NAACL 2024

The 8th Workshop on Online Abuse and Harms (WOAH)

Proceedings of the Workshop

June 20, 2024

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-105-6

Introduction

Digital technologies have brought many benefits for society, transforming how people connect, communicate and interact with each other. However, they have also enabled abusive and harmful content such as hate speech and harassment to reach large audiences, and for their negative effects to be amplified. The sheer amount of content shared online means that abuse and harm can only be tackled at scale with the help of computational tools. However, detecting and moderating online abuse and harms is a difficult task, with many technical, social, legal and ethical challenges. The Workshop on Online Harms and Abuse (WOAH) is the leading workshop dedicated to research addressing these challenges.

WOAH invites paper submissions from a wide range of fields, including natural language processing, machine learning, computational social sciences, law, politics, psychology, sociology and cultural studies. We explicitly encourage interdisciplinary submissions, technical as well as non-technical submissions, and submissions that focus on under-resourced languages. We also invite non-archival submissions for in progress work and reports from civil society to facilitate a meeting space between academic researchers and civil society.

This year marks the eighth edition of WOAH, which is co-located with NAACL 2024 in Mexico City, Mexico. The special theme for this year's edition is "**online harms in the age of large language models**". Highly capable large language models (LLMs) are now widely deployed and easily accessible by millions across the globe. Without proper safeguards, these LLMs will readily follow malicious instructions and generate toxic content. Even the safest LLMs can be exploited by bad actors for harmful purposes. With this theme, we invite submissions that explore the implications of LLMs for the creation, dissemination and detection of harmful online content. We are interested in how to stop LLMs from following malicious instructions and generating toxic content, but also how they could be used to improve content moderation and enable countermeasures like personalised counterspeech.

We received 56 submissions, of which 33 were accepted for presentation at the workshop. These papers will be presented at an in-person poster session on the day of the workshop. Authors who are unable to attend in person will instead give a virtual lightning talk describing their work. The workshop day will also include keynote talks from Alicia Parrish (Google), Yacine Jernite (Hugging Face), Seraphina Goldfarb-Tarrant (Cohere), Apostol Vassilev (NIST), and Lama Ahmad (OpenAI). Finally, we will close the day by inviting the keynote speakers to participate in a panel discussion on this year's special theme.

We thank all our participants and reviewers for their work, and our sponsors for their support. We hope you enjoy this year's WOAH and the research published in these proceedings.

Paul, Yi-Ling, Debora, Aida, Agostina, Flor, and Zeerak

Sponsors

WOAH is grateful for support from the following sponsors:

Diamond Tier

OpenAI

Gold Tier

JIGSAW

Google

Organizing Committee

Workshop Organiser

Paul Röttger, Bocconi University
Yi-Ling Chung, The Alan Turing Institute
Aida Mostafazadeh Davani, Google Research
Debora Nozza, Bocconi University
Flor Miriam Plaza-del-Arco, Bocconi University
Zeerak Talat, Mohamed bin Zayed University of Artificial Intelligence

Program Committee

Chairs

Agostina Calabrese, The University of Edinburgh
Yi-Ling Chung, The Alan Turing Institute
Aida Mostafazadeh Davani, Google Research
Debora Nozza, Bocconi University
Flor Miriam Plaza-del-Arco, Bocconi University
Paul Röttger, University of Oxford
Zeerak Talat, Mohamed bin Zayed University of Artificial Intelligence

Program Committee

Gavin Abercrombie, Heriot Watt University
Prabhat Agarwal, Pinterest
Syed Sarfaraz Akhtar, Apple Inc
Jisun An, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington
Ion Androutsopoulos, Athens University of Economics and Business
Naomi Appelman, University of Amsterdam
Hiromi Arai, RIKEN AIP
Thushari Atapattu, University of Adelaide
Giuseppe Attanasio, Bocconi University
Nikolay Babakov, Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela
Murali Raghu Babu Balusu, Georgia Institute of Technology
Francesco Barbieri, Snap Inc.
Renata Barreto, Berkeley Law
Thales Bertaglia, Maastricht University
Vishal Bhalla, Fourie
Helena Bonaldi, Fondazione Bruno Kessler
Peter Bourgonje, Saarland University
Noah Broestl, University of Oxford, Google Research
Ana-Maria Bucur, Interdisciplinary School of Doctoral Studies
Tommaso Caselli, Rijksuniversiteit Groningen
Amanda Cercas Curry, Bocconi University
Canyu Chen, Illinois Institute of Technology
Corinne David, Emakia
Ona De Gibert, University of Helsinki
Pieter Delobelle, KU Leuven, Department of Computer Science
Daryna Dementieva, Technical University of Munich
Kelly Dennis, University of Connecticut
Athiya Deviyani, Carnegie Mellon University
Mark Diaz, Google
Nemanja Djuric, Aurora Innovation
Tj Elmas, University of Edinburgh
Fatma Elsafoury, Fraunhofer research institute
Micha Elsner, The Ohio State University
Hugo Jair Escalante, INAOE

Elisabetta Fersini, University of Milano-Bicocca
Komal Florio, University of Torino
Simona Frenda, Università degli Studi di Torino
Zee Fryer, Google
Jay Gala, AI4Bharat (IIT Madras)
Bj&ouml;rn Gamb&auml;ck, Norwegian University of Science and Technology
Deep Gandhi, University of Alberta
Achyutarama Ganti, Oakland University
Joshua Garland, Arizona State University
Shlok Gilda, University of Florida
Lee Gillam, University of Surrey
Tonei Glavinic, Dangerous Speech Project
Jen Golbeck, University of Maryland
Darina Gold, Fraunhofer IIS
Janis Goldzycher, University of Zurich
Julia Guo, Columbia University
Udo Hahn, Friedrich-Schiller-Universitaet Jena
Alex Hanna, Google
Niclas Hertzberg, AI Sweden
Muhammad Okky Ibrohim, University of Turin
Tim Isbister, AI Sweden
Alvi Md. Ishmam, PhD student
Abraham Israeli, Ben Gurion University of the Negev
Abhinav Jain, amazon.com
Srecko Joksimovic, University of South Australia
Prashant Kapil, Indian Institute of Technology
Mohammad Aflah Khan, IIIT Delhi
Urja Khurana, Vrije Universiteit Amsterdam
Mamoru Komachi, Hitotsubashi University
Vasiliki Kougia, University of Vienna
Gokul Karthik Kumar, Technology Innovation Institute
Jana Kurrek, McGill University
Sandra Kübler, Indiana University
Lucy Lin, Spotify
Yunfei Long, University of Essex
Tanjim Mahmud, Kitami Institute of Technology, Japan
Nina Markl, University of Essex
Antonis Maronikolakis, Ludwig-Maximillians-University of Munich
Michele Mastromattei, University of Rome Tor Vergata
Sarah Masud, LCS2, IIITD
Puneet Mathur, University of Maryland College Park
Diana Maynard, University of Sheffield
Susan Mcgregor, Columbia University
Andreea Moldovan, University of Bucharest
Mainak Mondal, Institute of Engineering and Management
Angeliki Monnier, Université de Lorraine
Manuel Montes, INAOE
Smruthi Mukund, Amazon
Isar Nejadgholi, National Research Council Canada
Shaoliang Nie, Meta Inc
Brahmani Nutakki, Saarland University

Ali Omrani, University of Southern California
Kartikey Pant, Salesforce
Viviana Patti, University of Turin, Dipartimento di Informatica
Parth Patwa, University of California Los Angeles
Matúš Pikuliak, Kempelen Institute of Intelligent Technologies
Vinodkumar Prabhakaran, Google
Michal Ptaszynski, Kitami Institute of Technology
Yusu Qian, Apple
Krithika Ramesh, Microsoft Research India
Manikandan Ravikiran, Hitachi India R&D
Georg Rehm, DFKI
Bjorn Ross, University of Edinburgh
Paolo Rosso, Universitat Politècnica de València
Nazanin Sabri, University of California San Diego
Haji Mohammad Saleem, McGill University
Salim Sazzed, Old Dominion University
Tyler Schnoebelen, Decoded AI
Mina Schütz, Austrian Institute of Technology GmbH
Haitham Seelawi, Adarga ltd.
Nishant Shah, ArtEZ University of the Arts
Qinlan Shen, Oracle
Jeffrey Sorensen, Google Jigsaw
Ankit Srivastava, OryxLabs
Vivian Stamou, Institute for Language and Speech Processing
Nicolas Suzor, Queensland University of Technology
Kejsi Take, New York University
Zahidur Talukder, University of Texas at Arlington
Sajedul Talukder, Southern Illinois University
Joel Tetreault, Dataminr
Zuoyu Tian, Indiana University
Sara Tonelli, FBK
Dimitrios Tsarapatsanis, University of York
Avijit Vajpayee, Amazon
María Estrella Vallecillo Rodríguez, Universidad de Jaén
Francielle Vargas, University of São Paulo
Vaibhav Varshney, TCS Research
Elodie Vialle, Berkman Klein Center at Harvard / PEN America
Serena Villata, Université Côte d'Azur, CNRS, Inria, I3S
Piek Vossen, Vrije Universiteit Amsterdam
Ruyuan Wan, University of Notre Dame
Ingmar Weber, Saarland University
Michael Wiegand, Alpen-Adria-Universitaet Klagenfurt
Zach Wood-Doughty, Northwestern University
Yi Zheng, University of Edinburgh

Table of Contents

Program

Thursday, June 20, 2024

09:00 - 09:15 *Opening Remarks*

09:15 - 09:45 *Invited Talk 1 - Alicia Parrish*

09:45 - 10:15 *Invited Talk 2 - Yacine Jernite*

10:15 - 10:30 *Mini Break*

10:30 - 11:00 *Invited Talk 3 - Apostol Vassilev*

11:00 - 12:30 *In-Person Poster Session*

Investigating radicalisation indicators in online extremist communities
Christine De Kock and Eduard Hovy

Detection of Conspiracy Theories Beyond Keyword Bias in German-Language Telegram Using Large Language Models
Milena Pustet, Elisabeth Steffen and Helena Mihaljevic

EkoHate: Abusive Language and Hate Speech Detection for Code-switched Political Discussions on Nigerian Twitter
Comfort Ilevbare, Jesujoba Alabi, David Ifeoluwa Adelani, Firdous Bakare, Oluwatoyin Abiola and Oluwaseyi Adeyemo

A Study of the Class Imbalance Problem in Abusive Language Detection
Yaqi Zhang, Viktor Hangya and Alexander Fraser

HausaHate: An Expert Annotated Corpus for Hausa Hate Speech Detection
Francielle Vargas, Samuel Guimarães, Shamsuddeen Hassan Muhammad, Diego Alves, Ibrahim Said Ahmad, Idris Abdulmumin, Diallo Mohamed, Thiago Pardo and Fabrício Benevenuto

VIDA: The Visual Incel Data Archive. A Theory-oriented Annotated Dataset To Enhance Hate Detection Through Visual Culture
Selenia Anastasi, Florian Schneider, Chris Biemann and Tim Fischer

Does Prompt Engineering Matter for LLM-based Toxicity and Rumor Stance Detection? Evidence from a Large-scale Experiment
Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn and Libby Hemphill

Towards a Unified Framework for Adaptable Problematic Content Detection via Continual Learning
Ali Omrani, Alireza Salkhordeh Ziabari, Preni Golazizian, Jeffrey Sorensen and Morteza Dehghani

From Linguistics to Practice: a Case Study of Offensive Language Taxonomy in Hebrew
Chaya Liebeskind, Marina Litvak and Natalia Vanetik

Estimating the Emotion of Disgust in Greek Parliament Records
Vanessa Lislevand, John Pavlopoulos, Panos Louridas and Konstantina Dritsa

Simple LLM based Approach to Counter Algospeak
Jan Fillies and Adrian Paschke

Harnessing Personalization Methods to Identify and Predict Unreliable Information Spreader Behavior
Shaina Ashraf, Fabio Gruschka, Lucie Flek and Charles Welch

Robust Safety Classifier Against Jailbreaking Attacks: Adversarial Prompt Shield
Jinhwa Kim, Ali Derakhshan and Ian Harris

Improving aggressiveness detection using a data augmentation technique based on a Diffusion Language Model
Antonio Reyes-Ramírez, Mario Aragón, Fernando Sánchez-Vega and Adrian López-Monroy

The Mexican Gayze: A Computational Analysis of the Attitudes towards the LGBT+ Population in Mexico on Social Media Across a Decade
Scott Andersen, Segio-Luis Ojeda-Trueba, Juan Vásquez and Gemma Bel-Enguix

X-posing Free Speech: Examining the Impact of Moderation Relaxation on Online Social Networks
Arvindh Arun, Saurav Chhatani, Jisun An and Ponnurangam Kumaraguru

The Uli Dataset: An Exercise in Experience Led Annotation of oGBV
Arnav Arora, Maha Jinadoss, Cheshta Arora, Denny George, Brindaalakshmi , Haseena Khan, Kirti Rawat, Div , Ritash and Seema Mathur

Towards Interpretable Hate Speech Detection using Large Language Model-extracted Rationales
Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth and Huan Liu

Thursday, June 20, 2024 (continued)

A Bayesian Quantification of Aporophobia and the Aggravating Effect of Low–Wealth Contexts on Stigmatization
Ryan Brate, Marieke Van Erp and Antal Van Den Bosch

Toxicity Classification in Ukrainian
Daryna Dementieva, Valeriia Khylenko, Nikolay Babakov and Georg Groh

A Strategy Labelled Dataset of Counterspeech
Aashima Poudhar, Ioannis Konstas and Gavin Abercrombie

AGORA: a Language Model for Safe Speech-to-Text Conversion
Victor Cruz and Laurence Liang

Improving Covert Toxicity Detection by Retrieving and Generating References
Dong-Ho Lee, Hyundong Cho, Woojeong Jin, Jihyung Moon, Sungjoon Park, Paul Röttger, Jay Pujara and Roy Ka-wei Lee

Subjective Isms? On the Danger of Conflating Hate and Offence in Abusive Language Detection
Amanda Cercas Curry, Gavin Abercrombie and Zeerak Talat

From Languages to Geographies: Towards Evaluating Cultural Bias in Hate Speech Datasets
Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott Hale and Paul Röttger

SGHateCheck: Functional Tests for Detecting Hate Speech in Low-Resource Languages of Singapore
Ri Chi Ng, Nirmalendu Prakash, Ming Shan Hee, Kenny Tsu Wei Choo and Roy Ka-wei Lee

[Findings] Tokenization Matters: Navigating Data-Scarce Tokenization for Gender Inclusive Language Technologies
Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter and Rahul Gupta

[Main Conference] An Interactive Framework for Profiling News Media Sources
Nikhil Mehta and Dan Goldwasser

[Main Conference] MISGENDERMENDER: A Community-Informed Approach to Interventions for Misgendering
Tamanna Hossain, Sunipa Dev and Sameer Singh

[Main Conference] Exploring Cross-Cultural Differences in English Hate Speech Annotations: From Dataset Construction to Analysis
Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collado, Juho Kim and Alice Oh

12:30 - 13:45 *Lunch Break*

13:45 - 14:15 *Invited Talk 4 - Seraphina Goldfarb-Tarrant*

14:15 - 14:30 *Outstanding Paper Talks*

14:30 - 15:15 *Lightning Talks for Remote Attendants*

Adversarial Nibbler - A novel crowdsourcing procedure for detecting harmful content in t2i models
Jessica Quaye, Alicia Parrish, Oana Inel, Charvi Rastogi, Hannah Kirk, Minsuk Kahng, Erin Van Liemt, Max Bartolo, Jess Tsang and Justin White

Does Prompt Engineering Matter for LLM-based Toxicity and Rumor Stance Detection? Evidence from a Large-scale Experiment
Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn and Libby Hemphill

Introducing the Public Protection Data Programme
Samantha Lundrigan, Timothy Mcsweeney and Tabossan Sedighi

Comparing LLM ratings of conversational safety with human annotators
Rajiv Movva, Pang Wei Koh and Emma Pierson

Visual and Textual Narrative Analysis of the anti-femicide Movement in Mexico
Laura Dozal

Web Retrieval Agents for Evidence-Based Misinformation Detection
Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Mauricio Rivera, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany and Kellin Pelrine

AGORA: a Language Model for Safe Speech-to-Text Conversion
Victor Cruz and Laurence Liang

Thursday, June 20, 2024 (continued)

	AustroTox: A Dataset for Target-Based Austrian German and English Offensive Language Detection Pia Pachinger, Janis Goldzycher, Anna Maria Planitzer, Wojciech Kusa and Allan Hanbury
15:15 - 15:45	*Invited Talk 5 - Lama Ahmad*
15:45 - 16:30	*Coffee Break*
16:30 - 17:30	*Panel Discussion*
17:30 - 17:40	*Closing Remarks*