the stego image is almost the same.

The higher the PSNR and SSIM [10] values, the more is the quality and the similarity of the two images, respectively. To better understand our model, we have considered a total of 20 high-quality images, performed the above methods discussed

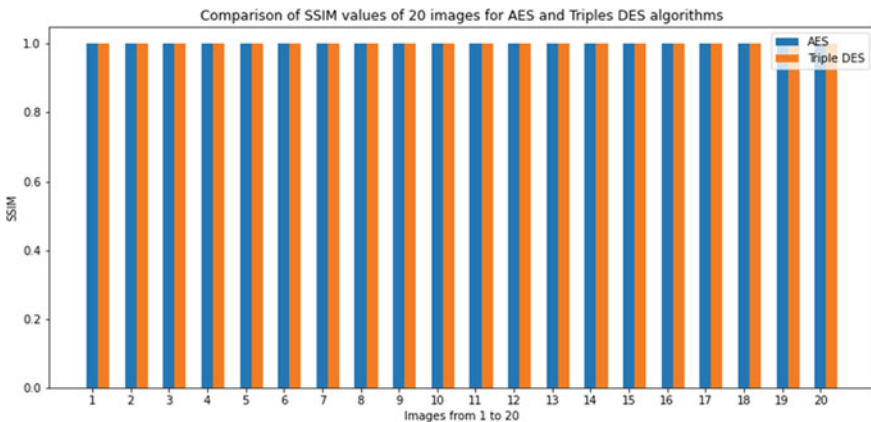


Fig. 6 The comparative depiction of SSIM values for 20 images, when the encryption was done using AES and TDEA algorithms

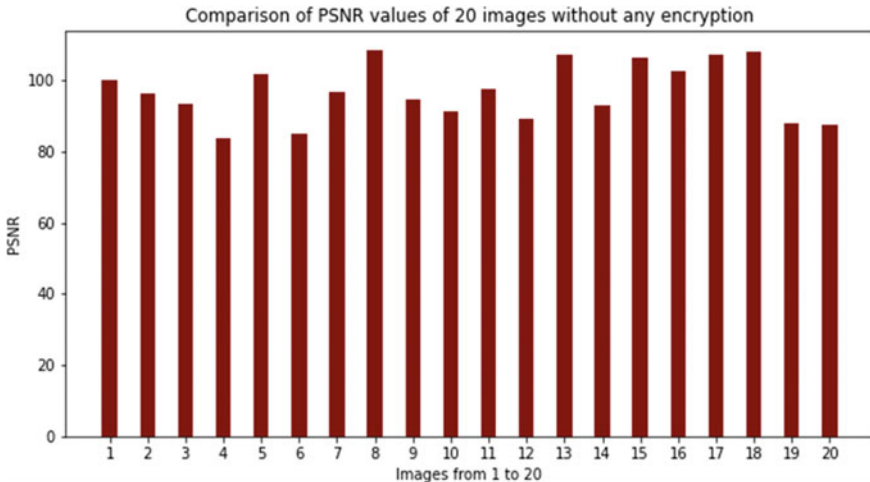


Fig. 7 The comparative depiction of PSNR values for 20 images, when the secret message was not encrypted at all and was directly placed into the image

in the implementation part, and later evaluated them using the PSNR and SSIM metrics.

We have also considered a scenario where the user might want to hide the text without encryption. In such a case, Fig. 7 shows that the PSNR values range from 80 to 100, indicating a high variation if the secret message is not encrypted using any cryptographic algorithm. Keeping in mind that the sender sometimes has to use different image formats, we have also performed the above operations listed in the implementation part for three different image formats—‘JPG / JPEG,’ ‘PNG,’ ‘TIFF’.

In every image format that has been considered, Triple-DES encrypted stego images have a slight edge over the AES ones in terms of their corresponding PSNR values. As far as the SSIM values are concerned, the difference in the case of both the algorithms is very minimal. The above analysis shows us that the model is very robust and flexible, as the user can choose any of the three mentioned file formats, and the difference is not noticeable at all (as shown in Figs. 8, 9).

6 Conclusion and Future Scope

Image encryption models are by far the most considered ones in the secure image processing domain. Image steganography began to mix with the science of computer vision with the introduction of Image steganography, which is also being studied by traditional computer vision researchers. The regions of picture steganography are broadened by combining research fields.

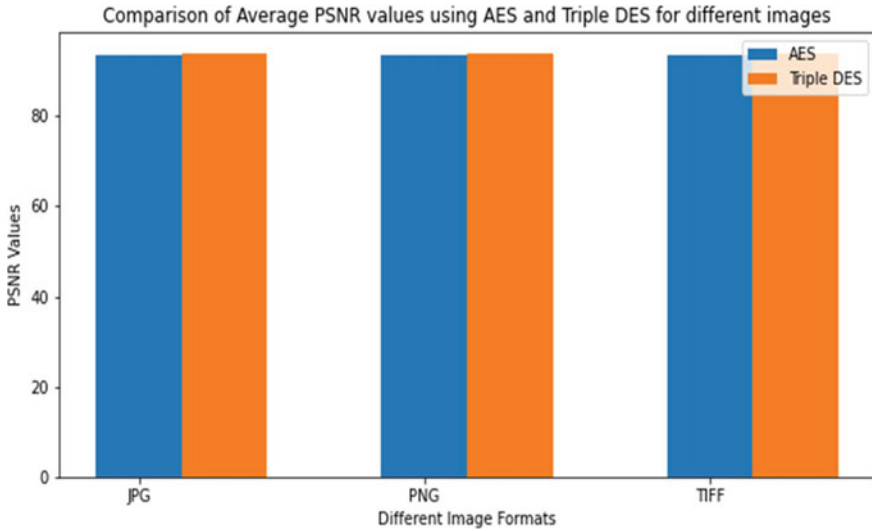


Fig. 8 The comparative depiction of Average PSNR values for different image file formats–JPG/JPEG, PNG, TIFF, when the message is encrypted using AES and TDEA

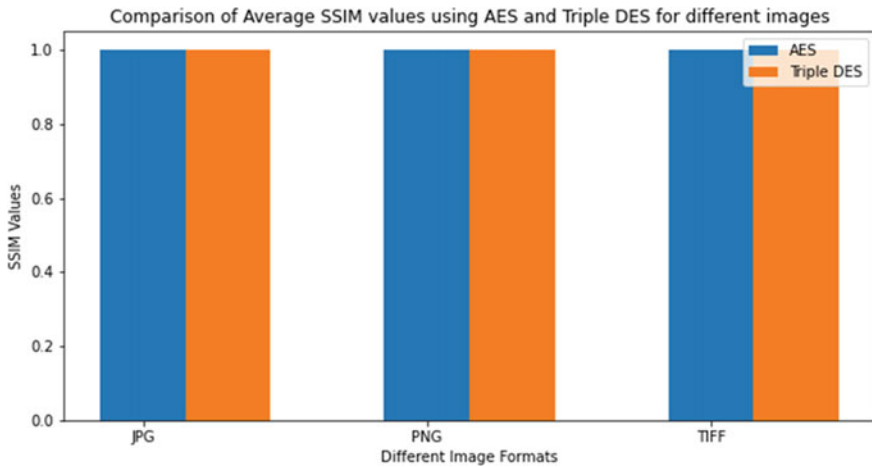


Fig. 9 The comparative depiction of Average SSIM values for different image file formats–JPG/JPEG, PNG, TIFF, when the message is encrypted using AES and TDEA

To the best of our knowledge, we have demonstrated Image Steganography using a unique logic of converting the letters of the text into their 8-bit binary equivalent via their ASCII codes, using two different encryption techniques – AES and Triple-DES, to obtain a stego image which has a similarity index. Also, to ensure an advanced level of security, we have executed Image Encryption on top of this stego image so

that no unauthorized user (attacker/hacker) will be able to decrypt the image and later decode the text. Moreover, the generated file from the image encryption is in '.enc', a standard encoding file for secure file transfers. Any attacker trying to trap this file in between the transmission will not be able to understand what type of a file this is (i.e., a.doc file, pdf file, image file, video file, etc.). We have later evaluated our work using Image Evaluation metrics such as PSNR and SSIM and deduced that Triple-DES is the safest and most viable algorithm for Image Steganography.

Nonetheless, the results obtained from the graphs are of high significance and non-trivial. Also, AES is very close to the results obtained from Triple-DES encrypted stego images. With the help of this project, one can easily send secured messages that too in an image format, with cryptographic algorithms of their choice. For the user to have a real-time experience, we have also created a UI / UX interface to make it look much more appealing.

The proposed project is still under development, i.e., as a part of future enhancement [4], one can extend this paper to implement other encryption techniques such as RSA, Blowfish and Two fish Ciphers, etc. A person trying this paper can also evaluate their images using different metrics such as RMSE [17], DSSIM, FSIM, etc.

References

1. Malathi P, Manoj M, Manoj R, Vaikunth Raghavan, Vinodhini RE, (2017 October) Highly improved DNA based steganography. <https://doi.org/10.1016/j.procs.2017.09.151>
2. Mathivanan P, Balaji Ganesh A, November (2020). QR code-based color image stego-crypto technique using dynamic bit replacement and logistic map. <https://doi.org/10.1016/j.jileo.2020.165838>
3. Chin-Nung Yang, Shen-Chieh Hsu, Cheonshik Kim (2016 November) Improving stego image quality in image interpolation-based data hiding. <https://doi.org/10.1016/j.csi.2016.10.005>
4. Shounak Shastri, Thanikaiselvan V, (2019 March) Dual image reversible data hiding using trinary assignment and center folding strategy with low distortion. <https://doi.org/10.1016/j.jvcir.2019.03.022>
5. Kamal AHM, Mohammad M. Islam (2017 January) Enhancing embedding capacity and stego image quality by employing multi predictors. <https://doi.org/10.1016/j.jjisa.2016.08.005>
6. Ching-Chiu Lin (2011 February) An information hiding scheme with minimal image distortion. <https://doi.org/10.1016/j.csi.2011.02.003>
7. Chetouani A, Beghdadi A, Deriche M (Aug.2010). Statistical modeling of image degradation based on quality metrics. <https://doi.org/10.1109/ICPR.2010.180>
8. Kadhim IJ, Premaratne P, Vial PJ, Halloran B, February. (2019) Comprehensive survey of image steganography: Techniques. Eval Trends Futur Res. <https://doi.org/10.1016/j.neucom.2018.06.075>
9. Abdel OF, Wahab AI, Hussein HFA, Hamed HM, Kelash AAM, Khalaf, 23 March. (2021) Efficient combination of RSA cryptography. lossy and lossless compression steganography techniques to hide data. <https://doi.org/10.1016/j.procs.2021.02.002>
10. Yucel Inan (2018 17 October) Assessment of the image distortion in using various bit lengths of steganographic LSB. <https://doi.org/10.1051/itmconf/20182201026>
11. Alexander Wong, May. 2012. Perceptual Structure Distortion Ratio: An image quality metric based on robust measures of complex phase order. <https://doi.org/10.1109/CRV.2012.15>

12. Ren-Er Y, Zhiwei Z, Shun T, Shilei D (Jan.2014). image steganography combined with DES encryption Pre-processing. <https://doi.org/10.1109/ICMTMA.2014.80>
13. Manoj Kumar Ramaiya, Naveen Hemrajani, Anil Kishore Saxena (2013 Feb.) Security improvisation in image steganography using DES. <https://doi.org/10.1109/IAdCC.2013.6514379>
14. Bhagya Pillai, Mundra Mounika, Pooja J Rao, Padmamala Sriram (2016 Sept.). Image steganography method using K-means clustering and encryption techniques. <https://doi.org/10.1109/ICACCI.2016.7732209>
15. Nidhi Menon, Vaithiyathan (2017 Dec) A survey on image steganography. <https://doi.org/10.1109/TAPENERGY.2017.8397274>
16. Manohar N, Peetla Vijay Kumar (2020 May) Data encryption & decryption using steganography. <https://doi.org/10.1109/ICICCS48265.2020.9120935>
17. S. Sravani R, Ranjith (2021 July) Image steganography for confidential data communication. <https://doi.org/10.1109/ICCCNT51525.2021.9579814>
18. Asha Asok, Poornima Mohan (2019 April) Implementation and comparison of different data hiding techniques in image steganography. <https://doi.org/10.1109/ICOEI.2019.8862750>
19. Anusha M, Bhanu KN, Divyashree D (2020 July). Secured communication of text and audio using image steganography. <https://doi.org/10.1109/ICESC48915.2020.9155715>
20. Preethi P, Prakash G (2021) Secure Fusion of Crypto-Stegano Based Scheme for Satellite Image Application. In: 2021 Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, (pp 1–6). <https://doi.org/10.1109/ASIANCON51346.2021.9544752>

Correction to: QuantumRNG, A Random Number Generator Using One Qubit



Dara Ekanth, Bheemanathy Saketh Chandra, and Meena Belwal

Correction to:
Chapter “QuantumRNG, A Random Number Generator Using One Qubit” in: S. Jain et al. (eds.), *Semantic Intelligence*, Lecture Notes in Electrical Engineering 964,
https://doi.org/10.1007/978-981-19-7126-6_10

In the original version of the book, the following belated correction has been incorporated in chapter “QuantumRNG, A Random Number Generator Using One Qubit”: The chapter author name “Bhemmanathy Saketh Chandra” has been changed to “Bheemanathy Saketh Chandra” in Chapter 10 and Table of Contents.

The correction chapter and the book has been updated with the changes.

The updated version of this chapter can be found at
https://doi.org/10.1007/978-981-19-7126-6_10