

6th Workshop on Online Abuse and Harms (WOAH 2022)

Online
14 July 2022

ISBN: 978-1-7138-5646-7

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© (2022) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2022)

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>Separating Hate Speech and Offensive Language Classes via Adversarial Debiasing</i> Shuzhou Yuan, Antonis Maronikolakis and Hinrich Schütze	1
<i>Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions</i> Mana Ashida and Mamoru Komachi	11
<i>GreaseVision: Rewriting the Rules of the Interface</i> Siddhartha Datta, Konrad Kollnig and Nigel Shadbolt	24
<i>Improving Generalization of Hate Speech Detection Systems to Novel Target Groups via Domain Adaptation</i> Florian Ludwig, Klara Dolos, Torsten Zesch and Eleanor Hobley	29
<i>“Zo Grof !”: A Comprehensive Corpus for Offensive and Abusive Language in Dutch</i> Ward Ruitenbeek, Victor Zwart, Robin Van Der Noord, Zhenja Gnezdilov and Tommaso Caselli	40
<i>Counter-TWIT: An Italian Corpus for Online Counterspeech in Ecological Contexts</i> Pierpaolo Goffredo, Valerio Basile, Biancamaria Cepollaro and Viviana Patti	57
<i>StereoKG: Data-Driven Knowledge Graph Construction For Cultural Knowledge and Stereotypes</i> Awantee Deshpande, Dana Ruiter, Marius Mosbach and Dietrich Klakow	67
<i>The subtle language of exclusion: Identifying the Toxic Speech of Trans-exclusionary Radical Feminists</i> Christina Lu and David Jurgens	79
<i>Lost in Distillation: A Case Study in Toxicity Modeling</i> Alyssa Chvasta, Alyssa Lees, Jeffrey Sorensen, Lucy Vasserman and Nitesh Goyal	92
<i>Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words</i> Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou and Stella Markantonatou	102
<i>Free speech or Free Hate Speech? Analyzing the Proliferation of Hate Speech in Parler</i> Abraham Israeli and Oren Tsur	109
<i>Resources for Multilingual Hate Speech Detection</i> Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña and Valentina Proust	122
<i>Enriching Abusive Language Detection with Community Context</i> Haji Mohammad Saleem, Jana Kurrek and Derek Ruths	131
<i>A Comprehensive Dataset for German Offensive Language and Conversation Analysis</i> Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel and Dirk Labudde	143
<i>Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models</i> Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat and Bertie Vidgen	154
<i>Distributional properties of political dogwhistle representations in Swedish BERT</i> Niclas Hertzberg, Robin Cooper, Elina Lindgren, Björn Rönnerstrand, Gregor Rettenegger, Ellen Breitholtz and Asad Sayeed	170

<i>Hate Speech Criteria: A Modular Approach to Task-Specific Hate Speech Definitions</i>	
Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos and Antske Fokkens . . .	176
<i>Accounting for Offensive Speech as a Practice of Resistance</i>	
Mark Diaz, Razvan Amironesei, Laura Weidinger and Iason Gabriel	192
<i>Towards a Multi-Entity Aspect-Based Sentiment Analysis for Characterizing Directed Social Regard in Online Messaging</i>	
Joan Zheng, Scott Friedman, Sonja Schmer-galunder, Ian Magnusson, Ruta Wheelock, Jeremy Gottlieb, Diana Gomez and Christopher Miller	203
<i>Flexible text generation for counterfactual fairness probing</i>	
Zee Fryer, Vera Axelrod, Ben Packer, Alex Beutel, Jilin Chen and Kellie Webster	209
<i>Users Hate Blondes: Detecting Sexism in User Comments on Online Romanian News</i>	
Andreea Moldovan, Karla Csürös, Ana-maria Bucur and Loredana Bercuci	230
<i>Targeted Identity Group Prediction in Hate Speech Corpora</i>	
Pratik Sachdeva, Renata Barreto, Claudia Von Vacano and Chris Kennedy	231
<i>Revisiting Queer Minorities in Lexicons</i>	
Krithika Ramesh, Sumeet Kumar and Ashiqur Khudabukhsh	245
<i>HATE-ITA: New Baselines for Hate Speech Detection in Italian</i>	
Debora Nozza, Federico Bianchi and Giuseppe Attanasio	252