Cyber Social Threats 2021: AI, Covid-19 Vaccine, Detection and Countering Strategies

Ugur Kursuncu[1] **Jeremy Blackburn**[2] **Yelena Mejova**[3] **Megan Squire**[4] **Amit Sheth**[1]
[1]AI Institute, University of South Carolina, SC, USA
[2]Department of Computer Science, Binghamton University, NY, USA
[3]ISI Foundation, Turin, Italy
[4]Department of Computer Science, Elon University, NC, USA
kursuncu@mailbox.sc.edu, jblackbu@binghamton.edu, yelenamejova@acm.org, msquire@elon.edu, amit@sc.edu

Abstract

In recent years, online platforms have been utilized for promoting harmful content and behavior such as extremism, harassment, mis/disinformation, human trafficking, gender-based violence among others affecting our society, often leading to real-world events. Such content and behaviors are inherently complex, making the recognition of these narratives challenging for researchers as well as social media companies. The Cyber Social Threats (CySoc) Workshop 2021 aimed to facilitate a rich forum for researchers and practitioners from both academia and industry in the areas of computing and social science, to discuss novel avenues for research on interdisciplinary aspects of harmful communications on social media, to exchange ideas and experiences, and to identify new opportunities for collaboration. The second edition of this workshop was well-received and very successful in providing a platform to explore future directions, possible research ideas, and collaborations, as it was for the first edition last year (2020).

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

Introduction

Online platforms have been a prominent communication medium being used on a daily basis, which also introduced novel challenges due to their misuse by malicious actors and organizations. These cyber social threats often significantly impact the well-being of individuals as well as communities and our society at large, and they include online extremism (Kursuncu et al. 2019a; Aldera et al. 2021), cyberbullying, harassment (Golbeck 2018; Wijesiriwardene et al. 2020), fake news (Shu et al. 2019; Safadi et al. 2020), human trafficking, and gender-based violence among many others (Kursuncu et al. 2019b; 2021; Sheth, Shalin, and Kursuncu 2021; Dwivedi et al. 2018). The misuse of technology has been particularly rampant during the COVID-19 pandemic, as the efforts in the spread of misinformation on COVID-19 dramatically increased. Such malicious actors and organizations have effectively utilized technology to spread their propaganda to recruit them to their "mission" that often results in real-world events of social significance. For instance, white supremacy groups in the U.S. recruited young individuals through social media and similar means, with subsequent real-world acts of violence such as the Capitol Riot on January 6, 2021. On the other hand, 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). Their efforts intensified even further during the pandemic, persuading their targets to intentionally spread COVID-19 and risk their lives or carry out attacks on hospitals caring for infected patients (Kruglanski et al. 2020). In another vein, prior to the pandemic, as of 2019, 65% of social media users have been exposed to misinformation on major platforms, and indulge in online misinformation campaigns. A 2020 Pew Research survey on COVID-19 related misinformation found that 29% of adults in the U.S. reported that they believe misinformation on how the novel coronavirus originated. Recent research indicates that the prevalence of low-credibility content on COVID-19 is comparable with mainstream and reliable sources and bots are playing an important role in creating and amplifying such content (Yang, Torres-Lugo, and Menczer 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.

These online (*cyber*) *social* communications that are *threats* (Kursuncu et al. 2020) to the well-being of individuals and the society, are challenging to measure, detect and counter due to their multi-dimensional characteristics (Kursuncu et al. 2019a; Kursuncu 2018; Kursuncu et al. 2019b). Prior work fails to capture fundamental nuances to reveal the actual meaning of the content. As per the sensitive nature of these problems and its implications for individuals and communities, the proposed solutions require reliable and fair algorithms.

The CySoc workshop this year aimed to facilitate a rich forum for researchers and practitioners from both academia

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

and industry in the areas of computing and social science, to discuss novel avenues for research on interdisciplinary aspects of harmful communications on social media, to exchange ideas and experiences, and to identify new opportunities for collaboration. This workshop brought together researchers and practitioners in computing 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.

The workshop had three main themes: (i) methodological approaches for detection and prediction of content, users, and communities, for modeling the use of language and detection of harmful content and users disseminating this content with their distinct roles in the respective community; (ii) development of strategies for countering harmful narratives along with identification of tactics and strategic dissemination of harmful content in a social network employed by malicious actors or organizations, (iii) ethical considerations, and handling bias with their implications in the potential deployment of computational models. For all of these themes, Covid-19 related research and works have received particular attention due to its urgency.

Specific research topics and discussions in addressing the challenges on 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, human trafficking, ethical considerations on social media analytics. 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. Two demonstrations for applications on online extremism and Covid-19 vaccine adoption were timely as per their relevance to ongoing pressing challenges we face as society (see Section Demo Presentations). Further, the workshop featured a panel discussion on “The Role of AI in Countering Cyber Social Threats” that received particular attention from the participants (see Section Panel Discussion).

Keynotes

The workshop had two invited keynote speakers who discussed the online framing of real-world protests and recent advances in research beyond detection, from interdisciplinary perspectives. Brooke Foucault Welles from Northeastern University, presented her talk, ”The Battle for Baltimore: Networked Counterpublics and the Contested Framing of Urban Unrest” (Foucault Welles and Jackson 2019) with a case study around the protests in Baltimore, stating the significance of *#HashtagActivism* for potential intervention in popular discourse on racism and policing in the U.S. She explained how communications in this contested online network evolved as users constructed meaning and debated questions of protest and race. She shared their findings from the study as justice-oriented messages spread even in such contested networks and these networks incubate social justice messages and extreme and hateful ideas. She also emphasized that one’s action to engage and amplify helps drive attention to the issues that they care the most. Huan Liu from Arizona State University, discussed AI-enabled strategies to combat cyber social threats in his keynote. Specifically, he focused on two problems, disinformation/fake news and cyberbullying and discussed the challenges in detecting them. He shared findings from his group’s research as explainability and causality are essential in fast detection of fake news (Cui et al. 2019). Then, he described the next frontiers of countering strategies as mitigation of disinformation/fake news and bias on online platforms and how to develop novel algorithms and models accordingly. He highlighted causal analysis as an avenue that can mitigate bias, increase transparency and improve fairness (Cheng et al. 2021).

Panel Discussion

We hosted a panel discussion on ”The Role of AI in Countering Cyber Social Threats” featuring Joan Donovan from Harvard University, Fil Menczer from Indiana University and Alexandra Olteanu from Microsoft Research as panelists, and Aleszu Bajak from USA Today as the moderator. AI-enabled/augmented/based technologies and techniques will likely be increasingly used by social media companies to counter cyber social threats such as extremism, cyberbullying, harassment, misinformation, etc. We facilitated a discussion around different perspectives of this topic for challenges and opportunities from a computing perspective, social impact/implications perspective, and ethics perspective.

Contributions

We have received 13 submissions, all of which were reviewed by at least three multi-disciplinary program committee (PC) members (25 in total) in the fields of computer and information sciences, political science, law, sociology, psychology, crime science, international affairs and communications. Seven 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 Vaccine. The authors (Pierri et al. 2021) aim to understand the relationship between the online public conversations on Facebook and Twitter and Covid-19 vaccine acceptance in the real world by studying the spread of low-credibility and high-credibility news related to vaccines. They found that there is a non-negligible amount of spread of low-credibility information on both platforms about the COVID-19 vaccine in Italy, and YouTube is a potential source of misinformation. In particular, they highlight the most popular YouTube videos as a source of low-credibility information as they find harmful and misleading content about vaccines. While this study is limited to Italy, the methods are applicable to other regions of the world. The outcome of this work provides various avenues for future research directions, such as comparison of the vaccine-related conversations with public data (e.g., statistics on vaccination rates). The authors make their data available for the research community.

Toxicity. (Fortuna et al. 2021) empirically studies news articles with highly toxic comments and their relation to the news content. The authors use both news content and metadata to classify news as toxic utilizing GBDT and BERT as the classification model. In annotation of the data, they use Perspective API to first label 'comments' of a news as toxic or not, then label a news as 'toxicity associated' if it has more than the median number of toxic comments. While toxicity in the news articles has been a well-studied problem, this study brings a new perspective by analyzing the comments and their associated metadata. The dataset was also made public by the authors.

Misinformation Dissemination. The authors (Basu and Sen 2021) propose a graph-based framework for identifying and monitoring users that are sources of misinformation. They compare the proposed framework with prior work that is using epidemiological models, and show that the task of identifying such users is formally intractable. Then, they provide approximation algorithms to identify users who disseminate misinformation on social networks using suitably placed detection monitors. The experimental validation shows its applicability to practice. The critical highlights of their approach are; (i) scalability of the approximation algorithms to 32000 node graphs, (ii) how they can avoid assessing every node, without requiring knowledge of the dissemination dynamics in the underlying graph network of users.

Online User Migration. (Davies et al. 2021) proposes two mathematically principled approaches for understanding and analyzing user migration patterns on Reddit at macro- and micro-levels through spatio-temporal patterns. The authors use COVID-19 related data to qualitatively illustrate the results of the proposed approaches for modeling and predicting online user behaviors. While their experiments show the effectiveness of the proposed methods, the authors indicate coupling this modeling approach for the migration on Reddit with natural language processing as their future work.

Human Trafficking. The authors of this paper (Upadhayay, Lodhia, and Behzadan 2021) focus on human trafficking developing an automated analytics pipeline for aggregating and de-duplicating news articles related to human trafficking, identifying information on perpetrators and victims/survivors from within the text, and developing a question answering (QA) system to enable queries of the text about the people involved in each case. As human trafficking is understudied from a data science perspective, potential benefits of their approach is invaluable.

Organized Inauthentic Online Activities. The authors (Tatang et al. 2021) collect data from a crowd-working platform that offers fake 'likes' through manipulation campaigns in exchange with micro-payments. They found the majority of these campaigns were primarily on Facebook ($\sim$81%) followed by YouTube, Instagram and other social networks. They also created and advertised a campaign to identify these accounts that provide fake 'likes' to analyze the nature of such manipulation campaigns. Once these accounts were identified, the authors reached out to these accounts to survey their motivations, benefits and views about providing fake 'likes'. The authors provide a discussion around the implications of these campaigns on society, including politics and product reliability, and ethical aspects.

Anti-semitism Dataset. This dataset paper (Jikeli et al. 2021) provides a contribution by rigorous development of a novel annotated anti-semitic and non-anti-semitic dataset, which is critical to scientific progress. The authors particularly reveal given that there have been relatively few studies that have explored anti-semitism as a form of hate in social media through machine learning approaches.

Demo Presentations

The workshop hosted two interesting demonstrations on very timely issues, online extremism and Covid-19 vaccine adoption.

Welton Chang and Eric Curwin from Human Rights First, presented a platform, "Extremism Explorer", that collects and analyzes content from multiple social networks, allowing researchers to research online violent hate speech in real time. The system collects data from across social media, forums, and chat rooms online, and identifies violent hate speech. The Extremism Explorer is currently used by 50+ researchers and NGOs that investigate online extremism. The presenters also discussed the origins of the system and the future of the platform.

Matthew DeVerna from Indiana University presented a tool, "CoVaxxy", for visualizing the relationship between COVID-19 vaccine adoption and online (mis)information (DeVerna et al. 2021). In this demonstration, they demonstrated CoVaxxy, a web-based dashboard built by the Observatory on Social Media. Using the growing data set that powers the dashboard, they also presented preliminary results from a deeper investigation into the relationship between vaccine uptake/hesitancy and misinformation in U.S. states and counties.

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 understanding self-hate and characteristics of language of self-deprecation, investigating the content and network of actors that disseminate hate speech towards Jewish and Muslim communities. Further, for combatting human trafficking, the need to develop tools that will facilitate real-time data collection, detection and response capabilities. 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.

Jeremy Blackburn. Assistant Professor, Department of Computer Science at Binghamton University. NY, USA.

Yelena Mejova. Senior Research Scientist, ISI Foundation, Turin, Italy.

Megan Squire. Professor, Department of Computer Science at Elon University, NC, USA.

Amit Sheth. Founding Director, Artificial Intelligence Institute, University of South Carolina. SC, USA.

Acknowledgement

We thank our workshop program committee members[2] for their helpful reviews and support.

References

Aldera, S.; Emam, A.; Al-Qurishi, M.; Alrubaian, M.; and Alothaim, A. 2021. Online extremism detection in textual content: A systematic literature review. *IEEE Access* 9:42384–42396.

Basu, K., and Sen, A. 2021. Epidemiological model independent misinformation source identification. *CySoc Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM).*

Cheng, L.; Guo, R.; Shu, K.; and Liu, H. 2021. Causal understanding of fake news dissemination on social media.

Cui, L.; Shu, K.; Wang, S.; Lee, D.; and Liu, H. 2019. defend: A system for explainable fake news detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2961–2964.

Davies, C.; Ashford, J.; Espinosa-Anke, L.; Preece, A.; Turner, L. D.; Whitaker, R. M.; Srivatsa, M.; and Felmlee, D. 2021. Multi-scale user migration on reddit. *CySoc Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM).*

DeVerna, M. R.; Pierri, F.; Truong, B. T.; Bollenbacher, J.; Axelrod, D.; Loynes, N.; Torres-Lugo, C.; Yang, K.-C.; Menczer, F.; and Bryden, J. 2021. Covaxxy: A collection of english-language twitter posts about covid-19 vaccines. *arXiv preprint arXiv:2101.07694.*

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.

Fortuna, P.; Cruz, L. B.; Maia, R.; Cortez, V.; and Nunes, S. 2021. Toxicity-associated news classification: The impact of metadata and content features. *CySoc Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM).*

Foucault Welles, B., and Jackson, S. J. 2019. The battle for# baltimore: Networked counterpublics and the contested framing of urban unrest. *International Journal of Communication* 13:1699.

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

Golbeck, J. 2018. *Online harassment.* Springer.

Jikeli, G.; Awasthi, D.; Axelrod, D.; Miehling, D.; Wagh, P.; and Joeng, W. 2021. Detecting anti-jewish messages on social media. building an annotated corpus that can serve as a preliminary gold standard. *CySoc Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM).*

Kruglanski, A. W.; Gunaratna, R.; Ellenberg, M.; and Speckhard, A. 2020. Terrorism in time of the pandemic: exploiting mayhem. *Global Security: Health, Science and Policy* 5(1):121–132.

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.

Kursuncu, U.; Mejova, Y.; Blackburn, J.; and Sheth, A. 2020. Cyber social threats 2020 workshop meta-report: Covid-19, challenges, methodological and ethical considerations. *CySoc Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM).*

Kursuncu, U.; Purohit, H.; Agarwal, N.; and Sheth, A. 2021. When the bad is good and the good is bad: Understanding cyber social health through online behavioral change. *IEEE Internet Computing* 25(1):6–11.

Kursuncu, U. 2018. *Modeling the persona in persuasive discourse on social media using context-aware and knowledge-driven learning*. Ph.D. Dissertation, University of Georgia.

Pierri, F.; Tocchetti, A.; Corti, L.; Giovanni, M.; Pavanetto, S.; Brambilla, M.; and Ceri, S. 2021. Vaccinitaly: monitoring italian conversations around vaccines on twitter and facebook. *CySoc Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM).*

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.*

Sheth, A.; Shalin, V. L.; and Kursuncu, U. 2021. Defining and detecting toxicity on social media: Context and knowledge are key. *arXiv preprint arXiv:2104.10788.*

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

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 395–405.

Tatang, D.; Kreißel, P.; Sehring, M.; Quinkert, F.; Degeling, M.; and Holz, T. 2021. Likes are not likes a crowdworking platform analysis. *CySoc Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM).*

Upadhayay, B.; Lodhia, Z. A. M.; and Behzadan, V. 2021. Combating human trafficking via automatic osint collection, validation and fusion. *CySoc Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM).*

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.*

Yang, K.-C.; Torres-Lugo, C.; and Menczer, F. 2020. Prevalence of low-credibility information on twitter during the covid-19 outbreak. *CySoc Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM).*