EMNLP 2018

Second Workshop on Abusive Language Online

Proceedings of the Workshop, co-located with EMNLP 2018

October 31, 2018
Brussels, Belgium

Sponsors

Primary Sponsor

HUMAN SECURITY COLLABORATORY HS COLLAB

ASU Global Security Initiative Arizona State University

Platinum Sponsors

figure eight

Gold Sponsors

amazon

Bloomberg®

facebook

Google

REUTERS

SPLC Southern Poverty Law Center

Silver Sponsors

data.world

The New York Times

Bronze Sponsors

FACTMATA

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-948087-68-1

Introduction

Interaction amongst users on social networking platforms can enable constructive and insightful conversations and civic participation; however, on many sites that encourage user interaction, verbal abuse has become commonplace. Abusive behavior such as cyberbullying, hate speech, and scapegoating can poison the social climates within online communities. The last few years have seen a surge in such abusive online behavior, leaving governments, social media platforms, and individuals struggling to deal with the consequences.

As a field that works directly with computational analysis of language, the NLP community is uniquely positioned to address the difficult problem of abusive language online; encouraging collaborative and innovate work in this area is the goal of this workshop. The first year of the workshop saw 14 papers presented in a day-long program including interdisciplinary panels and active discussion. In this second edition, we have aimed to build on the success of the first year, maintaining a focus on computationally detecting abusive language and encouraging interdisciplinary work. Reflecting the growing research focus on this topic, the number of submissions received more than doubled from 22 in last year's edition of the workshop to 48 this year.

The workshop will be broken into four sessions throughout the day. In the first session, we are delighted to have two invited speakers from beyond the NLP community joining us to share their unique perspectives and expertise:

> **Mikki Kendall**
> *The Gamification of Hate*
> Mikki Kendall has written for The Washington Post, Boston Globe, Time, Ebony, Essence, and other online and print markets. Born and raised in Chicago, her books *Hood Feminism* and *Amazons, Abolitionists, and Activists: A Graphic History of Women's Fight For their Rights* will be published by Penguin Random House in 2019. Having experienced online harassment, she has worked on projects related to abusive online cultures for nearly a decade.
>
> **Maryant Fernández Pérez**
> *The Damaging Effect of Privatised Law Enforcement in Tackling Illegal Content*
> Maryant Fernández Pérez is a Senior Policy Advisor at European Digital Rights (EDRi) and a lawyer admitted to the Madrid Bar association. She leads EDRi's work on surveillance and law enforcement, freedom of expression and intermediary liability, net neutrality, digital trade, transparency, internet governance and international engagement. Maryant is the author of several publications and has been a speaker at multiple conferences in Europe and around the world.

In the second session, a panel of experts both from within and outside of the NLP community will debate and frame the major issues facing the computational analysis of abusive language online, particularly as relevant to the morning's talks. This panel will be followed by a period for a discussion amongst all attendees.

The third session will be used for sharing the research results archived in these proceedings, presented as posters to encourage discussion. Finally, in the fourth session, the panelists, speakers, and participants will return to give feedback on what they've seen and heard, leading into a synthesizing discussion amongst all attendees facilitated by workshop organizer Jacqueline Wernimont. With this format we aim to open a space for synergies between the talks, panels, and discussions throughout the day and encourage interdisciplinary approaches to future work in the field.

The submissions to be presented at the workshop represent a compelling diversity of methods, topics, and approaches to the difficult problem of abusive language online, including embedding-based, adversarial, and neural models; the creation of new datasets from diverse sources such as WhatsApp and white supremacist forums; in-depth error analysis and classification interpretability analysis; and studies of languages beyond English such as Slovene, Croatian, and code-mixed Hindi and English. The workshop received 48 paper submissions, of which 21 were accepted, for an acceptance rate of 43%.

In organizing this workshop we collaborated with StackOverflow to curate a dataset of moderated comments, proposed as an unshared task. This dataset was ultimately utilized by one of the accepted papers and will hopefully encourage more work moving forward in close collaboration with industry partners. We have also reached an agreement with the journal First Monday to publish a special issue resulting from the joint proceedings of this workshop and the previous edition, wherein a subset of the papers will be nominated and the authors given an opportunity to expand them into full journal articles.

In closing, we wish to extend our sincere gratitude to our sponsors for their generous financial contributions and our reviewers for their time and expertise, without which this workshop would not have been possible.

- Zeerak, Jacque, Vinod, Darja, Ruihong, and Rob

Organizers:

Darja Fišer, University of Ljubljana & the Jožef Stefan Institute
Ruihong Huang, Texas A&M University
Vinodkumar Prabhakaran, Stanford University
Rob Voigt, Stanford University
Zeerak Waseem, University of Sheffield
Jacqueline Wernimont, Dartmouth College

Program Committee:

Ion Androutsopoulos, Department of Informatics, Athens University of Economics and Business, Greece
Veronika Bajt, Peace Institute, Slovenia
Susan Benesch, Berkman Klein Center, United States of America
Darina Benikova, University of Duisburg-Essen, Germany
Joachim Bingel, University of Copenhagen, Denmark
Ariane Chan, Data.world, United States of America
Wendy Chun, Simon Fraser University, Canada
Thomas Davidson, Cornell University, United States of America
Kelly Dennis, University of Connecticut, United States of America
Lucas Dixon, Jigsaw (Google), United States of America
Nemanja Djuric, Uber, United States of America
Yanai Elazar, Bar-Ilan University, Israel
Paula Fortuna, University of Porto, Portugal
Maya Ganesh, Leuphana University of Lüneburg, Germany
Tassie Gniady, Indiana University Bloomington, United States of America
Genevieve Gorrell, Sheffield University, United Kingdom
Hugo Jair Escalante, National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico
Björn Gambäck, Norwegian University of Science and Technology, Norway
Lee Gillam, University of Surrey, United Kingdom
Vojko Gorjanc, University of Ljubljana, Slovenia
Erica Greene, Jigsaw (Google), United States of America
Seda Gurses, KU Leuven, Belgium
Mareike Hartmann, University of Copenhagen, Denmark
Manoel Horta Ribeiro, Universidade Federal de Minas Gerais, Brazil
Joris Van Hoboken, Vrije Universiteit Brussels, Belgium
Veronique Hoste, University of Ghent, Belgium
Dirk Hovy, Bocconi University, Italy
Dan Jurafsky, Stanford, United States of America
George Kennedy, Intel, United States of America
Neža Kogovšek Šalomon, Peace Institute, Slovenia
Els Lefever, University of Ghent, Belgium
Chuan-Jie Lin, National Taiwan Ocean University, Taiwan
Nikola Ljubešić, Jožef Stefan Institute, Slovenia
Elizabeth Losh, William and Mary, United States of America
Prodromos Malakasiotis, StrainTek, Greece
Shervin Malmasi, Harvard University, United States of America
Diana Maynard, University of Sheffield, United Kingdom

Kathleen McKoewn, Columbia University, United States of America
Mainack Mondal, Max Planck Institute for Software Systems, Germany
Hamdy Mubarak, Qatar Computing Research Institute, Qatar
Smruthi Mukund, A9.com Inc, United States of America
Kevin Munger, New York University, United States of America
Preslav Nakov, Qatar Computing Research Institute, Qatar
Chikashi Nobata, Apple, United States of America
Gustavo Paetzold, Federal University of Technology - Paraná, Brasil
John Pavlopoulos, StrainTek, Greece
Daniel Preoțiuc-Pietro, Bloomberg, United States of America
Michal Ptaszynski, University of Duisburg-Essen, Germany
Vladan Radosavljevic, OLX Group, Argentina
Georg Rehm, Deutsche Forschungszentrum für Künstliche Intelligenz, Germany
Björn Ross, University of Duisburg-Essen, Germany
Masoud Rouhizadeh, Stony Brook University & University of Pennsylvania, United States of America
Niloofar Safi Samghabadi, University of Houston, United States of America
Christina Sauper, Facebook, United States of America
Xanda Schofield, Cornell, United States of America
Caroline Sinders, Wikimedia Foundation, United States of America
Maite Taboada, Simon Fraser University, Canada
Dennis Yi Tenen, Columbia University, United States of America
Dimitrios Tsarapatsanis, University of Sheffield, United Kingdom
Ingmar Weber, Qatar Computing Research Institute, Qatar
Amanda Williams, University of Bristol, United Kingdom
Michael Wojatzki, University of Duisburg-Essen, Germany
Lilja Øvrelid, University of Oslo, Norway

Invited Speakers:

Mikki Kendall, Writer and Diversity Consultant
Maryant Fernandez Perez, Senior Policy Advisor of European Digital Rights

Table of Contents

Conference Program

Wednesday, October 31, 2018

9:00–9:15 ***Opening Remarks***

9:15–10:00 *Invited Speaker*
Mikki Kendall

10:00–10:45 *Invited Speaker*
Maryant Fernandez Perez

10:45–11:15 ***Break***

11:15–12:15 ***Panel***

12:15–12:45 ***Plenary Discussions***

12:45–14:00 ***Lunch***

14:00–15:30 ***Posters***

14:00–15:30 *Social contagion in ethnic abusive swearing during periods of increased terrorist activity*
Christoph Spörlein

14:00–15:30 *Neural Character-based Composition Models for Abuse Detection*
Pushkar Mishra, Helen Yannakoudakis and Ekaterina Shutova

14:00–15:30 *Hate Speech Dataset from a White Supremacy Forum*
Ona de Gibert, Naiara Perez, Aitor García Pablos and Montse Cuadros

14:00–15:30 *A Review of Standard Text Classification Practices for Multi-label Toxicity Identification of Online Content*
Isuru Gunasekara and Isar Nejadgholi

Wednesday, October 31, 2018 (continued)

14:00–15:30 *Predictive Embeddings for Hate Speech Detection on Twitter*
Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown and Susan McGregor

14:00–15:30 *Challenges for Toxic Comment Classification: An In-Depth Error Analysis*
Betty van Aken, Julian Risch, Ralf Krestel and Alexander Löser

14:00–15:30 *Aggression Detection on Social Media Text Using Deep Neural Networks*
Vinay Singh, Aman Varshney, Syed Sarfaraz Akhtar, Deepanshu Vijay and Manish Shrivastava

14:00–15:30 *Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying*
Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini and Enrico Piras

14:00–15:30 *Improving Moderation of Online Discussions via Interpretable Neural Models*
Andrej Švec, Matúš Pikuliak, Marian Simko and Maria Bielikova

14:00–15:30 *Aggressive language in an online hacking forum*
Andrew Caines, Sergio Pastrana, Alice Hutchings and Paula Buttery

14:00–15:30 *The Effects of User Features on Twitter Hate Speech Detection*
Elise Fehn Unsvåg and Björn Gambäck

14:00–15:30 *Interpreting Neural Network Hate Speech Classifiers*
Cindy Wang

14:00–15:30 *Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks*
Rijul Magu and Jiebo Luo

14:00–15:30 *Comparative Studies of Detecting Abusive Language on Twitter*
Younghun Lee, Seunghyun Yoon and Kyomin Jung

14:00–15:30 *Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs*
sima sharifirad, Borna Jafarpour and Stan Matwin

14:00–15:30 *Learning Representations for Detecting Abusive Language*
Magnus Sahlgren, Tim Isbister and Fredrik Olsson

Wednesday, October 31, 2018 (continued)

14:00–15:30 *Datasets of Slovene and Croatian Moderated News Comments*
Nikola Ljubešić, Tomaž Erjavec and Darja Fišer

14:00–15:30 *Cross-Domain Detection of Abusive Language Online*
Mladen Karan and Jan Šnajder

14:00–15:30 *Did you offend me? Classification of Offensive Tweets in Hinglish Language*
Puneet Mathur, Ramit Sawhney, Meghna Ayyar and Rajiv Shah

14:00–15:30 *Decipherment for Adversarial Offensive Language Detection*
Zhelun Wu, Nishant Kambhatla and Anoop Sarkar

14:00–15:30 *The Linguistic Ideologies of Deep Abusive Language Classification*
Michael Castelle

15:30–16:00 ***Break***

16:00–17:15 ***Plenary Discussions and Future Agendas***

17:15–17:30 ***Wrapup of Discussions***

17:30–17:45 ***Closing Remarks***