ACL 2019

The Third Workshop on Abusive Language Online

Proceedings of the Workshop

August, 1, 2019
Florence, Italy

Platinum Sponsor

UCLA

(Center for Critical Internet Research)

Google

Gold Sponsors

facebook

Silver Sponsor

ELEMENT AI

Bronze Sponsor

AYLIEN

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 978-1-950737-43-7

Introduction

The last few years have seen a surge in attention to various forms of abuse such as cyberbullying, hate speech, and scapegoating occurring on online platforms. At the same time, there has been a rise in interest in using Natural Language Processing (NLP) to address these issues at scaale. However, in order to develop robust, long-term solutions for this problem, we require perspectives from diverse disciplines ranging from psychology, law, gender studies, communications, and critical race theory. Our goal with the Abusive Language Workshop is to provide a platform to facilitate the interdisciplinary conversations and collaborations necessary to thoughtfully address the issue of abuse at scale.

Each year, we choose a theme for our workshop that guides the talks and panel discussions at the workshop. In previous years we focused on the policy aspect of online abuse and the stories and experiences of those who have received large amounts of online abuse. The themes do not limit the original research presented at the workshop, rather it helps frame the research presented through the lens of its potential to address the concerns of the theme. For this year, we have chosen to focus on *human content rating*, the practice of annotating and moderating data - an aspect which is often unspoken, assumed, and often forms the basis of the research conducted.

Human judgments of online abuse are critical for building training data for automated models, human-in-the-loop solutions that rely on crowd workers' ratings along with automated moderation, and embedding the evaluations of models into the cultural fabric. Thus, human ratings in the context of toxicity in language raise important questions around the various socio-cultural biases that affect those ratings, but also on the impact it has on the psychological safety of the raters themselves. In order to situate our conversation around this theme, we have confirmed four keynote speakers and panelists who are leading experts on content moderation, crowd work, and the impact of algorithmic solutions on people:

Katherine Lo, University of California, Irvine

Kat Lo is the Content Moderation Program Lead at Meedan and visiting researcher at the University of California, Irvine specializing in online moderation and harassment. Lo consults with technology, social media, and game companies on platform policy and enforcement. She also serves on the advisory board for nonprofits and advocacy organizations that focus on online harassment and mental health.

Safiya Noble, University of California, Los Angeles

Dr. Safiya Umoja Noble is an Associate Professor at UCLA in the Departments of Information Studies and African American Studies. She is the author of a best-selling book on racist and sexist algorithmic bias in commercial search engines, entitled Algorithms of Oppression: How Search Engines Reinforce Racism (NYU Press), which has been widely-reviewed in journals and periodicals including the Los Angeles Review of Books, featured in the New York Public Library 2018 Best Books for Adults(non-fiction), and recognized by Bustle magazine as one of "10 Books about Race to Read Instead of Asking a Person of Color to Explain Things to You". Safiya is the recipient of a Hellman Fellowship and the UCLA Early Career Award. Her academic research focuses on the design of digital media platforms on the internet and their impact on society. Her work is both sociological and interdisciplinary, marking the ways that digital media impacts and intersects with issues of race, gender, culture, and technology.

Sarah T. Roberts, University of California, Los Angeles

Roberts researches information work and workers, and is a leading global authority on "commercial content moderation," the term she coined to describe the work of those responsible *for making sure media content posted to commercial websites fit within legal, ethical, and the site's own guidelines and standards. She is frequently consulted on matters of policy, worker welfare, and governance related to moderation issues. She is a 2018 Carnegie Fellow and winner of the 2018 EFF Barlow Pioneer Award in recognition of her work on commercial content moderation. Her book, "Behind the Screen: Content Moderation in the Shadows of Social Media", will be released on June 25 2019 (Yale University Press).*

Nithum Thain, Jigsaw

Nithum Thain is a Software Engineer at Google Jigsaw. He works on the Conversation-AI effort that leverages Machine Learning technologies to help improve online conversation. Previously, Nithum was a Lecturer at Berkeley in NLP and a Postdoc at Simon Fraser University. Nithum holds a PhD in Algorithmic Game Theory from McGill University under the supervision of Dr. Adrian Vetta and an MBA from Oxford University as a Rhodes Scholar.

In addition, we will have a multi-disciplinary panel discussion where experts will debate and contextualize the major issues facing computational analysis of abusive language online, with a specific focus on human raters' work. This session will be followed by a poster session that will facilitate discussions around the research papers described in these proceedings.

Continuing the success of the past two workshops, we received 41 submissions describing high quality original research. In order to encourage submissions from social science researchers, we had a separate track for non-archival work. We conducted a rigorous review process where each paper received reviews from at least three researchers, at least one of which was a non-NLP researcher working on a field relevant to the paper. After review, we selected 21 papers to be presented at the workshop as posters. These include 14 long papers, 5 short papers, 1 demo paper, and 1 non-archival extended abstract. The authors of all accepted papers will be given an opportunity to expand their work into full journal articles to be considered for publication in a forthcoming special issue on abusive language online in the journal *First Monday*.

The accepted papers deal with a wide array of topics, both proposing new techniques to better detect abuse, as well as extending abuse detection to more languages and types of abuse. Three of the accepted papers bring social science perspectives on this issue, a significant improvement compared to last two iterations of the workshop. Our proceedings is also geographically diverse: representing work from 14 different countries: United States, United Kingdom, Italy, Canada, Netherlands, Australia, Indonesia, Portugal, Turkey, Germany, Croatia, Norway, India, and Switzerland (as per contact authors' affiliation).

With this, we welcome you to the 3rd Workshop on Abusive Language Online and look forward to the conversations and your participation.

Joel, Sarah, Vinod, and Zeerak

Organizers:

Vinodkumar Prabhakaran, Google
Sarah T. Roberts, University of California, Los Angeles
Joel Tetreault, Grammarly
Zeerak Waseem, University of Sheffield

Tunde Adefioye, KVS, Belgium
Mark Alfano, Delft University of Technology, Netherlands
Hind Almerekhi, Qatar Foundation, Qatar
Jisun An, Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar
Renata Barreto, Berkeley Law, United States
Elizabeth Belding, UC Santa Barbara, United States
Joachim Bingel, University of Copenhagen, Denmark
Peter Bourgonje, Potsdam University, Germany
Andrew Caines, University of Cambridge, United Kingdom
Pedro Calais, UFMG, Brazil, Brazil
Michael Castelle, University of Warwick, United Kingdom
Eshwar Chandrasekharan, Georgia Institute of Technology, United States
Wendy Chun, Brown University, United States
Isobelle Clarke, University of Birmingham, United Kingdom
Montse Cuadros, Vicomtech, Spain
Tyrus Cukavac, Bloomberg, United States
Aron Culotta, Illinois Institute of Technology, United States
Kareem Darwish, Qatar Computing Research Institute, Qatar Foundation, Qatar
Thomas Davidson, Cornell University, United States
Ona de Gibert Bonet, University of the Basque Country, Spain
Kelly Dennis, University of Connecticut, United States
Lucas Dixon, Jigsaw/Google, United States
Nemanja Djuric, Uber ATG, United States
Jacob Eisenstein, Georgia Institute of Technology, United States
Elisabetta Fersini, University of Milano-Bicocca, Italy
Darja Fišer, University of Ljubljana, Slovenia
Paula Fortuna, INESC TEC and Pompeu Fabra University, Portugal
Maya Ganesh, Leuphana University, Germany
Sara E. Garza, FIME-UANL, Mexico
Ryan Georgi, University of Washington, United States
Lee Gillam, University of Surrey, United Kingdom
Tonei Glavinic, Dangerous Speech Project, Spain
Genevieve Gorrell, University of Sheffield, United Kingdom
Erica Greene, The New York Times, United States
Alex Hanna, Google, United States
Mareike Hartmann, University of Copenhagen, Denmark
Christopher Homan, Rochester Institute of Technology, United States
Manoel Horta Ribeiro, Universidade Federal de Minas Gerais, Brazil
Hossein Hosseini, Department of Electrical Engineering, University of Washington, United States
Veronique Hoste, Ghent University, Belgium
Ruihong Huang, Computer Science and Engineering, Texas A&M University, United States
Dan Jurafsky, Stanford University, United States

Mladen Karan, University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, Croatia
Anna Kasunic, Carnegie Mellon University, United States
Christian Katzenbach, Humboldt Institute for Internet and Society, Germany
George Kennedy, Intel, United States
Ralf Krestel, Hasso Plattner Institute, Univteersity of Potsdam, Germany
Haewoon Kwak, Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar
Els Lefever, LT3, Ghent University, Belgium
Nikola Ljubešić, Jožef Stefan Institute, Slovenia
Elizabeth Losh, William and Mary, United States
Pranava Madhyastha, Imperial College London, United Kingdom
Walid Magdy, The University of Edinburgh, United Kingdom
Rijul Magu, Conduent, United States
Prodromos Malakasiotis, Athens University of Economics and Business Informatics Department, Greece
Shervin Malmasi, Harvard Medical School, United States
Puneet Mathur, Netaji Subhas Institute of Technology, India
Diana Maynard, University of Sheffield, United Kingdom
Yashar Mehdad, Airbnb, United States
Rada Mihalcea, University of Michigan, United States
Pushkar Mishra, Facebook, United Kingdom
Mainack Mondal, Cornell Tech, United States
Hamdy Mubarak, Qatar Computing Research Institute, Qatar Foundation, Qatar
Smruthi Mukund, Amazon, United States
Smaranda Muresan, Columbia University, United States
Isar Nejadgholi, Researcher, Canada
Iva Nenadic, Research Associate, Italy
Chikashi Nobata, Apple Inc., United States
Gustavo Henrique Paetzold, Federal University of Technology, Brazil
Alexis Palmer, University of North Texas, United States
Viviana Patti, University of Turin, Dipartimento di Informatica, Italy
John Pavlopoulos, Athens University of Economics and Business, Greece
Seeta Pena Gangadharan, London School of Economics and Political Science, United Kingdom
Christopher Potts, Stanford University, United States
Daniel Preoţiuc-Pietro, Bloomberg, United States
Michal Ptaszynski, Kitami Institute of Technology, Japan
Georg Rehm, DFKI, Germany
Julian Risch, Hasso Plattner Institute, University of Potsdam, Germany
Melissa Robinson, University of North Texas, United States
Carolyn Rose, Carnegie Mellon University, United States
Björn Ross, University of Duisburg-Essen, Germany
Paolo Rosso, Universitat Politècnica de València, Spain
Niloofar Safi Samghabadi, University of Houston, United States
Magnus Sahlgren, RISE, Sweden
Christina Sauper, Facebook, United States
Tyler Schnoebelen, Decoded AI, United States
Alexandra Schofield, Cornell University, United States
Marian Simko, Slovak University of Technology in Bratislava, Slovakia
Caroline Sinders, Convocation Design + Research, United States
Alison Sneyd, The University of Sheffield, United Kingdom
Jeffrey Sorensen, Jigsaw, United States

Rachele Sprugnoli, FBK / University of Trento, Italy
Linnet Taylor, Tilburg University, Netherlands
Achint Thomas, Embibe, India
Sara Tonelli, FBK, Italy
Dimitrios Tsarapatsanis, University of York, United Kingdom
Betty van Aken, Beuth University of Applied Sciences Berlin, Germany
Joris Van Hoboken, Vrije Universiteit Brussel, Belgium
Anna Vartapetiance, Surrey Centre for Cyber Security / University of Surrey, United Kingdom
Erik Velldal, University of Oslo, Norway
Rob Voigt, Stanford University, United States
Cindy Wang, Stanford University, United States
Ingmar Weber, Qatar Computing Research Institute, Qatar
Jacque Wernimont, Arizona State University, United States
Michael Wojatzki, Language Technology Lab, University of Duisburg-Essen, Germany
Helen Yannakoudakis, University of Cambridge, United Kingdom
Seunghyun Yoon, Seoul National University, Republic of Korea
Aleš Završnik, Institute of criminology at the Faculty of Law Ljubljana, Slovenia
Torsten Zesch, Language Technology Lab, University of Duisburg-Essen, Germany
Andrej Švec, Slovak University of Technology in Bratislava, Slovakia

Additional Reviewers:

Daniela Agostinho, University of Copenhagen, Denmark
Nanna Thylstrup, Copenhagen Business School, Denmark

Invited Speaker:

Katherine Lo, University of California, Irvine
Safiya Noble, University of California, Los Angeles
Sarah T. Roberts, University of California, Los Angeles

Panelists:

Katherine Lo, University of California, Irvine
Nithum Thain, Jigsaw

Table of Contents

Workshop Program

Thursday, August 01, 2019

9:00–10:35 **Session 1**

9:00–9:15 *Opening Remarks*

9:15–9:55 *Keynote 1: Katherine Lo*

9:55-10:35 *Keynote 2: Safiya Noble*

10:35–11:00 ***Break***

11:00–12:00 **Session 2: Panel Discussion**

12:00–13:30 ***Lunch***

13:30–15:10 **Session 3: Posters**

13:30–14:20 **Poster Session A**

Subversive Toxicity Detection using Sentiment Information
Eloi Brassard-Gourdeau and Richard Khoury

Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification
Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore and Goran Predovic

Detecting harassment in real-time as conversations develop
Wessel Stoop, Florian Kunneman, Antal van den Bosch and Ben Miller

Racial Bias in Hate Speech and Abusive Language Detection Datasets
Thomas Davidson, Debasmita Bhattacharya and Ingmar Weber

Automated Identification of Verbally Abusive Behaviors in Online Discussions
Srecko Joksimovic, Ryan S. Baker, Jaclyn Ocumpaugh, Juan Miguel L. Andres, Ivan Tot, Elle Yuan Wang and Shane Dawson

Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter
Muhammad Okky Ibrohim and Indra Budi

The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy
Qinlan Shen and Carolyn Rose

Pay "Attention" to your Context when Classifying Abusive Language
Tuhin Chakrabarty, Kilol Gupta and Smaranda Muresan

Challenges and frontiers in abusive content detection
Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale and Helen Margetts

A Hierarchically-Labeled Portuguese Hate Speech Dataset
Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner and Sérgio Nunes

A System to Monitor Cyberbullying based on Message Classification and Social Network Analysis
Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli and Serena Villata

14:20–15:10 Poster Session B

L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language
Hala Mulki, Hatem Haddad, Chedi Bechikh Ali and Halima Alshabani

At the Lower End of Language—Exploring the Vulgar and Obscene Side of German
Elisabeth Eder, Ulrike Krieg-Holz and Udo Hahn

Preemptive Toxic Language Detection in Wikipedia Comments Using Thread-Level Context
Mladen Karan and Jan Šnajder

Neural Word Decomposition Models for Abusive Language Detection
Sravan Bodapati, Spandana Gella, Kasturi Bhattacharjee and Yaser Al-Onaizan

Thursday, August 01, 2019 (continued)

A Platform Agnostic Dual-Strand Hate Speech Detector
Johannes Skjeggestad Meyer and Björn Gambäck

Detecting Aggression and Toxicity using a Multi Dimension Capsule Network
Saurabh Srivastava and Prerna Khurana

An Impossible Dialogue! Nominal Utterances and Populist Rhetoric in an Italian Twitter Corpus of Hate Speech against Immigrants
Gloria Comandini and Viviana Patti

"Condescending, Rude, Assholes": Framing gender and hostility on Stack Overflow
Sian Brooke

Online aggression from a sociological perspective: An integrative view on determinants and possible countermeasures
Sebastian Weingartner and Lea Stahel

15:10–15:40 ***Break***

Session 4

15:40–16:20 *Keynote 3: Sarah T. Roberts*

16:20–17:20 *Interdisciplinary Work Proposals & Plenary Discussion*

17:20–17:35 *Closing Remarks*