Detecting AI-Generated Content on Social Media

Attacking the Flow of Misinformation on X

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Maxwell Penders

12/13/2024

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Caitlin D. Wylie, Department of Engineering and Society

Introduction

In the silicon age, we have unprecedented tools for sharing information, beginning with the internet. The internet has not only contributed immensely to the world's economic growth, but it is also expected to further enhance our quality of life in the future (Manyika & Roxburgh, 2011; Rainie, 2019). Yet, as new information-creation technologies like large language models (LLMs) become widely accessible, the challenge of distinguishing truth from misinformation (unintentionally misleading false information) and disinformation (intentionally misleading false information) grows increasingly complex.

While there are arguments against misinformation from many different backgrounds, one simple case can be made from a pragmatic standpoint: it is physically impossible for an individual to be an expert in more than a handful of areas, let alone thousands. Yet we still need to make decisions in all sorts of domains—ranging from our own health and financial investments to the safety and ethical sourcing of the products we buy. In the latter case, we typically depend on reliable consumer reports, regulatory agencies, and investigative journalists to help us navigate complex global supply chains. However, if misinformation spreads about product safety or certifications, it becomes harder to distinguish between legitimate claims and falsehoods. This, in turn, undermines our ability to make educated purchases and reduces the expected value of our choices. As our information becomes less reliable, the certainty we can have in the outcomes of our decisions decreases (Duijf, 2021).

My two research papers, titled "Detecting AI-Generated Content on Social Media" and "Attacking the Flow of Misinformation on X," tackle this pressing issue from different perspectives. The technical research paper delves into developing algorithms that can detect content generated by LLMs, specifically online in the social media realm, aiming to strengthen our ability to identify potentially deceptive information, which could be the key to preventing malevolent actors in a time of crisis such as Covid-19. I plan to conduct a brief literature review, taking the methods that work best from different realms of AI content detection while attempting optimizations that make these models more feasible at the internet scale. Meanwhile, my STS research paper aims to understand human-non-human interactions related to the spread of mis- and dis-information on X and determine points within these networks where intervention can be used to disrupt this spread through a document analysis using an actor-network theoretical framework (Law, 1992).

By examining these two aspects, I aim to contribute to the larger effort of optimizing the spread of accurate information throughout society, ultimately supporting a well-informed global community. My research could be applied in both practical settings such as providing tooling for social media platforms to automatically flag AI-generated content and in political settings, where models could inform policy used to enhance the resilience of social networks against disinformation.

Technical Topic

The rapid rise in LLM-generated content poses significant challenges, with education and academia being among the first to respond due to students' early adoption of the technology (Paustian & Slinger, 2024). However, an additional area of need is on social media platforms, where AI-generated propaganda has been found to be highly persuasive (Goldstein et al., 2024). As the amount of AI-generated online content continues to grow, the harm to society increases. In academia, a recent analysis of over 950 thousand papers found that anywhere from 6.3% to

17.5% of papers, depending on field, included some modification by a LLM (Liang et al., 2024). Due to the extreme levels of damage that may be done by these fabricated papers – from the enactment of problematic policies as well as from diminishing the public's respect for research and science – the academic AI detection realm has significantly more research than that in social media. In fact, using off-the-shelf machine learning tools, scientists have been able to correctly determine if a sample essay was created by a human or an AI with an accuracy rate of 99% (Desaire et al., 2023).

In general, successful AI writing detection algorithms have used specific probabilistic features of AI- vs. Human-written samples to determine if it was likely AI-written. For example, AI-written content may be far more likely to contain apostrophes used as single quotation marks as compared to a human's writing. As compared to an AI, it seems also as if a human writes far longer paragraphs. When comparing two otherwise similar papers, the one with more paragraphs should be more likely to be AI-generated. However, this illuminates a problem with currently existing AI-detection methods: the required number of words from writing samples is large when compared to the average word count of posts on most social media websites. In academia, this is less of an issue, as the word count of published articles is much more significant. While usually not as wordy as published work, assignments in education usually meet that invisible threshold for high-stakes submissions like term papers and long programming assignments. To tackle this issue, I aim to detect AI at the user-level by observing the entire history of a user's posts in order to increase the writing samples available to the detection algorithm.

Another policy-side issue for AI detection algorithms is setting the threshold for labeling a user as a bot. If the threshold likelihood is set too low, real people will be flagged as false positives, harming the reputation and enjoyability of the platform. If the threshold likelihood is set too high, orders of magnitude more bots will not be detected due to false negatives. The optimal choice here for a clean experience should be a threshold likelihood on the higher end, with a somewhat time-intensive appeal system. For a bot account, it is likely far easier just to create a new bot account when banned, but the odd human user who is flagged should be willing to simply wait 24 hours to have their account reinstated. With this system in mind, my technical work is to find features that work well and fit them onto the structure of social media.

STS Topic

On X, the impact of false information can be deadly. During the height of Covid-19, over 700 people died in Iran after ingesting toxic methanol as a result of misinformation initially pushed by Donald Trump (*Iran*, n.d.). In the United States, nearly three-quarters of the population were exposed to Covid-19 misinformation, with the more highly exposed groups being far less likely to get vaccinated for the virus (Neely et al., 2022). The cost of people not getting vaccinated is estimated to be more than 230 thousand lives in the United States alone (Jia et al., 2023).

Much like a virus, the spread of misinformation is contagious. On X, false information is far more likely to be retweeted than true information, leading to a spread of misinformation by genuine users (Vosoughi et al., 2018). Just like how the asymptomatic cases allowed Covid-19 to spread without much notice, the regular users who fall prey to disinformation accounts increase the impact of that disinformation far more than the bot account could on its own.

In the aim of combatting this widespread issue, I will answer the research question: How do human and non-human actors interact in the spread of misinformation and disinformation on X, and what points of intervention exist to disrupt these networks? The theoretical framework through which I will conduct this work is actor-network theory, which will trace the relationships between actors such as users, algorithms, content moderators, bots, and platform policies, revealing how misinformation and disinformation are enabled or countered by their interactions (Law, 1992).

My method to research available evidence is a document analysis, exploring how concepts like mis- and dis-information are discussed across academia, media, and public discourse (Mogalakwe, 2006). From this, I will identify key narratives, influential actors, and gaps in current understanding regarding misinformation on X by different stakeholders in the issue.

The first type of evidence which I will review are information spread models. From contagion models to Markov bridge models to agent-based simulations, a large amount of relevant research has been done, though these still require a translation from the mathematical model's formulation to a qualitative model appropriate for actor-network theory (Beskow & Carley, 2019; Jin et al., 2013; Luo et al., 2022). The second type of evidence I will review are studies evaluating the impact of specific countermeasures, such as the addition of the community notes feature (Chuai et al., 2024). This evidence will help me determine which countermeasures are the most effective, as well as providing helpful constants that fit beneficially into the information spread models. Finally, I will review relevant policy suggestions made by academics, think tanks, and law professionals. These will help inform my overall process and provide analysis by different types of experts. They also are a shortcut to understanding avenues for policy enactment, as just because some action could technically be taken does not mean it could feasibly be taken.

Misinformation on X poses a critical public health and societal threat. By adopting an actor-network theory framework, this research will trace the complex web of interactions that enable, constrain, and counteract the diffusion of harmful content. Through a thorough literature review—encompassing information spread models, countermeasure evaluations, and policy analyses—this work seeks to illuminate both the structural dynamics that foster disinformation and the strategic interventions capable of curtailing its reach. Ultimately, understanding these multifaceted relationships and their points of potential intervention is key to developing solutions that are both effective and implementable, thereby mitigating the human toll of misinformation on X and beyond.

Conclusion

AI-generated posts, and the misinformation contained in them pose significant threats to society, potentially more damaging than any fictional AI threat like a Terminator-style robot. Our society, particularly the US, relies heavily on the rapid transfer of information to educate, inform, and drive intellectual growth. As misinformation becomes more pervasive, its impacts seep into our daily lives, influencing the way we interact, make decisions, and shape our beliefs.

By addressing two aspects of this problem – improving the detection of AI-generated content and understanding how misinformation spreads – I hope to contribute to the development of a more informed and resilient community. My work is a small step toward mitigating the negative effects of misinformation and supporting a transition to a society where information is reliable, and trust is restored.

Beskow, D. M., & Carley, K. M. (2019). Agent Based Simulation of Bot Disinformation Maneuvers in Twitter. *2019 Winter Simulation Conference (WSC)*, 750–761. https://doi.org/10.1109/WSC40007.2019.9004942

Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). *Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter?* https://doi.org/10.1145/3686967

Desaire, H., Chua, A. E., Isom, M., Jarosova, R., & Hua, D. (2023). Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science*, *4*(6). https://doi.org/10.1016/j.xcrp.2023.101426

Duijf, H. (2021). Should one trust experts? *Synthese*, *199*(3), 9289–9312. https://doi.org/10.1007/s11229-021-03203-7

Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda? *PNAS Nexus*, *3*(2), 1–7. https://doi.org/10.1093/pnasnexus/pgae034

Iran: Over 700 dead after drinking alcohol to cure coronavirus. (n.d.). Al Jazeera. Retrieved September 22, 2024, from https://www.aljazeera.com/news/2020/4/27/iran-over-700-dead-after-drinking-alcohol-to-cure-coronavirus

Jia, K. M., Hanage, W. P., Lipsitch, M., Johnson, A. G., Amin, A. B., Ali, A. R., Scobie, H. M., & Swerdlow, D. L. (2023). Estimated preventable COVID-19-associated deaths due to non-vaccination in the United States. *European Journal of Epidemiology*, *38*(11), 1125. https://doi.org/10.1007/s10654-023-01006-3

Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on Twitter. *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, 1–9. https://doi.org/10.1145/2501025.2501027

Law, J. (1992). Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity. *Systems Practice*, *5*(4), 379–393. https://doi.org/10.1007/BF01059830

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. Y. (2024). *Mapping the Increasing Use of LLMs in Scientific Papers* (arXiv:2404.01268). arXiv. https://doi.org/10.48550/arXiv.2404.01268

Luo, R., Nettasinghe, B., & Krishnamurthy, V. (2022). Echo Chambers and Segregation in Social Networks: Markov Bridge Models and Estimation. *IEEE Transactions on Computational Social Systems*, *9*(3), 891–901. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2021.3091168

Manyika, J., & Roxburgh, C. (2011). *The great transformer: The impact of the Internet on economic growth and prosperity* (p. 10). McKinsey Gloabal Institute. https://www.mckinsey.com/~/media/mckinsey/industries/technology%20media%20and%20telecommunications/high%20tech/our%20insights/the%20great%20transformer/mgi_impact_of_internet_on_economic_growth.pdf

Mogalakwe, M. (2006). The Use of Documentary Research Methods in Social Research. *African Sociological Review / Revue Africaine de Sociologie*, *10*(1), 221–230.

Neely, S. R., Eldredge, C., Ersing, R., & Remington, C. (2022). Vaccine Hesitancy and Exposure to Misinformation: A Survey Analysis. *Journal of General Internal Medicine*, *37*(1), 179–187. https://doi.org/10.1007/s11606-021-07171-z

Paustian, T., & Slinger, B. (2024). Students are using large language models and AI detectors can often detect their use. *Frontiers in Education*, *9*. https://doi.org/10.3389/feduc.2024.1374889

Rainie, K. S., Janna Anderson and Lee. (2019, October 28). 4. The internet will continue to make life better. *Pew Research Center*. https://www.pewresearch.org/internet/2019/10/28/4-the-internet-will-continue-to-make-life-better/

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559