Summer 8-31-2020

Cyber Social Threats 2020 Workshop Meta-Report: COVID-19, Challenges, Methodological and Ethical Considerations

Ugur Kursuncu
University of South Carolina - Columbia

Yelena Mejova
ISI Foundation

Jeremy Blackburn
Binghamton University--SUNY

Amit Sheth
University of South Carolina - Columbia

Publication Info

Published in *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*, Summer 2020.

Cyber Social Threats 2020 Workshop Meta-report: COVID-19, Challenges, Methodological and Ethical Considerations

Ugur Kursuncu[1] **Yelena Mejova**[2] **Jeremy Blackburn**[3] **Amit Sheth**[1]
[1]AI Institute, University of South Carolina, SC, USA
[2]ISI Foundation, Turin, Italy
[3]Department of Computer Science at Binghamton University, NY, USA
kursuncu@mailbox.sc.edu, yelenamejova@acm.org, jblackbu@binghamton.edu, amit@sc.edu

Abstract

Online platforms have been increasingly misused by ill-intentioned actors, affecting our society, often leading to real-world events of social significance. On the other hand, recognizing the narratives related to harmful behaviors is challenging due to its complex and sensitive nature. The Cyber Social Threats Workshop 2020 aimed to stimulate research for the challenges on methodological and ethical considerations in developing novel approaches to analyze online harmful conversations, concerning social, cultural, emotional, communicative, and linguistic aspects. It provided a forum to bring together researchers and practitioners from both academia and industry in the areas of computational social sciences, social network analysis and mining, natural language processing, computational linguistics, human-computer interaction, and cognitive scientists to present their related, fundamental research and emerging applications, to exchange ideas and experiences, and to identify new opportunities for collaboration.

Keywords— Cyber Social Threats, Harassment, Extremism, Misinformation, Disinformation, Fake News, Hate Speech, Dataset, Tool, Social Media

Introduction

The role of social media as a prime, daily communication tool is coincident with a sharp rise in the misuse of online platforms, threatening our society in large. These platforms have been implicated for promoting misinformation (Pennycook et al. 2020; Ghenai and Mejova 2018) radicalization (Meleagrou-Hitchens, Alexander, and Kaderbhai 2017; Kursuncu et al. 2019a), harassment and cyber-bullying (Chatzakou et al. 2017; Wijesiriwardene et al. 2020), fake news, human trafficking, drug dealing, and gender-based violence among many others (Dwivedi et al. 2018; Purohit and Pandey 2019), with a significant impact on the well-being of individuals as well as communities. Further, malicious organizations (e.g., terrorist or extremist groups) have been exploiting social media for propagating misinformation and their pernicious propaganda materials to persuade individuals and eventually recruit them to their ideology, often leading to real-world events of social significance. For instance, since 2011, 300 Americans attempted or traveled to Syria and Iraq to join extremist groups[1], while the terror attacks were linked to online extremist content, consumed on social media by supporters living in the West (Frampton, Fisher, and Prucha 2017). In another vein, a 2017 Pew Research survey on online harassment found that 66% of adult Internet users claim to have observed online harassment and nearly 18% of Americans claim to have faced severe forms of harassment online such as physical threats, sexual harassment or stalking (Anderson 2017). In addition, as of 2019, 65% of social media users have been exposed to misinformation on major platforms[2], and indulge in online misinformation campaigns; millions of malicious accounts participated in online discussions in the 2016 U.S. presidential election spreading false information (Safadi et al. 2020). Apart from these examples of harmful communication content, excessive use of online tools, specifically among young individuals, can often have negative effects leading to addiction and mental health disorders.

The communications related to these problems are complex concerning their language and contextual characteristics, making recognition of such narratives challenging for researchers as well as social media companies (Kumar and Shah 2018; Wu et al. 2019; Kursuncu et al. 2019b). Most of the existing computational approaches fail to capture fundamental nuances in the language of these communications, owing to two prominent challenges: ambiguity and sparsity. Sole, data level bottom-up analysis has been unsuccessful in revealing the actual meaning of the content, resulting in false alarms. Considering the sensitive nature of these problems and its implications for individuals and communities, a potential solution requires reliable algorithms for modeling such communications.

To meet these challenges, this workshop aimed to stimulate research on understanding social, cultural, emotional, communicative, and linguistic aspects of harmful conversations on social media and developing novel approaches to analyze, interpret and understand them using computation.

[1]https://extremism.gwu.edu/isis-america

[2]CIGI-IPSOS GLobal Survey: Internet Security & Trust 2019. Social media, fake news & algorithms.

This workshop brought together researchers and practitioners in computer and social sciences from both academia and industry to discuss and exchange ideas on understanding social, psychological, cultural, communicative, and linguistic aspects of harmful content while leading the discussion on building novel computational methods to reliably detect, derive meaning, interpret, understand and counter them. Moreover, the workshop provided participants opportunities to identify new collaborations across disciplines.

Key Challenges

The workshop emphasized three key challenges: (1) detection and prediction of content, users, and communities, (2) countering harmful narratives, (3) ethical considerations, and handling bias. Modeling the use of language and detection of harmful content and users disseminating this content with their distinct roles in the respective community, is the first challenge in the pursuit of overcoming the cyber social threats. Further, preventing the spread of the content generated by malicious actors or organizations is essential; hence, the identification of tactics and strategic dissemination of the content in a social network and the development of countering narratives and their dissemination, emerges as another key challenge in countering such harmful activities online. On the other hand, ethical considerations in developing and potential deployment of computational models, mandates renewed attention because of its implications in the individual and societal life. Specific research topics and discussions in dealing with the online harmful content included; misinformation and disinformation (e.g., epidemics of fake news, images and videos, specifically on COVID-19), online extremism, harassment and cyberbullying, hate speech, gender-based violence, mental health implications of social media, ethical considerations on social media analytics, relationship of the social web and mainstream news media. The workshop received contributions spanning these challenges and topics employing quantitative and/or qualitative, analytical, theoretical approaches. Diverse nature of the received papers, such as research, dataset, tool demo, was the driving factor of the fruitful discussions.

Keynotes

The workshop had two invited keynote speakers who provided diverse insight into both academic and industrial perspectives on dealing with potentially harmful behaviors. Alexandra Olteanu from Microsoft Research, presented her talk, "Challenges to Measuring Objectionable Behavior Online by Humans and Machines," emphasizing the lack of understanding in the implications of computational models on online harmful behaviors, such as inadvertently reinforcing and amplifying them (Olteanu et al. 2019). She highlighted that the existing techniques focus on the issues that we already know about, while there is a lack of understanding in the techniques for preempting future issues that may not yet be on the researchers' radar. She shared reflections on the reasons that these issues continue to persist, such as the reliance on proxy measurements, data generation and collection processes, training, and testing datasets construction, and the design of evaluation metrics. Mikey Cohen from the Network Contagion Research Institute, shared his motivation behind an unexpected career change from being a leader at Netflix to fighting hate online. He described the challenges to deciphering and resolving the seemingly insurmountable and ever-changing problems online and how a collective effort between academia and the industry in Silicon Valley would address them.

Contributions

All submissions were reviewed by at least three multidisciplinary program committee (PC) members (21 in total) in the fields of computer science, information science, political science, sociology, psychology, and communications. Eight papers were accepted based on the quality of the rigor of analysis, results and presentation, and we provide a brief description for each contribution below.

COVID-19. The authors (Yang, Torres-Lugo, and Menczer 2020) analyze the flows of information on Twitter related to Covid-19 for the detection of bots and misinformation. The authors provide a comparison of spreading of news from low credibility media outlets with the New York Times and the Center for Disease. The authors' important findings include that the prevalence of misinformation is comparable to accurate information, and there appears to be a coordinated distribution of misinformation by groups of accounts.

(Lynnette and Yuan 2020) presents a preliminary analysis of a Telegram channel on COVID-19 from Singapore. The choice of platform, Telegram, is of interest for the research community, as studies on the content from closed chat rooms like Telegram groups are not many. The paper provides an analysis of several dimensions such as sentiment, misinformation, topics, and psychological features. Their findings include that government-identified misinformation is rare on group chats (0.05% of all messages) while participants of these communications were mostly approached towards such misinformation with scepticism (0.4% of all messages, almost 10 times more than messages discussing misinformation).

As per the emergence of the COVID-19 pandemic and relevant misinformation dissemination efforts on online platforms, (Shahi and Nandini 2020) provides a multilingual news dataset on COVID-19. The authors provide a detailed description of this new dataset, aiming to address the main challenges in identifying misinformation in online news. They also develop a classifier for the purpose of fact-checking.

The authors (Abrahams and Aljizawi 2020) investigate bots on Twitter during the COVID-19 pandemic, specifically for the region of the Middle East. They built a new dataset 149 pandemic-related keywords in Arabic, Turkish, and Persian, in order to analyze the prevalence of bots. They provide their finding on the prevalence of anomalous accounts as an indicator of bots, and suggest that suspected bot network(s) may not be necessarily due to state-backed interference.

Ethics. (Allison 2020) addresses ethical and methodological considerations in the collection and analysis of social media data, which include operationalizing research foci, recruiting participants, data collection, researcher well-being,

and research dissemination. The author provides a significant contribution to the foundational stage regarding the attempt to build an agreed-upon set of protocols for maintaining ethical standards and best practices when conducting research on (and with) toxic online communities. The study outlines pertinent methodological and ethical considerations around the study of antisocial, subversive, and toxic online communities and behaviors. The author also points out that most research in this field considers individual communities, platforms, and/or behaviors in isolation despite increasing evidence suggesting that toxic and subversive communities and behaviors can be interconnected and difficult to delimit or delineate. Finally, the author highlights an approach that would encourage the careful consideration of methodological and ethical challenges in the design, execution, and presentation of research in this direction. These points are also useful in the context of developing human subject protocols for researchers studying hate speech and radicalization online or similar problems.

Tool. The authors (Bevensee et al. 2020) introduce a new, easy-to-use toolkit to analyze posts for preliminary analysis and visualization of activity and trends on social media platforms. The tool enables end-user accessibility to large-scale analysis methods for social media analytics, for researchers, journalists, activists, and other non-computer scientists. They have integrated data from three social media platforms for actionable and real-time analyses analysis that offers scope for cross-platform comparison. The paper also describes two case studies to show the toolkit in use, to demonstrate the usefulness of the framework.

Fake News. (Wakamiya and Aramaki 2020) provides a multimodal approach for fake news detection addressing the question of how fake news travels over time after the initial spike, utilizing the temporal features extracted from social networking services (SNSs). The proposed approach encodes information from different domains and uses cross-attention to aggregate three different input sources. To encode temporal information, the authors convert the time series information of posts to infectiousness which captures the user behavior in the meta-level. Their ablation study between with and without temporal feature points to the importance of the temporal features.

Social Contagion. The authors (Ling and Stringhini 2020) measure reactions to the death of a Korean pop star Sulli on Twitter. They examine how social distance affects tone, timing and duration of engagement in public discussion of a tragedy. They focus on the differences in community membership (K-pop community/not, Asian/non-Asian) of users, and whether the users posted original content, retweeted, or were Twitter-verified. Their findings include that the membership affected the tone and duration of engagement on this topic, as the in-group shows a longer attention period and more frequent in-group interactions compared to the out-group.

Synthesis & Future Directions

At the end of the workshop, the participants attended a synthesis exercise session where they brainstormed ideas that are found most important, urgent, and high-impact for potential future research and collaborations. These ideas included efforts to mitigate doxxing (publishing private information about individuals online) and mental health problems induced by COVID-19. Further, the greater transparency and fairness specifically on the issues related to cyber social threats, such as bot detection, evaluation of therapies for those affected by hate speech, and the actionable application of research to policy changes were identified as pressing issues to be addressed in the future. The participants expressed their interest in collaborating on the identified problems and areas, as well as participating in future workshops to be organized.

Workshop Organization

The organizers of this workshop brought distinct interdisciplinary backgrounds and synergy, spanning multiple career stages including research institutes and academic departments.

Ugur Kursuncu. Postdoctoral Fellow, Artificial Intelligence Institute, University of South Carolina. SC, USA.
Yelena Mejova. Research Leader, ISI Foundation, Turin, Italy.
Jeremy Blackburn. Assistant Professor, Department of Computer Science at Binghamton University. NY, USA.
Amit Sheth. Founding Director, Artificial Intelligence Institute, University of South Carolina. SC, USA.

Acknowledgement

We thank our workshop program committee members[3] for their helpful reviews and support.

References

Abrahams, A., and Aljizawi, N. 2020. Middle east twitter bots and the covid-19 infodemic. In *International Workshop on Cyber Social Threats 2020. ICWSM 2020.*

Allison, K. R. 2020. Navigating negativity in research: Methodological and ethical considerations in the study of antisocial, subversive and toxic online communities and behaviours. In *International Workshop on Cyber Social Threats 2020. ICWSM 2020.*

Anderson, M. 2017. Key takeaways on how americans view and experience online harassment. *Pew Research Center*.

Bevensee, E.; Aliapoulios, M.; Dougherty, Q.; Baumgartner, J.; McCoy, D.; and Blackburn, J. 2020. Smat: The social media analysis toolkit. In *International Workshop on Cyber Social Threats 2020. ICWSM 2020.*

Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, 13–22. ACM.

Dwivedi, Y. K.; Kelly, G.; Janssen, M.; Rana, N. P.; Slade, E. L.; and Clement, M. 2018. Social media: The good, the bad, and the ugly. *Information Systems Frontiers* 20(3):419–423.

[3] http://CySoc.aiisc.ai/

Frampton, M.; Fisher, A.; and Prucha, N. 2017. The new netwar. *Policy Exchange: Westminster, London*.

Ghenai, A., and Mejova, Y. 2018. Fake cures: user-centric modeling of health misinformation in social media. *Proceedings of the ACM on human-computer interaction* 2(CSCW):1–20.

Kumar, S., and Shah, N. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

Kursuncu, U.; Gaur, M.; Castillo, C.; Alambo, A.; Thirunarayan, K.; Shalin, V.; Achilov, D.; Arpinar, I. B.; and Sheth, A. 2019a. Modeling islamist extremist communications on social media using contextual dimensions: Religion, ideology, and hate. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–22.

Kursuncu, U.; Gaur, M.; Lokala, U.; Thirunarayan, K.; Sheth, A.; and Arpinar, I. B. 2019b. Predictive analysis on twitter: Techniques and applications. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*. Springer. 67–104.

Ling, C., and Stringhini, G. 2020. Examining the impact of social distance on the reaction to a tragedy a case study on sulli's death. In *International Workshop on Cyber Social Threats 2020. ICWSM 2020.*

Lynnette, N. H. X., and Yuan, L. J. 2020. Is this pofma? analysing public opinion and misinformation in a covid-19 telegram group chat. In *International Workshop on Cyber Social Threats 2020. ICWSM 2020.*

Meleagrou-Hitchens, A.; Alexander, A.; and Kaderbhai, N. 2017. The impact of digital communications technology on radicalization and recruitment. *International Affairs* 93(5):1233–1249.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kiciman, E. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2:13.

Pennycook, G.; McPhetres, J.; Zhang, Y.; Lu, J. G.; and Rand, D. G. 2020. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31(7):770–780.

Purohit, H., and Pandey, R. 2019. Intent mining for the good, bad, and ugly use of social web: concepts, methods, and challenges. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*. Springer. 3–18.

Safadi, H.; Li, W.; Rahmati, P.; Soleymani, S.; Kursuncu, U.; Kochut, K.; and Sheth, A. 2020. Curtailing fake news propagation with psychographics. *Available at SSRN 3558236*.

Shahi, G. K., and Nandini, D. 2020. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. In *International Workshop on Cyber Social Threats 2020. ICWSM 2020. arXiv:2006.11343*.

Wakamiya, T. M. S., and Aramaki, E. 2020. Fake news detection using temporal features extracted via point process. In *International Workshop on Cyber Social Threats 2020. ICWSM 2020. arXiv:2007.14013*.

Wijesiriwardene, T.; Inan, H.; Kursuncu, U.; Gaur, M.; Shalin, V. L.; Thirunarayan, K.; Sheth, A.; and Arpinar, I. B. 2020. Alone: A dataset for toxic behavior among adolescents on twitter. *International Conference on Social Informatics (SocInfo 2020). arXiv:2008.06465*.

Wu, L.; Morstatter, F.; Carley, K. M.; and Liu, H. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter* 21(2):80–90.

Yang, K.-C.; Torres-Lugo, C.; and Menczer, F. 2020. Prevalence of low-credibility information on twitter during the covid-19 outbreak. In *International Workshop on Cyber Social Threats 2020. ICWSM 2020. arXiv:2004.14484v2*.