WOAH 2021

The 5th Workshop on Online Abuse and Harms

Proceedings of the Workshop

August 6, 2021
Bangkok, Thailand (online)

Platinum Sponsors

facebook

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 978-1-954085-59-6

Message from the Organisers

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. Already marginalised and vulnerable communities are often disproportionately at risk of receiving such abuse, compounding other social inequalities and injustices. The Workshop on Online Abuse and Harms (WOAH) convenes research into these issues, particularly work that develops, interrogates and applies computational methods for detecting, classifying and modelling online abuse.

Technical disciplines such as machine learning and natural language processing (NLP) have made substantial advances in creating more powerful technologies to stop online abuse. Yet a growing body of work shows the limitations of many automated detection systems for tackling abusive online content, which can be biased, brittle, low performing and simplistic. These issues are magnified by the lack of explainability and transparency. And although WOAH is collocated with ACL and many of our papers are rooted firmly in the field of machine learning, these are not purely engineering challenges, but raise fundamental social questions of fairness and harm. For this reason, we continue to emphasise the need for inter-, cross- and anti- disciplinary work by inviting contributions from a range of fields, including but not limited to: NLP, machine learning, computational social sciences, law, politics, psychology, network analysis, sociology and cultural studies. In this fifth edition of WOAH we direct the conversation at the workshop through our theme: Social Bias and Unfairness in Online Abuse Detection Systems. Continuing the tradition started in WOAH 4, we have invited civil society, in particular individuals and organisations working with women and marginalised communities, to submit reports, case studies, findings, data, and to record their lived experiences through our civil society track. Our hope is that WOAH provides a platform to facilitate the interdisciplinary conversations and collaborations that are needed to effectively and ethically address online abuse.

Speaking to the complex nature of the issue of online abuse, we are pleased to invite Leon Derczynski, currently an Associate Professor at ITU Copenhagen who works on a range of topics in Natural Language Processing; Deb Raji, currently a Research Fellow at Mozilla who researches AI accountability and auditing; Murali Shanmugavelan, currently a researcher at the Centre for Global Media and Communications at SOAS (London) to deliver keynotes. We are grateful to all our speakers for being available, and look forward to the dialogues that they will generate. On the day of WOAH the invited keynote speakers will give talks and then take part in a multi-disciplinary panel discussion to debate our theme and other issues in computational online abuse research. This will be followed by paper Q&A sessions, with facilitated discussions. Due to the virtual nature of this edition of the workshop, we have gathered papers into thematic panels to allow for more in-depth and rounded discussions.

In this edition of the workshop, we introduce our first official Shared Task for fine-grained detection of hateful memes, in recognition of the ever-growing complexity of human communication. Memes and their communicative intent can be understood by humans because we jointly understand the text and pictures. In contrast, most AI systems analyze text and image separately and do not learn a joint representation. This is both inefficient and flawed, and such systems are likely to fail when a non-hateful image is combined with non-hateful text to produce content that is nonetheless still hateful. For AI to detect this sort of hate it must learn to understand content the way that people do: holistically.

Continuing the success of past editions of the workshop, we received 48 submissions. Following a rigorous review process, we selected 24 submissions to be presented at the workshop. These include 13 long papers, 7 short papers, 3 shared-task system descriptions, and 1 extended abstract. The accepted papers cover a wide array of topics: Understanding the dynamics and nature of online abuse; BERTology: transformer-based modelling of online abuse; Datasets and language resources for online abuse; Fairness, bias and understandability of models; Analysing models to improve real-world performance; Resources for non-English languages. We are hugely excited about the discussions which will take place around these works. We are grateful to everyone who submitted their research and to our excellent team of reviewers.

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

Aida, Bertie, Douwe, Lambert, Vinod and Zeerak.

Organizing Committee

Aida Mostafazadeh Davani, University of Southern California
Douwe Kiela, Facebook AI Research
Mathias Lambert, Facebook AI Research
Bertie Vidgen, The Alan Turing Institute
Vinodkumar Prabhakaran, Google Research
Zeerak Waseem, University of Sheffield

Program Committee

Syed Sarfaraz Akhtar, Apple Inc (United States)
Mark Alfano, Macquarie University (Australia)
Pinkesh Badjatiya, International Institute of Information Technology Hyderabad (India)
Su Lin Blodgett, Microsoft Research (United States)
Sravan Bodapati, Amazon (United States)
Andrew Caines, University of Cambridge (United Kingdom)
Tuhin Chakrabarty, Columbia University (United States)
Aron Culotta, Tulane University (United States)
Thomas Davidson, Cornell University (United States)
Lucas Dixon, Google Research (France)
Nemanja Djuric, Aurora Innovation (United States)
Paula Fortuna, "TALN, Pompeu Fabra University" (Portugal)
Lee Gillam, University of Surrey (United Kingdom)
Tonei Glavinic, Dangerous Speech Project (Spain)
Marco Guerini, Fondazione Bruno Kessler (Italy)
Udo Hahn, Friedrich-Schiller-Universität Jena (Germany)
Alex Harris, The Alan Turing Institute (United Kingdom)
Christopher Homan, Rochester Institute of Technology (United States)
Muhammad Okky Ibrohim, Universitas Indonesia (Indonesia)
Srecko Joksimovic, University of South Australia (Australia)
Nishant Kambhatla, Simon Fraser University (Canada)
Brendan Kennedy, University of Southern California (United States)
Ashiqur KhudaBukhsh, Carnegie Mellon University (United States)
Ralf Krestel, "Hasso Plattner Institute, University of Potsdam" (Germany)
Diana Maynard, University of Sheffield (United Kingdom)
Smruthi Mukund, Amazon (United States)
Isar Nejadgholi, National Research Council Canada (Canada)
Shaoliang Nie, Facebook Inc (United States)
Debora Nozza, Bocconi University (Italy)
Viviana Patti, "University of Turin, Dipartimento di Informatica" (Italy)
Matúš Pikuliak, Kempelen Institute of Intelligent Technologies (Slovakia)
Michal Ptaszynski, Kitami Institute of Technology (Japan)
Georg Rehm, DFKI (Germany)
Julian Risch, deepset.ai (Germany)
Björn Ross, University of Edinburgh (United Kingdom)
Paul Röttger, University of Oxford (United Kingdom)
Niloofar Safi Samghabadi, Expedia Inc. (United States)
Qinlan Shen, Carnegie Mellon University (United States)
Jeffrey Sorensen, Google Jigsaw (United States)

Laila Sprejer, The Alan Turing Institute (United Kingdom)
Sajedul Talukder, Southern Illinois University (United States)
Linnet Taylor, Tilburg University (Netherlands)
Tristan Thrush, Facebook AI Research (FAIR) (United States)
Sara Tonelli, FBK (Italy)
Dimitrios Tsarapatsanis, University of York (United Kingdom)
Avijit Vajpayee, Amazon (United States)
Joris Van Hoboken, Vrije Universiteit Brussel and University of Amsterdam (Belgium)
Ingmar Weber, Qatar Computing Research Institute (Qatar)
Jing Xu, Facebook AI (United States)
Seunghyun Yoon, Adobe Research (United States)
Aleš Završnik, Institute of criminology at the Faculty of Law Ljubljana (Slovenia)
Torsten Zesch, "Language Technology Lab, University of Duisburg-Essen" (Germany)

Table of Contents

Conference Program

August 6, 2021

August 6, 2021

15:00–15:10 ***Opening Remarks***

15:10–15:40 **Keynote Session I**

15:10–15:55 *Keynote I*
Leon Derczynski

15:55–16:40 *Keynote II*
Murali Shanmugavelan

16:40–16:45 ***Break***

16:45–18:10 **Paper Presentations**

16:45–17:10 ***1-Minute Paper Storm***

17:10–17:40 **Paper Q & A Panels I**

17:10–17:40 ***BERTology: transformer-based modelling of online abuse***

Exploiting Auxiliary Data for Offensive Language Detection with Bidirectional Transformers
Sumer Singh and Sheng Li

Modeling Profanity and Hate Speech in Social Media with Semantic Subspaces
Vanessa Hahn, Dana Ruiter, Thomas Kleinbauer and Dietrich Klakow

HateBERT: Retraining BERT for Abusive Language Detection in English
Tommaso Caselli, Valerio Basile, Jelena Mitrović and Michael Granitzer

August 6, 2021 (continued)

[Findings] Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech
Wanzheng Zhu, Suma Bhat

17:10–17:40 ***Analysing models to improve real-world performance***

Multi-Annotator Modeling to Encode Diverse Perspectives in Hate Speech Annotations
Aida Mostafazadeh Davani, Mark Díaz and Vinodkumar Prabhakaran

Memes in the Wild: Assessing the Generalizability of the Hateful Memes Challenge Dataset
Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski and Yuki M Asano

Measuring and Improving Model-Moderator Collaboration using Uncertainty Estimation
Ian Kivlichan, Zi Lin, Jeremiah Liu and Lucy Vasserman

[Findings] Detecting Harmful Memes and Their Targets
Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, Tanmoy Chakraborty

[Findings] Survival text regression for time-to-event prediction in conversations
Christine De Kock, Andreas Vlachos

17:40–18:10 ***Resources for non-English languages***

DALC: the Dutch Abusive Language Corpus
Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman and Malvina Nissim

Offensive Language Detection in Nepali Social Media
Nobal B. Niraula, Saurab Dulal and Diwa Koirala

MIN_PT: An European Portuguese Lexicon for Minorities Related Terms
Paula Fortuna, Vanessa Cortez, Miguel Sozinho Ramalho and Laura Pérez-Mayos

17:40–18:10 Paper Q & A Panels II

17:40–18:10 *Fairness, bias and understandability of models*

Fine-Grained Fairness Analysis of Abusive Language Detection Systems with CheckList
Marta Marchiori Manerba and Sara Tonelli

Improving Counterfactual Generation for Fair Hate Speech Detection
Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren and Morteza Dehghani

Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon
Samira Zad, Joshuan Jimenez and Mark Finlayson

Mitigating Biases in Toxic Language Detection through Invariant Rationalization
Yung-Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung-yi Lee, Yun-Nung Chen and Shang-Wen Li

Fine-grained Classification of Political Bias in German News: A Data Set and Initial Experiments
Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider and Georg Rehm

17:40–18:10 *Datasets and language resources for online abuse*

Jibes & Delights: A Dataset of Targeted Insults and Compliments to Tackle Online Abuse
Ravsimar Sodhi, Kartikey Pant and Radhika Mamidi

Context Sensitivity Estimation in Toxicity Detection
Alexandros Xenos, John Pavlopoulos and Ion Androutsopoulos

A Large-Scale English Multi-Label Twitter Dataset for Cyberbullying and Online Abuse Detection
Semiu Salawu, Jo Lumsden and Yulan He

Toxic Comment Collection: Making More Than 30 Datasets Easily Accessible in One Unified Format
Julian Risch, Philipp Schmidt and Ralf Krestel

August 6, 2021 (continued)

[Findings] CONDA: a CONtextual Dual-Annotated dataset for in-game toxicity understanding and detection
Henry Weld, Guanghao Huang, Jean Lee, Tongshu Zhang, Kunze Wang, Xinghong Guo, Siqu Long, Josiah Poon, Soyeon Caren Han

17:40–18:10 ***Understanding the dynamics and nature of online abuse***

When the Echo Chamber Shatters: Examining the Use of Community-Specific Language Post-Subreddit Ban
Milo Trujillo, Sam Rosenblatt, Guillermo de Anda Jáuregui, Emily Moog, Briane Paul V. Samson, Laurent Hébert-Dufresne and Allison M. Roth

Targets and Aspects in Social Media Hate Speech
Alexander Shvets, Paula Fortuna, Juan Soler and Leo Wanner

Abusive Language on Social Media Through the Legal Looking Glass
Thales Bertaglia, Andreea Grigoriu, Michel Dumontier and Gijs van Dijck

18:10–18:20 ***Break***

18:20–19:00 ***Multi-Word Expressions and Online Abuse Panel***

19:00–19:15 ***Break***

19:15–19:45 **Keynote Session II**

19:15–20:00 *Keynote III*
Deb Raji

20:00–20:45 *Keynote Panel*
Deb Raji, Murali Shanmugavelan, Leon Derczynski

20:45–21:00 ***Break***

August 6, 2021 (continued)

21:00–21:45 Shared Task Session

Findings of the WOAH 5 Shared Task on Fine Grained Hateful Memes Detection
Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen and Zeerak Waseem

VL-BERT+: Detecting Protected Groups in Hateful Multimodal Memes
Piush Aggarwal, Michelle Espranita Liman, Darina Gold and Torsten Zesch

Racist or Sexist Meme? Classifying Memes beyond Hateful
Haris Bin Zia, Ignacio Castro and Gareth Tyson

Multimodal or Text? Retrieval or BERT? Benchmarking Classifiers for the Shared Task on Hateful Memes
Vasiliki Kougia and John Pavlopoulos

21:45–22:00 *Closing Remarks*