Advances in AI for Safety, Equity, and Well-Being on Web and Social Media: Detection, Robustness, Attribution, and Mitigation

Srijan Kumar

Georgia Institute of Technology
srijan@gatech.edu

In the talk, I shall describe my lab's recent advances in AI, applied machine learning, and data mining to combat malicious actors (sockpuppets, ban evaders, etc.) and dangerous content (misinformation, hate, etc.) on web and social media platforms. My vision is to create a trustworthy online ecosystem for everyone and create the next generation of socially-aware methods that promote health, equity, and safety. Broadly, in my research, I have created novel *graph, content (NLP, multimodality), and adversarial machine learning* methods leveraging terabytes of data to detect, predict, and mitigate online threats. My interdisciplinary research innovates socio-technical solutions that I achieve by amalgamating computer science with social science theories. I am also passionate about putting my research into practice – my lab's models have been deployed on Flipkart, influenced Twitter's Birdwatch, and now being deployed on Wikipedia. My research has started a paradigm shift from the current slow and reactive approach against online harms to agile, proactive, and whole-of-society solutions. My talk will outline four thrusts of my research:

(1) Detection of harmful content and malicious actors across platforms, languages, and modalities: Going beyond the standard practice of studying "English text on Twitter", my research seeks to address the grand challenge of addressing fundamental problems rooted deeply across platforms (Micallef et al. 2022), languages (Verma et al. 2022b), and modalities (Verma et al. 2022b,c) (images, videos, text).

(2) Robustifying detection models against adversarial actors by predicting future malicious activities: My work is pushing the boundaries by developing the first adversarial learning techniques to proactively forecast adversary behavior to fool detection models. Next, we improve model robustness against manipulation. My lab has investigated the vulnerabilities of models deployed on some of the biggest platforms: Facebook's TIES bad actor detector (He, Ahamad, and Kumar 2021), Twitter's Birdwatch misinformation detector (Mujumdar and Kumar 2021), and Wikipedia's ban evasion (Niverthi, Verma, and Kumar 2022).

(3) Attributing the impact of harmful content and the role of recommender systems: My lab has created data-driven techniques to establish the impact of online harm in the real world (e.g., exacerbating mental health (Verma et al. 2022a)). Further, to understand the algorithmic impact on harm, my recent work has illustrated how adversaries can manipulate recommendation systems (Oh et al. 2022).

(4) Mitigation techniques to counter misinformation by professionals and non-expert crowds: My lab is building one of the first socio-technical counter-misinformation systems to simultaneously empower professionals (to enable faster, accurate, and focused fact-checking) and non-experts (Micallef et al. 2020).

References

He, B.; Ahamad, M.; and Kumar, S. 2021. Petgen: Personalized text generation attack on deep sequence embedding-based classification models. In *ACM SIGKDD*.

Micallef, N.; He, B.; Kumar, S.; Ahamad, M.; and Memon, N. 2020. The role of the crowd in countering misinformation: A case study of the COVID-19 infodemic. In *Big Data*. IEEE.

Micallef, N.; Sandoval-Castañeda, M.; Cohen, A.; Ahamad, M.; Kumar, S.; and Memon, N. 2022. Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts and Videos. In *AAAI ICWSM*.

Mujumdar, R.; and Kumar, S. 2021. HawkEye: a robust reputation system for community-based counter-misinformation. In *IEEE/ACM ASONAM*.

Niverthi, M.; Verma, G.; and Kumar, S. 2022. Characterizing, Detecting, and Predicting Online Ban Evasion. In *ACM Web Conference*.

Oh, S.; Ustun, B.; McAuley, J.; and Kumar, S. 2022. Rank List Sensitivity of Recommender Systems to Interaction Perturbations. In *ACM CIKM*.

Verma, G.; Bhardwaj, A.; Aledavood, T.; De Choudhury, M.; and Kumar, S. 2022a. Examining the impact of sharing COVID-19 misinformation online on mental health. *Scientific Reports*, 12(1): 1–9.

Verma, G.; Mujumdar, R.; Wang, Z. J.; De Choudhury, M.; and Kumar, S. 2022b. Overcoming Language Disparity in Online Content Classification with Multimodal Learning. In *AAAI ICWSM*.

Verma, G.; Vinay, V.; Rossi, R. A.; and Kumar, S. 2022c. Robustness of Fusion-based Multimodal Classifiers to Cross-Modal Content Dilutions. In *EMNLP*.