

5th Workshop on Online Abuse and Harms (WOAH 2021)

Held online due to COVID-19

Bangkok, Thailand
6 August 2021

ISBN: 978-1-7138-3391-8

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© (2021) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2021)

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>Exploiting Auxiliary Data for Offensive Language Detection with Bidirectional Transformers</i> Sumer Singh and Sheng Li	1
<i>Modeling Profanity and Hate Speech in Social Media with Semantic Subspaces</i> Vanessa Hahn, Dana Ruiter, Thomas Kleinbauer and Dietrich Klakow	6
<i>HateBERT: Retraining BERT for Abusive Language Detection in English</i> Tommaso Caselli, Valerio Basile, Jelena Mitrović and Michael Granitzer	17
<i>Memes in the Wild: Assessing the Generalizability of the Hateful Memes Challenge Dataset</i> Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski and Yuki M Asano	26
<i>Measuring and Improving Model-Moderator Collaboration using Uncertainty Estimation</i> Ian Kivlichan, Zi Lin, Jeremiah Liu and Lucy Vasserman	36
<i>DALC: the Dutch Abusive Language Corpus</i> Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman and Malvina Nissim	54
<i>Offensive Language Detection in Nepali Social Media</i> Nobal B. Niraula, Saurab Dulal and Diwa Koirala	67
<i>MIN_PT: An European Portuguese Lexicon for Minorities Related Terms</i> Paula Fortuna, Vanessa Cortez, Miguel Sozinho Ramalho and Laura Pérez-Mayos	76
<i>Fine-Grained Fairness Analysis of Abusive Language Detection Systems with CheckList</i> Marta Marchiori Manerba and Sara Tonelli	81
<i>Improving Counterfactual Generation for Fair Hate Speech Detection</i> Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren and Morteza Dehghani	92
<i>Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon</i> Samira Zad, Joshuan Jimenez and Mark Finlayson	102
<i>Mitigating Biases in Toxic Language Detection through Invariant Rationalization</i> Yung-Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung-yi Lee, Yun-Nung Chen and Shang-Wen Li	114
<i>Fine-grained Classification of Political Bias in German News: A Data Set and Initial Experiments</i> Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider and Georg Rehm	121
<i>Jibes & Delights: A Dataset of Targeted Insults and Compliments to Tackle Online Abuse</i> Ravsimar Sodhi, Kartikey Pant and Radhika Mamidi	132
<i>Context Sensitivity Estimation in Toxicity Detection</i> Alexandros Xenos, John Pavlopoulos and Ion Androutsopoulos	140
<i>A Large-Scale English Multi-Label Twitter Dataset for Cyberbullying and Online Abuse Detection</i> Semiu Salawu, Jo Lumsden and Yulan He	146

<i>Toxic Comment Collection: Making More Than 30 Datasets Easily Accessible in One Unified Format</i> Julian Risch, Philipp Schmidt and Ralf Krestel	157
<i>When the Echo Chamber Shatters: Examining the Use of Community-Specific Language Post-Subreddit Ban</i> Milo Trujillo, Sam Rosenblatt, Guillermo de Anda Jáuregui, Emily Moog, Briane Paul V. Samson, Laurent Hébert-Dufresne and Allison M. Roth	164
<i>Targets and Aspects in Social Media Hate Speech</i> Alexander Shvets, Paula Fortuna, Juan Soler and Leo Wanner	179
<i>Abusive Language on Social Media Through the Legal Looking Glass</i> Thales Bertaglia, Andreea Grigoriu, Michel Dumontier and Gijs van Dijck	191
<i>Findings of the WOAHA 5 Shared Task on Fine Grained Hateful Memes Detection</i> Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen and Zeerak Waseem	201
<i>VL-BERT+: Detecting Protected Groups in Hateful Multimodal Memes</i> Piush Aggarwal, Michelle Espranita Liman, Darina Gold and Torsten Zesch	207
<i>Racist or Sexist Meme? Classifying Memes beyond Hateful</i> Haris Bin Zia, Ignacio Castro and Gareth Tyson	215
<i>Multimodal or Text? Retrieval or BERT? Benchmarking Classifiers for the Shared Task on Hateful Memes</i> Vasiliki Kougia and John Pavlopoulos	220