WOAH 2020

The Fourth Workshop on Online Abuse and Harms

Proceedings of the Workshop

November 20, 2020 Online

Platinum Sponsors

Gold Sponsors

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-952148-79-8

Introduction

Digital technologies have brought myriad benefits for society, transforming how people connect, communicate and interact with each other. However, they have also enabled harmful and abusive behaviours to reach large audiences and for their negative effects to be amplified, including interpersonal aggression, bullying and hate speech. The negative effects are further compounded as marginalised and vulnerable communities are disproportionately at the risk of receiving abuse. As policymakers, civil society and tech companies devote more resources and effort to tackling online abuse, there is a pressing need for scientific research that critically and rigorously investigates how they are defined, detected, moderated and countered.

Over the last few years there has been a rise in interest in using Natural Language Processing (NLP) to address online abuse at scale. In order to develop robust, long-term technological solutions for this problem, we need perspectives from beyond computer science, including a diverse range of disciplines such as psychology, law, gender studies, communications, and critical race theory. Our goal with the Workshop on Online Abuse and Harms (WOAH, formerly the Workshop on Online Abusive Language, ALW) is to provide a platform to facilitate the interdisciplinary conversations and collaborations that are needed to effectively and ethically address online abuse.

Each year, we choose a theme for our workshop that guides the talks and panel discussions. In previous years we have focused on human content moderators, the policy aspects of tackling online abuse, and the stories and experiences of those who have received large amounts of online abuse. The themes do not limit what original research is presented at the workshop, rather it helps to frame the discussions by providing a particular lens. For this year, we have chosen to focus on Social Bias and Unfairness in Online Abuse Detection. We continue to highlight the need for research that emphasizes the disparate harms that content moderation systems create and propagate.

With content moderation systems being researched and deployed more widely, there is a growing need to critically consider how they unequally impact different communities and drive processes of marginalisation. To this effect, we have expanded the remit of WOAH, introducing a new track for civil society reports to encourage civil society actors to submit and present their work. Moreover, for practical considerations of bias in content moderation, we conceptualised *shared explorations* - a type of shared task that emphasizes the critical investigation of datasets over leaderboards.
To situate the Workshop around the theme of social biases, we have invited speakers to highlight different ways in which content moderation systems can be vehicles of oppression against marginalised people.

In addition, in the interest of increased engagement with the civil society, we organized a satellite session of the workshop at RightsCon 2020, the biggest international conference at the intersection of human rights and digital technologies. Our session consisted of a panel discussion including computer scientists, human rights scholars, and social scientists, and was attended by over 100 human rights scholars and activists from across the world. We include a RightsCon session report in the proceedings to summarize the main themes of insights emerged from the discussion in the satellite session.

During the Workshop we will have a multi-disciplinary panel discussion where experts will debate and contextualize the major issues facing computational analysis of online abuse, with a specific focus on systemic biases that are propagated by content moderation systems. This session will be followed by paper Q&A sessions, facilitating discussions around the research papers described in these proceedings. Due to the virtual nature of this edition of the workshop, we have gathered papers into five thematic panels to allow for more in-depth and rounded discussions.

Continuing the success of the past editions of the workshop, we received 53 submissions. Our submissions come overwhelmingly from academics with 42 submission from academics, 3 from civil society, and 7 from industry. To encourage submissions from social scientists and civil society, we constructed two submission tracks that allowed for unarchived submissions: Extended abstracts and reports. Following a rigorous review process, we selected 24 submissions to be presented at the workshop. These include 3 extended abstracts, 12 long papers, 4 short papers and 2 reports. The authors of all accepted papers are given an opportunity to expand their work into full journal articles, considered for publication in a forthcoming special issue on online abuse and harms in the journal *First Monday*.[1]

The accepted papers deal with a wide array of topics; critically investigating existing approaches to tackling online abuse, interrogating the uses and implications of classification systems, proposing new datasets, models and extending online abuse detection to new languages, contexts and types of abuse. Three of the accepted papers incorporate social science perspectives, a significant improvement compared with the last two iterations of WOAH. The five panels capture the most important focuses of the submissions: Methods for classifying online abuse (including Transformer-based models and new model architectures), Technical challenges in classifying online abuse, Biases in abusive content training datasets, New datasets and resources for tackling online abuse, and Ways of tackling online abuse.

The authors in these proceedings are also geographically diverse, representing work from 10 countries (based on affiliations): Australia, Canada, France, Germany, New Zealand, Norway, Russia, Turkey, United Kingdom, and the United States.

With this, we welcome you to the Fourth Workshop on Online Abuse and Harms and look forward to a day filled with spirited discussion and thought provoking research!

Bertie, Seyi, Vinod, and Zeerak

[1] https://www.workshopononlineabuse.com/cfp/first-monday-special-issue

Organizers:

Seyi Akiwowo, Glitch!
Bertram Vidgen, Alan Turing Institute
Vinodkumar Prabhakaran, Google
Zeerak Waseem, University of Sheffield

Program Committee:

Mark, Alfano, Delft University of Technology (Netherlands)
Naomi, Appelman, University of Amsterdam (Netherlands)
Veronika, Bajt, Peace Institute (Slovenia)
Renata, Barreto, Berkeley Law (United States)
Dan, Bateyko, Georgetown University Law Center (United States)
Peter, Bourgonje, DFKI (Germany)
Andrew, Caines, University of Cambridge (United Kingdom)
Michael, Castelle, University of Warwick (United Kingdom)
Tuhin, Chakrabarty, Columbia University (United States)
Montse, Cuadros, Vicomtech (Spain)
Aron, Culotta, Illinois Institute of Technology (United States)
Thomas, Davidson, Cornell University (United States)
Gretel Liz, De la Peña Sarracèn, Universidad Politècnica de València (Spain)
Nemanja, Djuric, Uber ATG (United States)
Yanai, Elazar, Bar-Ilan University (Israel)
Elisabetta, Fersini, University of Milano-Bicocca (Italy)
Paula, Fortuna, TALN, Pompeu Fabra University (Portugal)
Simona, Frenda, Universitat Politècnica de València (Spain)
Björn, Gambäck, Norwegian University of Science and Technology (Norway)
Maya, Ganesh, Leuphana University (Germany)
Sara E., Garza, FIME-UANL (Mexico)
Ryan, Georgi, University of Washington (United States)
Lee, Gillam, University of Surrey (United Kingdom)
Tonei, Glavinic, Dangerous Speech Project (Spain)
Genevieve, Gorrell, University of Sheffield (United Kingdom)
Erica, Greene, The New York Times (United States)
Marco, Guerini, Fondazione Bruno Kessler (Italy)
Udo, Hahn, Friedrich-Schiller-Universität Jena (Germany)
Alex, Hanna, Google (United States)
Alex, Harris, The Alan Turing Institute (United Kingdom)
Christopher, Homan, Rochester Institute of Technology (United States)
Hossein, Hosseini, Department of Electrical Engineering, University of Washington (United States)
Veronique, Hoste, Ghent University (Belgium)
Ruihong, Huang, Texas A&M University (United States)
Muhammad Okky, Ibrohim, Universitas Indonesia (Indonesia)
Srecko, Joksimovic, University of South Australia (Australia)
Nishant, Kambhatla, Simon Fraser University (Canada)
George, Kennedy, Intel (United States)
Ashiqur, KhudaBukhsh, Carnegie Mellon University (United States)
Ralf, Krestel, Hasso Plattner Institute, University of Potsdam (Germany)

Els, Lefever, LT3, Ghent University (Belgium)
Diana, Maynard, University of Sheffield (United Kingdom)
Smruthi, Mukund, Amazon (United States)
Isar, Nejadgholi, National Research Council Canada (Canada)
Viviana, Patti, University of Turin, Dipartimento di Informatica (Italy)
Umashanthi, Pavalanathan, Twitter (United States)
Matùš, Pikuliak, Kempelen Institute of Intelligent Technologies (Slovakia)
Michal, Ptaszynski, Kitami Institute of Technology (Japan)
Georg, Rehm, DFKI (Germany)
Julian, Risch, Hasso Plattner Institute, University of Potsdam (Germany)
Björn, Ross, University of Edinburgh (United Kingdom)
Paolo, Rosso, Universitat Politècnica de València (Spain)
Niloofar, Safi Samghabadi, University of Houston (United States)
Magnus, Sahlgren, RISE (Sweden)
Christina, Sauper, Facebook (United States)
Alexandra, Schofield, Harvey Mudd College (United States)
Qinlan, Shen, Carnegie Mellon University (United States)
Marian, Simko, Slovak University of Technology in Bratislava (Slovakia)
Vinay, Singh, International Institute of Information Technology, Hyderabad (India)
Sajedul, Talukder, Florida International University (United States)
Zahidur, Talukder, University of Texas at Arlington (United States)
Linnet, Taylor, Tilburg University (Netherlands)
Achint, Thomas, Embibe (India)
Sara, Tonelli, FBK (Italy)
Dimitrios, Tsarapatsanis, University of York (United Kingdom)
Avijit, Vajpayee, Educational Testing Service (United States)
Joris, Van Hoboken, Vrije Universiteit Brussel (Belgium)
Erik, Velldal, University of Oslo (Norway)
Ingmar, Weber, Qatar Computing Research Institute (Qatar)
Lucas, Wright, Cornell University (United States)
Fan, Yang, Nuance Communications (United States)
Seunghyun, Yoon, Seoul National University (Republic of Korea)
Aleš, Završnik, Institute of criminology at the Faculty of Law Ljubljana (Slovenia)
Torsten, Zesch, Language Technology Lab, University of Duisburg-Essen (Germany)
Andrej, Švec, Slido (Slovakia)

Additional Reviewers:

Mareike, Hartmann, University of Copenhagen (Denmark)

Invited Speaker:

André Brock, School of Literature, Media, and Communication, Georgia Tech
Alex Hannah, Google
Maliha Ahmed, Independent Researcher
Maria Y. Rodriguez, University at Buffalo (State University of New York)

Table of Contents

Online Abuse and Human Rights

Regular Contributions

Shared Exploration

Conference Program[1]

November 20, 2020

November 20, 2020

13:00–13:10 ***Opening Remarks***

13:10–15:40 **Keynotes**

13:10–13:55 *Keynote I*
André Brock

13:55–14:40 *Keynote II*
Alex Hanna and Maliha Ahmed

14:40–14:45 ***Break***

14:45–15:30 *Keynote III*
Maria Rodriguez

15:30–15:40 ***Break***

15:40–16:40 *Keynote Panel*
Maliha Ahmed, André Brock, Alex Hanna, Maria Rodriguez

16:40–17:00 ***Break***

[1] All times are given in UTC.

November 20, 2020 (continued)

17:00–17:45 Paper Q & A Panels I:

Panel I: Methods for classifying online abuse

A Novel Methodology for Developing Automatic Harassment Classifiers for Twitter
Ishaan Arora, Julia Guo, Sarah Ita Levitan, Susan McGregor and Julia Hirschberg

Using Transfer-based Language Models to Detect Hateful and Offensive Language Online
Vebjørn Isaksen and Björn Gambäck

Fine-tuning BERT for multi-domain and multi-label incivil language detection
Kadir Bulut Ozler, Kate Kenski, Steve Rains, Yotam Shmargad, Kevin Coe and Steven Bethard

HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language
Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile and Viviana Patti

Abusive Language Detection using Syntactic Dependency Graphs
Kanika Narang and Chris Brew

Panel II: Biases in datasets for abuse

Impact of politically biased data on hate speech classification
Maximilian Wich, Jan Bauer and Georg Groh

Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics
Hala Al Kuwatly, Maximilian Wich and Georg Groh

Investigating Annotator Bias with a Graph-Based Approach
Maximilian Wich, Hala Al Kuwatly and Georg Groh

Reducing Unintended Identity Bias in Russian Hate Speech Detection
Nadezhda Zueva, Madina Kabirova and Pavel Kalaidin

November 20, 2020 (continued)

Investigating Sampling Bias in Abusive Language Detection
Dante Razo and Sandra Kübler

Is your toxicity my toxicity? Understanding the influence of rater identity on perceptions of toxicity
Ian Kivlichan, Olivia Redfield, Rachel Rosen, Raquel Saxe, Nitesh Goyal and Lucy Vasserman

Panel III: Technical challenges in classifying online abuse

Attending the Emotions to Detect Online Abusive Language
Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar and Thamar Solorio

Enhancing the Identification of Cyberbullying through Participant Roles
Gathika Rathnayake, Thushari Atapattu, Mahen Herath, Georgia Zhang and Katrina Falkner

Developing a New Classifier for Automated Identification of Incivility in Social Media
Sam Davidson, Qiusi Sun and Magdalena Wojcieszak

[Findings] Hybrid Emoji-Based Masked Language Models for Zero-Shot Abusive Language Detection
Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli and Serena Villata

Countering hate on social media: Large scale classification of hate and counter speech
Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne and Mirta Galesic

17:45–18:00 ***Break***

November 20, 2020 (continued)

18:00–18:45 Paper Q & A Panels II:

Panel IV: Ways of tackling online abuse

Moderating Our (Dis)Content: Renewing the Regulatory Approach
Claire Pershan

Investigating takedowns of abuse on Twitter
Rosalie Gillett, Nicolas Suzor, Jean Burgess, Bridget Harris and Molly Dragiewicz

Six Attributes of Unhealthy Conversations
Ilan Price, Jordan Gifford-Moore, Jory Fleming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon and Jeffrey Sorensen

Free Expression By Design: Improving In-Platform Features & Third-Party Tools to Tackle Online Abuse
Viktorya Vilk, Elodie Vialle and Matt Bailey

A Unified Taxonomy of Harmful Content
Michele Banko, Brendon MacKeen and Laurie Ray

Panel V: New datasets for abuse

Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage
Jana Kurrek, Haji Mohammad Saleem and Derek Ruths

In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets
Kosisochukwu Madukwe, Xiaoying Gao and Bing Xue

Detecting East Asian Prejudice on Social Media
Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall and Rebekah Tromble

On Cross-Dataset Generalization in Automatic Detection of Online Abuse
Isar Nejadgholi and Svetlana Kiritchenko

[Findings] A little goes a long way: Improving toxic language classification despite data scarcity
Mika Juuti, Tommi Gröndahl, Adrian Flanagan and N. Asokan

November 20, 2020 (continued)

18:45–19:00 ***Break***

19:00–19:20 *The interplay between human rights and content moderation technologies*

Online Abuse and Human Rights: WOAH Satellite Session at RightsCon 2020
Vinodkumar Prabhakaran, Zeerak Waseem, Seyi Akiwowo and Bertie Vidgen

19:20–19:30 ***Closing Remarks***