

The 9th Workshop on Online Abuse and Harms (WOAH 2025)

Vienna, Austria
1 August 2025

ISBN: 979-8-3313-2422-3

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2025) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>A Comprehensive Taxonomy of Bias Mitigation Methods for Hate Speech Detection</i> Jan Fillies, Marius Wawerek and Adrian Paschke	1
<i>Sensitive Content Classification in Social Media: A Holistic Resource and Evaluation</i> Dimosthenis Antypas, Indira Sen, Carla Perez Almendros, Jose Camacho-Collados and Francesco Barbieri	17
<i>From civility to parity: Marxist-feminist ethics for context-aware algorithmic content moderation</i> Dayei Oh	32
<i>A Novel Dataset for Classifying German Hate Speech Comments with Criminal Relevance</i> Vincent Kums, Florian Meyer, Luisa Pivit, Uliana Vedenina, Jonas Wortmann, Melanie Siegel and Dirk Labudde	41
<i>Learning from Disagreement: Entropy-Guided Few-Shot Selection for Toxic Language Detection</i> Tommaso Caselli and Flor Miriam Plaza-del-Arco	53
<i>Debiasing Static Embeddings for Hate Speech Detection</i> Ling Sun, Soyoung Kim, Xiao Dong and Sandra Kübler	67
<i>Web(er) of Hate: A Survey on How Hate Speech Is Typed</i> Luna Wang, Andrew Caines and Alice Hutchings	77
<i>Think Like a Person Before Responding: A Multi-Faceted Evaluation of Persona-Guided LLMs for Countering Hate Speech.</i> Mikel Nguējio, Flor Miriam Plaza-del-Arco, Yi-Ling Chung, Danda Rawat and Amanda Cercas Curry	104
<i>HODIAT: A Dataset for Detecting Homophobic Hate Speech in Italian with Aggressiveness and Target Annotation</i> Greta Damo, Alessandra Teresa Cignarella, Tommaso Caselli, Viviana Patti and Debora Nozza	124
<i>Beyond the Binary: Analysing Transphobic Hate and Harassment Online</i> Anna Talas and Alice Hutchings	136
<i>Evading Toxicity Detection with ASCII-art: A Benchmark of Spatial Attacks on Moderation Systems</i> Sergey Berezin, Reza Farahbakhsh and Noel Crespi	153
<i>Debunking with Dialogue? Exploring AI-Generated Counterspeech to Challenge Conspiracy Theories</i> Mareike Lisker, Christina Gottschalk and Helena Mihaljević	163
<i>MisinfoTeleGraph: Network-driven Misinformation Detection for German Telegram Messages</i> Lu Kalkbrenner, Veronika Solopova, Steffen Zeiler, Robert Nickel and Dorothea Kolossa ...	179
<i>Catching Stray Balls: Football, fandom, and the impact on digital discourse</i> Mark Hill	192
<i>Exploring Hate Speech Detection Models for Lithuanian Language</i> Justina Mandravickaitė, Eglė Rimkienė, Mindaugas Petkevičius, Milita Songailaitė, Eimantas Zaranka and Tomas Krilavičius	206
<i>RAG and Recall: Multilingual Hate Speech Detection with Semantic Memory</i> Khoulood Mnassri, Reza Farahbakhsh and Noel Crespi	219

<i>Implicit Hate Target Span Detection in Zero- and Few-Shot Settings with Selective Sub-Billion Parameter Models</i>	
Hossam Boudraa, Benoit Favre and Raquel Urena	228
<i>Hate Speech in Times of Crises: a Cross-Disciplinary Analysis of Online Xenophobia in Greece</i>	
Maria Pontiki, Vasiliki Georgiadou, Lamprini Rori and Maria Gavriilidou	241
<i>Hostility Detection in UK Politics: A Dataset on Online Abuse Targeting MPs</i>	
Mugdha Pandya, Mali Jin, Kalina Bontcheva and Diana Maynard	254
<i>Detoxify-IT: An Italian Parallel Dataset for Text Detoxification</i>	
Viola De Ruvo, Arianna Muti, Daryna Dementieva and Debora Nozza	267
<i>Pathways to Radicalisation: On Research for Online Radicalisation in Natural Language Processing and Machine Learning</i>	
Zeerak Talat, Michael Sejr Schlichtkrull, Pranava Madhyastha and Christine De Kock	276
<i>Social Hatred: Efficient Multimodal Detection of Hatemongers</i>	
Tom Marzea, Abraham Israeli and Oren Tsur	284
<i>Blue-haired, misandriche, rabiata: Tracing the Connotation of 'Feminist(s)' Across Time, Languages and Domains</i>	
Arianna Muti, Sara Gemelli, Emanuele Moscato, Emilie Francis, Amanda Cercas Curry, Flor Miriam Plaza-del-Arco and Debora Nozza	299
<i>Towards Fairness Assessment of Dutch Hate Speech Detection</i>	
Julie Bauer, Rishabh Kaushal, Thales Bertaglia and Adriana Iamnitchi	312
<i>Between Hetero-Fatalism and Dark Femininity: Discussions of Relationships, Sex, and Men in the Femosphere</i>	
Emilie Francis	325
<i>Can LLMs Rank the Harmfulness of Smaller LLMs? We are Not There Yet</i>	
Berk Atıl, Vipul Gupta, Sarkar Snigdha Sarathi Das and Rebecca Passonneau	342
<i>Are You Trying to Convince Me or Are You Trying to Deceive Me? Using Argumentation Types to Identify Deceptive News</i>	
Ricardo Muñoz Sánchez, Emilie Francis and Anna Lindahl	355
<i>QGuard: Question-based Zero-shot Guard for Multi-modal LLM Safety</i>	
Taeyyeong Lee, Jeonghwa Yoo, Hyoungseo Cho, Soo Yong Kim and Yunho Maeng	373
<i>Who leads? Who follows? Temporal dynamics of political dogwhistles in Swedish online communities</i>	
Max Boholm, Gregor Rettenegger, Ellen Breitholtz, Robin Cooper, Elina Lindgren, Björn Rönnerstrand and Asad Sayeed	383
<i>Detecting Child Objectification on Social Media: Challenges in Language Modeling</i>	
Miriam Schirmer, Angelina Voggenreiter, Juergen Pfeffer and Agnes Horvat	396
<i>Can Prompting LLMs Unlock Hate Speech Detection across Languages? A Zero-shot and Few-shot Study</i>	
Faeze Ghorbanpour, Daryna Dementieva and Alexandar Fraser	413
<i>Multilingual Analysis of Narrative Properties in Conspiracist vs Mainstream Telegram Channels</i>	
Katarina Laken, Matteo Melis, Sara Tonelli and Marcos Garcia	426

<i>Hate Explained: Evaluating NER-Enriched Text in Human and Machine Moderation of Hate Speech</i> Andres Carvallo, Marcelo Mendoza, Miguel Fernandez, Maximiliano Ojeda, Lilly Guevara, Diego Varela, Martin Borquez, Nicolas Buzeta and Felipe Ayala	458
<i>Personas with Attitudes: Controlling LLMs for Diverse Data Annotation</i> Leon Fröhling, Gianluca Demartini and Dennis Assenmacher	468
<i>Graph of Attacks with Pruning: Optimizing Stealthy Jailbreak Prompt. Generation for Enhanced LLM Content Moderation</i> Daniel Schwarz, Dmitriy Bespalov, Zhe Wang, Ninad Kulkarni and Yanjun Qi	482
<i>A Modular Taxonomy for Hate Speech Definitions and Its Impact on Zero-Shot LLM Classification Performance</i> Matteo Melis, Gabriella Lapesa and Dennis Assenmacher	490
<i>Red-Teaming for Uncovering Societal Bias in Large Language Models</i> Chu Fei Luo, Ahmad Ghawanmeh, Kashyap Coimbatore Murali, Bhimshetty Bharat Kumar, Murli Jadhav, Xiaodan Zhu and Faiza Khan Khattak	522
<i>Using LLMs and Preference Optimization for Agreement-Aware HateWiC Classification</i> Sebastian Loftus, Adrian Mülthaler, Sanne Hoeken, Sina Zarrieß and Ozge Alacam	538