ACL 2025

The 9th Workshop on Online Abuse and Harms (WOAH)

Proceedings of the Workshop

August 1, 2025

The ACL organizers gratefully acknowledge the support from the following sponsors.

Platinium Level

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 significant benefits to society, transforming how people connect, communicate, and interact. However, these same technologies have also enabled the dissemination and amplification of abusive and harmful content, such as hate speech, harassment, and misinformation. Given the increased volume of content shared online, addressing abuse and harm at scale requires the use of computational tools. Nevertheless, detecting and moderating online abuse remains a complex task, which encompasses technical, social, legal, and ethical challenges. **The Workshop on Online Abuse and Harms (WOAH)** is the leading venue dedicated to addressing these challenges through interdisciplinary research and dialogue.

WOAH invites contributions from a broad range of fields, including natural language processing, machine learning, computational social science, law, political science, psychology, sociology, and cultural studies. We explicitly encourage interdisciplinary and cross-sectoral submissions, including both technical and non-technical work, as well as research focusing on under-resourced languages and marginalized communities. We also provide space for non-archival submissions and reports from civil society organizations to foster collaboration between academic researchers and practitioners working on the front lines of these issues.

The 9th edition of WOAH will take place on **August 1st, 2025**, as part of **ACL 2025 in Vienna, Austria**. The special theme for this edition is **Harms Beyond Hate Speech.** With this theme, we aim to broaden the conversation around online harms by exploring the complex and often overlooked ways in which harm is mediated through technology. This includes forms of technologically mediated inauthentic behavior, the role of digital systems in reshaping perceptions and influencing public discourse, and the risks these dynamics pose in inciting discrimination, hostility, violence, and even atrocities such as genocide. Additionally, this theme highlights the diversity of affected targets, calling attention to the ethical and methodological challenges that arise when developing computational interventions for such harms. We particularly encouraged contributions addressing critical topics such as child sexual abuse material, radicalization, misinformation, platform policies, security, and the political implications of computational approaches.

This year's program features a selection of high-quality papers, presented through poster sessions and lightning talks, alongside keynote presentations from distinguished researchers and practitioners in the field. We are also introducing two award categories: **Best Paper** and **Best Theme Paper**. In total, we received **72 archival** and **15 non-archival submissions**, of which we accepted 38 (53%) and 11 (73%), respectively. These works will be showcased through poster sessions and lightning talks, facilitating both in-person and online participation. The program also features keynote talks from Kate Sim (Children's Online Safety and Privacy Research), Cordelia Moore (independent trauma counsellor), and Francesco Barbieri (Meta).

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. We hope that WOAH 2025 serves as a platform for productive discussions, meaningful collaborations, and continued progress in addressing online abuse and harms.

Agostina, Christine, Debora, Flor, Zeerak, and Francielle

Organizing Committee

Workshop Organisers

Agostina Calabrese, University of Edinburgh
Christine de Kock, University of Melbourne
Debora Nozza, Bocconi University
Flor Miriam Plaza-del-Arco, LIACS, Leiden University
Zeerak Talat, University of Edinburgh
Francielle Vargas, University of São Paulo

Program Committee

Chairs

Agostina Calabrese, The University of Edinburgh
Christine De Kock, University of Melbourne
Debora Nozza, Bocconi University
Flor Miriam Plaza-del-Arco, Leiden University
Zeerak Talat, University of Edinburgh
Francielle Vargas, University of São Paulo

Program Committee

Gavin Abercrombie, Heriot Watt University
Oluwaseyi Adeyemo, Afe Babalola University
Syed Sarfaraz Akhtar, Apple Inc
Diego Alves, Saarland University
Jisun An, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington
Dimosthenis Antypas, Cardiff University
Mario Aragon, Universidade de Santiago de Compostela
Arnav Arora, University of Copenhagen
Shubham Atreja, University of Michigan School of Information
Nikolay Babakov, Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela
Baran Barbarestani, Vrije Universiteit Amsterdam
Valerio Basile, University of Turin
Prabh Simran Baweja, Apple Inc.
Gemma Bel-Enguix, Universidad Nacional Autónoma de México
Hagen Blix, New York University
Helena Bonaldi, Fondazione Bruno Kessler
Caroline Brun, Naver Labs Europe
Tommaso Caselli, Rijksuniversiteit Groningen
Camilla Casula, University of Trento / Fondazione Bruno Kessler
Amanda Cercas Curry, Bocconi University
Alessandra Teresa Cignarella, LT3, Ghent University
Ryan Cotterell, ETH Zürich
Greta Damo, Université Côte d'Azur
Arijit Das, Jadavpur University
Daryna Dementieva, Technical University of Munich
Ali Derakhshan, University of California Irvine
Jan Fillies, Freie Universität Berlin
Björn Gambäck Gambäck, Norwegian University of Science and Technology
Sara Gemelli, University of Pavia, University of Bergamo
Matteo Guida, The University of Melbourne
Balint Gyevnar, University of Edinburgh
Karina Halevy, Carnegie Mellon University
Eduard Hovy, University of Melbourne
Wenjia Hu, Carnegie Mellon University
Comfort Ilevbare, Afe Babalola University

Farane Jalali Farahani, Institute for Artificial Intelligence, University of Stuttgart
Mohammad Aflah Khan, IIIT Delhi
Sangyeop Kim, Seoul National University
Shaghayegh Kolli, Student
Ioannis Konstas, Heriot-Watt University
Katerina Korre, University of Bologna
Sandra Kübler, Indiana University
Dong-Ho Lee, University of Southern California
Roy Ka-wei Lee, Singapore University of Technology and Design
Els Lefever, LT3, Ghent University
Chaya Liebeskind, Jerusalem College of Technology , Lev Academic Center
Lucy Lin, Spotify
Ajay Malik, CU-R
Sayan Mandal, AMD
Marta Marchiori Manerba, Università di Pisa
Ilia Markov, Vrije Universiteit Amsterdam, CLTL
Diana Maynard, University of Sheffield
Helena Mihaljevic, Hochschule für Technik und Wirtschaft Berlin
Emanuele Moscato, Bocconi University
Arianna Muti, Bocconi University
Nazia Nafis, The University of Sheffield
Isar Nejadgholi, National Research Council Canada
Hellina Hailu Nigatu, UC Berkeley
Ayushi Nirmal, Arizona State University
Ali Omrani, University of Southern California
Matthias Orlikowski, Bielefeld University
Pia Pachinger, TU Wien
Lucia Passaro, University of Pisa
Viviana Patti, University of Turin, Dipartimento di Informatica
Siddhesh Pawar, Google
Nicolã Penzo, University of Trento, Fondazione Bruno Kessler
Anna Maria Planitzer, Political Communication Research Group, Department of Communication, University of Vienna
Nirmalendu Prakash, SUTD
Michal Ptaszynski, Kitami Institute of Technology
Jessica Quaye, Harvard University
Georg Rehm, DFKI
Bjorn Ross, University of Edinburgh
Hamidreza Saffari, Politecnico di Milano
Miriam Schirmer, Northwestern University
Michael Sejr Schlichtkrull, University of Cambridge
Indira Sen, RWTH Aachen
Mattia Setzu, University of Pisa
Mohammadamin Shaifiei, University of Milan
Akshay Singh, Indian Institute of Technology Roorkee
Jeffrey Sorensen, Google Jigsaw
Steffen Staab, University of Stuttgart and University of Southampton
Vivian Stamou, Archimedes AI
Elisabeth Steffen, HTW Berlin
Paul Thompson, National Centre for Text Mining, School of Computer Science, University of Manchester

Zuoyu Tian, Indiana University
Manuel Tonneau, University of Oxford, World Bank
Dimitrios Tsarapatsanis, University of York
Aatman Vaidya, Tattle Civic Tech
Avijit Vajpayee, Amazon
Juan Vasquez, Department of Computer Science, University of Colorado Boulder
Charles Welch, McMaster University
Guanqun Yang, Stevens Institute of Technology
Zachary Yang, McGill | Mila | Ubisoft
Jason Zhang, Student Researcher
Yi Zheng, University of Edinburgh

Table of Contents

Program

Friday, August 1, 2025

09:00 - 09:20 *Opening Remarks*

09:20 - 10:00 *Invited Talk 1 - Cordelia Moore*

10:00 - 10:30 *Best paper & Best theme paper*

From civility to parity: Marxist-feminist ethics for context-aware algorithmic content moderation
Dayei Oh

Catching Stray Balls: Football, fandom, and the impact on digital discourse
Mark Hill

10:30 - 10:50 *Mini Break*

10:50 - 11:30 *Invited Talk 2 - Francesco Barbieri*

11:30 - 12:10 *Invited Talk 3 - Kate Sim*

12:10 - 13:40 *Lunch Break*

13:40 - 15:10 *In-Person Poster Session*

A Comprehensive Taxonomy of Bias Mitigation Methods for Hate Speech Detection
Jan Fillies, Marius Wawerek and Adrian Paschke

Sensitive Content Classification in Social Media: A Holistic Resource and Evaluation
Dimosthenis Antypas, Indira Sen, Carla Perez Almendros, Jose Camacho-Collados and Francesco Barbieri

A Novel Dataset for Classifying German Hate Speech Comments with Criminal Relevance
Vincent Kums, Florian Meyer, Luisa Pivit, Uliana Vedenina, Jonas Wortmann, Melanie Siegel and Dirk Labudde

Learning from Disagreement: Entropy-Guided Few-Shot Selection for Toxic Language Detection
Tommaso Caselli and Flor Miriam Plaza-del-Arco

Friday, August 1, 2025 (continued)

Friday, August 1, 2025 (continued)

(Civil Society Reports) - Monitoring Hate Speech in Indonesia: An NLP-based Classification of Social Media Texts
Musa Wijanarko, Lucky Susanto, Prasetia Pratama, Derry Tanti Wijaya and Ika Idris

Who leads? Who follows? Temporal dynamics of political dogwhistles in Swedish online communities
Max Boholm, Gregor Rettenegger, Ellen Breitholtz, Robin Cooper, Elina Lindgren, Björn Rönnerstrand and Asad Sayeed

(Non-Archival Submission) A Multi-Labeled Dataset for Indonesian Discourse: Examining Toxicity, Polarization, and Demographics Information
Lucky Susanto, Musa Wijanarko, Prasetia Pratama, Zilu Tang, Fariz Akyas, Traci Hong, Ika Idris, Alham Fikri Aji and Derry Tanti Wijaya

Detecting Child Objectification on Social Media: Challenges in Language Modeling
Miriam Schirmer, Angelina Voggenreiter, Juergen Pfeffer and Agnes Horvat

Can Prompting LLMs Unlock Hate Speech Detection across Languages? A Zero-shot and Few-shot Study
Faeze Ghorbanpour, Daryna Dementieva and Alexandar Fraser

Alignment by Disagreement? Toward Investigating LLMs' Adaptation to Personal and Sociodemographic Variability in the Perception of Toxicity
Pia Pachinger, Anna Maria Planitzer, Julia Neidhardt, Allan Hanbury and Sophie Lecheler

Multilingual Analysis of Narrative Properties in Conspiracist vs Mainstream Telegram Channels
Katarina Laken, Matteo Melis, Sara Tonelli and Marcos Garcia

Hate Explained: Evaluating NER-Enriched Text in Human and Machine Moderation of Hate Speech
Andres Carvallo, Marcelo Mendoza, Miguel Fernandez, Maximiliano Ojeda, Lilly Guevara, Diego Varela, Martin Borquez, Nicolas Buzeta and Felipe Ayala

Personas with Attitudes: Controlling LLMs for Diverse Data Annotation
Leon Fröhling, Gianluca Demartini and Dennis Assenmacher

Graph of Attacks with Pruning: Optimizing Stealthy Jailbreak Prompt. Generation for Enhanced LLM Content Moderation
Daniel Schwarz, Dmitriy Bespalov, Zhe Wang, Ninad Kulkarni and Yanjun Qi

A Modular Taxonomy for Hate Speech Definitions and Its Impact on Zero-Shot LLM Classification Performance
Matteo Melis, Gabriella Lapesa and Dennis Assenmacher

Friday, August 1, 2025 (continued)

Using LLMs and Preference Optimization for Agreement-Aware HateWiC Classification
Sebastian Loftus, Adrian Mülthaler, Sanne Hoeken, Sina Zarrieß and Ozge Alacam

When Claims Evolve: Evaluating and Enhancing the Robustness of Embedding Models Against Misinformation Edits
Jabez Magomere, Emanuele La Malfa, Manuel Tonneau, Ashkan Kazemi and Scott A. Hale

Hatevolution: What Static Benchmarks Don't Tell Us
Chiara Di Bonaventura, Barbara McGillivray, Yulan He and Albert Meroño-Peñuela

15:10 - 15:40 *Lightning Talks for Remote Attendants*

Debiasing Static Embeddings for Hate Speech Detection
Ling Sun, Soyoung Kim, Xiao Dong and Sandra Kübler

Hate Speech in Times of Crises: a Cross-Disciplinary Analysis of Online Xenophobia in Greece
Maria Pontiki, Vasiliki Georgiadou, Lamprini Rori and Maria Gavriilidou

Hostility Detection in UK Politics: A Dataset on Online Abuse Targeting MPs
Mugdha Pandya, Mali Jin, Kalina Bontcheva and Diana Maynard

QGuard:Question-based Zero-shot Guard for Multi-modal LLM Safety
Taegyeong Lee, Jeonghwa Yoo, Hyoungseo Cho, Soo Yong Kim and Yunho Maeng

Anti-Phishing Layered Prompting (ALP): A Structured Few-Shot Approach to Enhance Webpage Phishing Detection
Atharva Bhargude, Ishan Gonehal, Chandler Haney, Dave Yoon, Kaustubh Vinn and Kevin Zhu

Red-Teaming for Uncovering Societal Bias in Large Language Models
Chu Fei Luo, Ahmad Ghawanmeh, Kashyap Coimbatore Murali, Bhimshetty Bharat Kumar, Murli Jadhav, Xiaodan Zhu and Faiza Khan Khattak

15:40 - 16:00 *Coffee Break*

16:00 - 17:00 *Panel Discussion*

Friday, August 1, 2025 (continued)

17:00 - 17:10 *Closing Remarks*