

Second Workshop on Abusive Language Online 2018

Brussels, Belgium
31 October 2018

ISBN: 978-1-5108-7419-0

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© (2018) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2019)

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>Neural Character-based Composition Models for Abuse Detection</i> Pushkar Mishra, Helen Yannakoudakis and Ekaterina Shutova	1
<i>Hate Speech Dataset from a White Supremacy Forum</i> Ona de Gibert, Naiara Perez, Aitor García Pablos and Montse Cuadros	11
<i>A Review of Standard Text Classification Practices for Multi-label Toxicity Identification of Online Content</i> Isuru Gunasekara and Isar Nejadgholi	21
<i>Predictive Embeddings for Hate Speech Detection on Twitter</i> Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown and Susan McGregor	26
<i>Challenges for Toxic Comment Classification: An In-Depth Error Analysis</i> Betty van Aken, Julian Risch, Ralf Krestel and Alexander Löser	33
<i>Aggression Detection on Social Media Text Using Deep Neural Networks</i> Vinay Singh, Aman Varshney, Syed Sarfaraz Akhtar, Deepanshu Vijay and Manish Shrivastava	43
<i>Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying</i> Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini and Enrico Piras	51
<i>Improving Moderation of Online Discussions via Interpretable Neural Models</i> Andrej Švec, Matúš Pikuliak, Marian Simko and Maria Bielikova	60
<i>Aggressive language in an online hacking forum</i> Andrew Caines, Sergio Pastrana, Alice Hutchings and Paula Buttery	66
<i>The Effects of User Features on Twitter Hate Speech Detection</i> Elise Fehn Unsvåg and Björn Gambäck	75
<i>Interpreting Neural Network Hate Speech Classifiers</i> Cindy Wang	86
<i>Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks</i> Rijul Magu and Jiebo Luo	93
<i>Comparative Studies of Detecting Abusive Language on Twitter</i> Younghun Lee, Seunghyun Yoon and Kyomin Jung	101
<i>Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs</i> sima sharifirad, Borna Jafarpour and Stan Matwin	107
<i>Learning Representations for Detecting Abusive Language</i> Magnus Sahlgren, Tim Isbister and Fredrik Olsson	115
<i>Datasets of Slovene and Croatian Moderated News Comments</i> Nikola Ljubešić, Tomaž Erjavec and Darja Fišer	124
<i>Cross-Domain Detection of Abusive Language Online</i> Mladen Karan and Jan Šnajder	132

<i>Did you offend me? Classification of Offensive Tweets in Hinglish Language</i> Puneet Mathur, Ramit Sawhney, Meghna Ayyar and Rajiv Shah	138
<i>Decipherment for Adversarial Offensive Language Detection</i> Zhelun Wu, Nishant Kambhatla and Anoop Sarkar	149
<i>The Linguistic Ideologies of Deep Abusive Language Classification</i> Michael Castelle	160