Fourth Workshop on Online Abuse and Harms (WOAH 2020)

Online
20 November 2020

ISBN: 978-1-7138-2007-9

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© (2020) 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

Online Abuse and Human Rights

Online Abuse and Human Rights: WOAH Satellite Session at RightsCon 2020
Vinodkumar Prabhakaran, Zeerak Waseem, Seyi Akiwowo and Bertie Vidgen 1

Regular Contributions

A Novel Methodology for Developing Automatic Harassment Classifiers for Twitter
Ishaan Arora, Julia Guo, Sarah Ita Levitan, Susan McGregor and Julia Hirschberg 7

Using Transfer-based Language Models to Detect Hateful and Offensive Language Online
Vebjørn Isaksen and Björn Gambäck 16

Fine-tuning BERT for multi-domain and multi-label incivil language detection
Kadir Bulut Ozler, Kate Kenski, Steve Rains, Yotam Shmargad, Kevin Coe and Steven Bethard . 28

HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language
Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile and Viviana Patti 34

Abusive Language Detection using Syntactic Dependency Graphs
Kanika Narang and Chris Brew 44

Impact of politically biased data on hate speech classification
Maximilian Wich, Jan Bauer and Georg Groh 54

Reducing Unintended Identity Bias in Russian Hate Speech Detection
Nadezhda Zueva, Madina Kabirova and Pavel Kalaidin 65

Investigating Sampling Bias in Abusive Language Detection
Dante Razo and Sandra Kübler 70

Attending the Emotions to Detect Online Abusive Language
Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar and Thamar Solorio . . 79

Enhancing the Identification of Cyberbullying through Participant Roles
Gathika Rathnayake, Thushari Atapattu, Mahen Herath, Georgia Zhang and Katrina Falkner . . . 89

Developing a New Classifier for Automated Identification of Incivility in Social Media
Sam Davidson, Qiusi Sun and Magdalena Wojcieszak 95

Countering hate on social media: Large scale classification of hate and counter speech
Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne and Mirta Galesic 102

Moderating Our (Dis)Content: Renewing the Regulatory Approach
Claire Pershan 113

Six Attributes of Unhealthy Conversations
Ilan Price, Jordan Gifford-Moore, Jory Fleming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon and Jeffrey Sorensen 114

A Unified Taxonomy of Harmful Content
Michele Banko, Brendon MacKeen and Laurie Ray 125

Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage
Jana Kurrek, Haji Mohammad Saleem and Derek Ruths 138

In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets
Kosisochukwu Madukwe, Xiaoying Gao and Bing Xue 150

Detecting East Asian Prejudice on Social Media
Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall and Rebekah Tromble 162

Shared Exploration

On Cross-Dataset Generalization in Automatic Detection of Online Abuse
Isar Nejadgholi and Svetlana Kiritchenko 173

Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics
Hala Al Kuwatly, Maximilian Wich and Georg Groh 184

Investigating Annotator Bias with a Graph-Based Approach
Maximilian Wich, Hala Al Kuwatly and Georg Groh 191