COLING

International Conference on Computational Linguistics

Proceedings of the Conference and Workshops

COLING

Volume 29 (2022), No. 12

Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)

The 29th International Conference on Computational Linguistics

October 12 - 17, 2022
Gyeongju, Republic of Korea

ISSN 2951-2093

Introduction

As the number of users and their web-based interaction has increased, incidents of a verbal threats, aggression and related behaviour like trolling, cyberbullying, and hate speech have also increased manifold globally. The reach and extent of the Internet have given such incidents unprecedented power and influence to affect the lives of billions of people. Such incidents of online abuse have not only resulted in mental health and psychological issues for users, but they have manifested in other ways, spanning from deactivating social media accounts to instances of self-harm and suicide.

To mitigate these issues, researchers have begun to explore the use of computational methods for identifying such toxic interactions online. In particular, Natural Language Processing (NLP) and ML-based methods have shown great promise in dealing with such abusive behaviour through early detection of inflammatory content.

In fact, we have observed an explosion of NLP-based research on offensive content in the last few years. This growth has been accompanied by the creation of new venues such as the WOAH and the TRAC workshop series. Community-based competitions, like tasks 5/6 at SemEval-2019, task 12 at SemEval-2020, and task 5/7 at SemEval-2021 have also proven to be extremely popular. In fact, because of the huge community interest, multiple workshops are being held on the topic in a single year. For example, in 2018 ACL hosted both the Abusive Language Online workshop (EMNLP) as well as TRAC-1 (COLING). Both venues achieved healthy participation with 21 and 24 papers, respectively. Interest in the topic has continued to grow since then and given its immense popularity, we are proposing a new edition of the workshop to support the community and further research in this area.

As in the earlier editions, TRAC focuses on the applications of NLP, ML and pragmatic studies on aggression and impoliteness to tackle these issues. As such the workshop also includes shared tasks on ‘Aggression Identification. The task consisted of two sub-tasks - (1) Bias, Threat and Aggression Identification in Context and (2) Generalising across domains - COVID-19. For task 1, the participants were provided with a "thread" of comments with information about the presence of different kinds of biases and threats (viz. gender bias, gendered threat and none, etc) and its discursive relationship to the previous comment as well as the original post (viz. attack, abet, defend, counter-speech and gaslighting). In a series/thread of comments, participants were required to predict the presence of aggression and bias in each comment, possibly making use of the context. In this task, a total dataset of approximately 60k comments (approximately 180k annotation samples) in Meitei, Bangla and Hindi, compiled in the ComMA Project, were provided for training and testing.

Both the workshop and the shared task received a very encouraging response from the community. The proceedings include 4 oral, 3 posters, and 2 system description papers. In addition to this, the workshop also includes 1 Demo to be presented in the workshop.
We would like to thank all the authors for their submissions and members of the Program Committee for their invaluable efforts in reviewing and providing feedback to all the papers. We would also like to thank all the members of the Organising Committee who have helped immensely in various aspects of the organisation of the workshop and the shared task.

Workshop Chairs

Workshop Chairs

Ritesh Kumar, Dr. Bhimrao Ambedkar University, India
Atul Kr. Ojha, University of Galway, Ireland & Panlingua Language Processing LLP, India
Marcos Zampieri, George Mason University, USA
Shervin Malmasi, Amazon Inc., USA
Daniel Kadar, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary

Assistant Organisers

Siddharth Singh, Dr. Bhimrao Ambedkar University, India
Shyam Ratan, Dr. Bhimrao Ambedkar University, India

Shared Task Organising Committee

Shervin Malmasi, Amazon Inc., USA
Siddharth Singh, Dr. Bhimrao Ambedkar University, India
Shyam Ratan, Dr. Bhimrao Ambedkar University, India
Ritesh Kumar, Dr. Bhimrao Ambedkar University, India
Atul Kr. Ojha, University of Galway, Ireland & Panlingua Language Processing LLP, India
Bharathi Raja Chakravarthi, University of Galway

Programme Committee

Atul Kr. Ojha, University of Galway, Ireland & Panlingua Language Processing LLP, India
Bharathi Raja Chakravarthi, University of Galway
Bornini Lahiri, Indian Institute of Technology-Kharagpur, India
Bruno Emanuel Martins, IST and INESC-ID
Cheng-Te Li, National Cheng Kung University, Taiwan
Chuan-Jie Lin, National Taiwan Ocean University, Taiwan
David Jurgens, University of Michigan
Denis Gordeev, The Russian Presidential Academy of National Economy and Public Administration under the President of the Russian Federation
Dennis Tenen, Columbia University, USA
Dhairya Dalal, University of Galway
Els Lefever, LT3, Ghent University, Belgium
Faneva Ramiandrisoa, IRIT
Han Liu, Cardiff University
Hugo Jair Escalante, INAOE, Mexico
Koustava Goswami, University of Galway
Liang-Chih Yu, Yuan Ze University, Taiwan
Lun-Wei Ku, Academia Sinica, Taiwan
Lütfiye Seda Mut Altın, Pompeu Fabra University Mainack Mondal, University of Chicago, USA
Manuel Montes-y-Gómez, INAOE, Mexico
Marco Guerini, Fondazione Bruno Kessler, Trento
Ming-Feng Tsai, National Chengchi University, Taiwan
Monojit Choudhury, Microsoft Turing
Nemanja Djuric, Aurora Innovation
Parth Patwa, Indian Institute of Information Technology, Sri City
Preslav Nakov, Qatar Computing Research Institute, Qatar
Priya Rani, University of Galway

Ritesh Kumar, Dr. B. R. Ambedakar University, India
Roman Klinger, University of Stuttgart, Germany
Ruifeng Xu, Harbin Institute of Technology, China
Saja Tawalbeh, University of Antwerp
Sara E. Garza, Universidad Autónoma de Nuevo León (UANL), Mexico
Shardul Suryawanshi, University of Galway
Shubhanshu Mishra, Twitter Inc.
Valerio Basile, University of Turin
Veronique Hoste, LT3, Ghent University, Belgium
Xavier Tannier, Université Paris-Sud, LIMSI, CNRS, France
Zeerak Waseem, University of Sheffield, UK

Invited Speaker

Valerio Basile, University of Turin, Italy

Valerio Basile is an Assistant Professor at the Computer Science Department of the University of Turin, Italy, member of the Content-centered Computing group and the Hate Speech Monitoring lab. His work spans across several areas such as: formal representations of meaning, linguistic annotation, natural language generation, commonsense knowledge, semantic parsing, sentiment analysis, and hate speech detection, perspectives and bias in supervised machine learning, from data creation to system evaluation. He is currently PI of the project BREAKhateDOWN "Toxic Language Understanding in Online Communication", and among the main proponents of the Perspectivist Data Manifesto: https://pdai.info

Title: The Evaluation of Language Models for Undesirable Language Analysis

Abstract:

In the past five years, the field of Natural Language Processing has seen several important changes and paradigm shifts. Methodologically, large neural language models have taken the spotlight as the new state of the art for most classification (and other kinds of) tasks. At the same time, the focus of research has opened up more and more to the study of pragmatics phenomena in natural language. Among these, toxic, abusive, offensive language, hate speech, and other undesirable phenomena have been subject of the development of specialized models, language resources, and evaluation campaigns.

In this talk, he will give a partial overview of the design and the results of large-scale evaluation efforts, in a multilingual perspective. Quantitative results on such subjective and hard-to-define phenomena should not be taken at a face value. Rather, the quality of benchmarks, and the annotated data behind them, should be carefully analysed. Finally, he will briefly introduce the perspectivist framework and its potential impact on the evaluation of models for undesirable language analysis.

Panelists: Amitava Das (Wipro AI), Stavros Assimakopoulos (University of Malta), Pilar G. Blitvich (University of North Carolina) and Bertie Vidgen (The Alan Turing Institute)

Table of Contents

Conference Program

Monday, October 17, 2022 (GMT+9)

09:00–09:15 Inaugural Session
Chair: Workshop Chairs

09:00–09:15 *Welcome*
Workshop Chairs

09:15–10:30 Q&A Session 1

09:15–09:30 *L3Cube-MahaHate: A Tweet-based Marathi Hate Speech Detection Dataset and BERT Models*
Hrushikesh Patil, Abhishek Velankar and Raviraj Joshi

09:30–09:45 *Which One Is More Toxic? Findings from Jigsaw Rate Severity of Toxic Comments*
Millon Das, Punyajoy Saha and Mithun Das

09:45–10:00 *Can Attention-based Transformers Explain or Interpret Cyberbullying Detection?*
Kanishk Verma, Tijana Milosevic and Brian Davis

10:00–10:15 *Bias, Threat and Aggression Identification Using Machine Learning Techniques on Multilingual Comments*
Kirti Kumari, Shaury Srivastav and Rajiv Ranjan Suman

10:15–10:30 *The Role of Context in Detecting the Target of Hate Speech*
Ilia Markov and Walter Daelemans

10:30–11:00 ***COFEE BREAK***

Monday, October 17, 2022 (GMT+9) (continued)

11:00–12:30 Q&A Session 2

11:00–11:25 *Annotating Targets of Toxic Language at the Span Level*
Baran Barbarestani, Isa Maks and Piek Vossen

11:25–11:50 *Is More Data Better? Re-thinking the Importance of Efficiency in Abusive Language Detection with Transformers-Based Active Learning*
Hannah Kirk, Bertie Vidgen and Scott Hale

11:50–12:15 *A Lightweight Yet Robust Approach to Textual Anomaly Detection*
Leslie Barrett, Robert Kingan, Alexandra Ortan and Madhavan Seshadri

14:00–15:00 Keynote Talk

14:00–15:00 *The Evaluation of Language Models for Undesirable Language Analysis*
Valerio Basile, University of Turin

15:00–16:00 ***COFEE BREAK***

16:00–17:00 Panel Discussion

16:00–17:00 *The Role of Pragmatics in Offensive and Aggressive Language Identification Research*
Amitava Das (Wipro AI), Stavros Assimakopoulos (University of Malta), Pilar G. Blitvich (University of North Carolina) and Bertie Vidgen (The Alan Turing Institute)

Monday, October 17, 2022 (GMT+9) (continued)

17:00–17:50 Q&A Session 3

17:00–17:25 *Detection of Negative Campaign in Israeli Municipal Elections*
Marina Litvak, Natalia Vanetik, Sagiv Talker and Or Machlouf

17:25–17:50 *Hypothesis Engineering for Zero-Shot Hate Speech Detection*
Janis Goldzycher and Gerold Schneider

17:50–18:00 Closing

17:50–18:00 *Vote of Thanks*
Workshop Chairs