Special Issue on Detecting, Understanding and Countering Online Harms

Arkaitz Zubiaga, Bertie Vidgen, Miriam Fernandez, Nishanth Sastry

November 22, 2021

This editorial article introduces the OSNEM special issue on Detecting, Understanding and Countering Online Harms. Whilst online social networks and media have revolutionised society, leading to unprecedented connectivity across the globe, they have also enabled the spread of hazardous and dangerous behaviours. Such 'online harms' are now a pressing concern for policymakers, regulators and big tech companies. Building deep knowledge about the scope, nature, prevalence, origins and dynamics of online harms is crucial for ensuring we can clean up online spaces. This, in turn, requires innovation and advances in methods, data, theory and research design – and developing multi-domain and multi-disciplinary approaches. In particular, there is a real need for methodological research that develops high-quality methods for detecting online harms in a robust, fair and explainable way.

With this motivation in mind, the present special issue attracted 20 submissions, of which 8 were ultimately accepted for publication in the journal. These submissions predominantly revolve around online misinformation and abusive language, with an even distribution between the two topics. In what follows, we introduce and briefly discuss the contributions of these accepted submissions.

1 Online misinformation

Four of the submissions dealt with online misinformation and associated phenomena, which we describe next.

The first paper on online misinformation is titled "Understanding the characteristics of COVID-19 misinformation communities through graphlet analysis", by James R. Ashford, Liam D. Turner, Roger M. Whitaker, Alun Preece and Diane Felmlee. This paper investigates the networks formed in subreddits involving misinformation. By using local and global metrics for the measurements, it focuses on the use of substructures called "graphlets" to analyse the networks. The study finds that these networks have strongly defined local features relating to node degree, which is in turn suggested as a useful metric to detect the potential misinformation. Authors argue that this network-based methodology for detecting misinformation offers the flexibility of being applied globally, given its language independence.

The second paper focuses on fake news detection, which is titled “Check-It: A plugin for Detecting Fake News on the Web”, by Demetris Paschalides, Chrysovalantis Christodoulou, Kalia Orphanou, Rafael Andreou, Alexandros Kornilakis, George Pallis, Marios D. Dikaiakos and Evangelos Markatos. The paper describes the development of a web browser plugin that enables detection of fake news while browsing the Internet, with the benefit that it respects user privacy. The Check-It plugin combines a variety of sources into a pipeline for fake news identification. The paper describes promising performance scores, both in terms on quantitative measurements through accuracy as well as through qualitative measurement based on user experience.

The next paper focuses on the timely topic of the COVID-19 pandemic and its association with misinformation. The paper is titled “The Networked Context of COVID-19 Misinformation: Informational Homogeneity on YouTube at the Beginning of the Pandemic”, by Daniel Röchert, Gautam Kishore Shahi, German Neubaum, Björn Ross and Stefan Stieglitz. In investigating misinformation during the pandemic, this paper focuses on analysing its impact on the video-sharing platform YouTube. By analysing three months worth of data (January-March 2020), the paper looks at the network structure of videos and their associated comments, looking at their evolution over time. Findings from this study suggest that the spread misinformation was often part of heterogeneous discussion networks, generally involving other topics not related to the misinformation.

The final paper on misinformation is also linked to the COVID-19 pandemic, in this case however focused on conspiracy narratives, which the authors define as combinations “of disinformation, misinformation and rumour that are especially effective in drawing people to believe in post-factual claims.” The paper is titled “Disinformed Social Movements and Trust in Government: A Large-Scale Mapping of Conspiracy Narratives as Online Harms during the COVID-19 Pandemic”, and authored by Philipp Darius and Michael Urquhart. The study looks at the communities formed around two established conspiracy narratives on the Twitter platform, *i.e.*, the anti-vax and anti-5G narratives.

2 Abusive language online

The remaining four papers revolve around abusive language on online platforms and associated phenomena such as hate speech, particularly focusing on their impact on and through social media platforms.

The first paper investigates the modelling of the network dynamics of aggressive behaviour online. The paper is titled “Modeling Aggression Propagation on Social Media”, and authored by Chrysoula Terizi, Despoina Chatzakou, Evaggelia Pitoura, Panayiotis Tsaparas and Nicolas Kourtellis. To study the network dynamics of aggressive behaviour in social media, the authors focus on opinion dynamics, by modelling how aggression propagates from one user to another. Through experiments on Twitter datasets, the study shows promising results in the detection of cyberaggression by exploiting network dynamics.

The next paper is on hate speech detection and is titled "Data Expansion using Back Translation and Paraphrasing for Hate Speech Detection", and authored by Djamila Romaissa Beddiar, Md Saroar Jahan and Mourad Oussalah. The paper is motivated around the need for building effective hate speech detection models in situations where labelled data is scarce. They propose to perform data augmentation to increase the training data, which is achieved by using methods for back translation and paraphrasing existing training data to generate new instances. Experimenting on five datasets, the paper presents promising results compared to other state-of-the-art hate speech detection models.

The third paper looks at ways in which users in social media modify their language to circumvent the flagging of their posts as abusive, in this case through the use of emotes to avoid using offensive keywords that can be easily monitored. The paper is titled "Understanding and Identifying the Use of Emotes in Toxic Chat on Twitch", and authored by Jaeheon Kim, Donghee Yvette Wohn and Meeyoung Cha. Indeed, using visual language through emotes to avoid the use of offensive keywords can easily mislead algorithms while the message is still easily understood by humans. This study focuses on analysing the effect of this phenomenon on the Twitch platform and builds a classifier to improve the detection of those cases where keywords are replaced by emotes.

The final paper related to abusive language is the one titled "Empowering NGOs in Countering Online Hate Messages" and authored by Yi-Ling Chung, Serra Sinem Tekiroğlu, Sara Tonelli and Marco Guerini. The authors propose to go beyond developing algorithms for hate speech detection to also define strategies to fight hate speech through the injection of counter-messages. They propose the use of natural language processing methods to provide a systematic approach to hatred management. The tool they present in this paper offers suggestions of auto-generated messages to be used as counter-messages to confront hate speech. After testing the tool with 100+ NGO operators, the study finds positive results, significantly decreasing the time needed to craft counter-messages.

Arkaitz Zubiaga[1], Bertie Vidgen[2], Miriam Fernandez[3], Nishanth Sastry[4]

[1] Queen Mary University of London, UK

[2] Alan Turing Institute, UK

[3] Open University, UK

[4] University of Surrey, UK

Email addresses: a.zubiaga@qmul.ac.uk (A. Zubiaga), bvidgen@turing.ac.uk (B. Vidgen), miriam.fernandez@open.ac.uk (M. Fernandez), n.sastry@surrey.ac.uk (N. Sastry)