Fake News and AI: Fighting Fire with Fire?

Kimiz Dalkir [1]

[1]McGill University, Montreal, Canada

Abstract

While AI can be used to generate fake news and misinformation, often with malicious intent, AI can also be used to detect and, ideally, prevent this type of content from being shared so quickly and so widely. This position paper discusses some of the ways in which AI creates fake news, how AI can help combat fake news and why a hybrid (person-machine) approach is recommended.

Keywords

Fake news, AI-generated fake news, AI-based solutions, hybrid solutions

1. Introduction

Fake news is not a new phenomenon as propaganda and false marketing claims have been around for decades if not centuries. What is new today is that fake news can reach so many more people around the world almost instantaneously, primarily due to the use of social media. Whereas in the past only powerful people or major corporations could generate false claims in a convincing manner, today social media lets anyone create and disseminate fake news. The risks are great as fake news can have an impact on financial and health decisions. There has been a proliferation of fake news around the current pandemic for example, with some seeking to profit by selling fake COVID-19 cures while others add fuel to anti-vaccine movements.

At the same time, the sphere of influence, that is to say, the people who can influence us, has shrunk to increasingly small "filter bubbles" or echo chambers. We tend to listen to and believe people who are most like us – our family and friends. This creates a problem of validity as we are listening less to authoritative sources but it also means we tend to interact with people who hold similar views to our own. [1]

Artificial intelligence has been very effective in the creation and dissemination of fake news, alternate facts and misinformation. At the same time, AI may be the best defense against this type of content, even if AI was used to generate it. The old adage of "fight fire with fire" is a good analogy to use for this defense. When a forest fire is spreading quickly out of control and devouring vast amounts of forestland, the best way to stop it is to purposefully set a smaller, controlled fire in its path. When the two meet, the fire is stopped at this point and it can no longer spread. Could this be a good metaphor to use to combat AI-generated fake news?

2. How is AI used to create and spread fake news?

There are a number of effective AI tools used to generate fake content and new ones are being added every day. These include smart tools that allow computers to pose as humans and help manipulate the public conversation [e.g. 2] text-generating AI that produces synthetic text ("readfakes") as well as

AIofAI'21: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, Montreal, CA
EMAIL: kimiz.dalkir@mcgill.ca (K. Dalkir)

bots that can quickly and inexpensively create deep fakes [3]. These technologies keep on improving, making it more difficult to identify let alone prevent fake content.

In the past, fake news was more text-centric but today it is just as likely to be multimedia (images and video). For example, it used to be that we could just search for duplicate photos on user profiles but Facebook had to manually take down a great many profiles that had AI-generated photos. They could not be easily detected as they are not exact duplicates, clearly demonstrating that AI-created fake content has become much more sophisticated [3].

Today there are even more dangerous AI bots that can wage propaganda wars and phishing campaigns even more effectively than humans can. This was shown as early as 2017 by [4] in which a company carried out an experiment to compare humans and AI software to see which could get users to click on more malicious Twitter links. The AI software easily won. The researchers trained their AI algorithm on the behaviors of social media users in order to create and send out their own AI-generated phishing bait.

Their AI system created and distributed more phishing tweets and were able to attain a higher success rate than the human team. The AI system sent out tweets to over 800 users at a rate of 6.75 tweets per minute and managed to lure 275 victims. In contrast, the human team sent out tweets to only 129 users at a rate of 1.08 tweets per minute and succeeded in tricking only 49 users. [2] also notes that AI-enabled computational propaganda and disinformation ar also proving to be quite effective and there are many more examples.

So AI is very good at deceiving human users by creating and disseminating fake content. Can AI also be just as good at detecting and ideally preventing the creation and sharing of fake content?

3. How is AI used to fight fake news?

Can we use AI to fight AI-generated fake news and misinformation? In 2018, Mike Zuckerberg predicted that in the next decade, AI would be the savior for the massive problems of scale that Facebook and others come up against when dealing with the global spread of junk content and manipulation [2].

Some examples include software that detects fake news using linguistic analysis [e.g. 5]; others that look at the similarity of the text in the headline and the article [e.g. 6]; and linguistic comparison of the language used in real news and deceptive news [e.g. 7]. Some AI researchers looked at sentiment analysis as most fake news uses very sensationalistic or emotional language [e.g. 8]. Others developed machine-learning tools that can detect bots by analyzing over 1000 features to identify it as a bot or a human being [2].

Machine learning can also help social media firms debunk false stories by fact checking and using evidence-based corrections to identify problematic content [2]. [9] note that people react emotionally to fake news using words such as “love” and “hate” quite frequently. These features in user reactions (e.g. posts, likes) that can then be used to identify potential fake news.

The dissemination pattern of fake news dissemination tends to be within echo chambers and these patterns can also be used to identify fake news. [3] discusses the growing number of tools that can detect fake content based on features such as the linguistic style while others can detect fake news by the way it spreads across the Internet or social media (the pattern of sharing can be different for valid vs. false content). In addition, the frequency of visual features differ between fake and real news pieces. The latter often contains more multimodal information such as images and videos.

More AI-based tools to combat fake news are being developed every day. However, it is unlikely that a purely AI or technology-driven approach will be sufficient to tackle the complexity of content – whether it be valid or false. Content understanding depends heavily on context and language - two areas where humans still outperform AI.

The more promising approaches in using AI to detect AI do not rely on AI alone. One example, proposed by [10] makes use of the nutritional information labels used on food products as a model. Instead of leaving it up to an intelligent system to designate content as true or fake, content labels could be automatically generated to inform the consumer on parameters such as authority, credibility, veracity, among others. The human reader then has the responsibility of “buying” this online content or not in the same way they would use a food label to decide whether or not it was good for them. Human

agency and decision-making are thus not compromised. Once again, this is the recommended hybrid approach that addresses the issues created by letting AI alone make such decision.

4. Recommendations

A hybrid human-AI approach is likely to offer the best solution to identifying and preventing fake news. [2] states that "AI is not perfect on its own." The corollary is that people are not perfect on their own either. Therefore, it is not going to be enough to "fight fire with fire" – AI will always do better if the human is in the loop too.

AI is not well equipped to evaluate statements for their truthfulness, but it can find signals, such as expressions of disbelief in the comment section [11]. In the hybrid Factmata system, humans are the experts who label the content used for the AI's training. Fake news tends to be more subtle now and only humans can detect the nuances. Satire, for example, continues to baffle AI-based systems. Humans are better equipped to understand the context, to apply everyday commonsense and to tease out what is fake as opposed to something that is true or that was not intended to deceive (e.g. overt satire vs. covert propaganda). Human expertise is needed to fully grasp the nature of the content and to then tag it accordingly. Factmata is training its AI to recognize politically biased content, false content, and hate speech using this hybrid human-machine approach [11].

Even with a hybrid human-machine approach, however, it is not enough to identify fake news. Nor is it enough to prevent further sharing and spreading of fake news. The ideal solution would be to prevent its creation and spread while at the same time using AI to help the consumers of content be in a better position decide what is valid and what is fake content. We also need to stop relying on fact-checks after the fact—that is, only after a false article has gone viral. We really need a better early warning system or, ideally, a way of preventing fake news from reaching users. The latter is sometimes referred to as inoculation strategies – either inoculating people so that they are better aware and more cautious and/or inoculating social media platforms so that fake content cannot be posted nor shared.

For example, a one-click function that is always present as an option on the screen to report suspected fake content would help "crowdsource" the truth. This was recently added by Facebook to allow users to report fake content. Facebook admitted, however, that the tool still relied on a large pool of human moderators to actually remove the fake content [2]. In fact, this type of user reporting can be additional data to be mined by machine learning systems to better detect fake news but also to better model the features that made users suspicious. Machine learning can then learn from these examples and begin to more proactively flag fake content.

[12] highlight some of the dangers in allowing AI alone as a means of combatting fake news. They emphasize that people need to be equipped with the means of assessing their own personal risks. They need to have the opportunity to reflect on a given piece of content and arrive at their own decisions as to whether it comes from a credible source or that is fake. In addition to potential loss of agency, the authors note many AI-based counter-measures could lead to violation of user privacy and open up another set of potential risks. For example, users may be asked to divulge personal information so that the AI system can build up a valid and useful model of them. As these intelligent systems can be quite persuasive, users may not be too careful about revealing personal information. The researchers propose conducting a form of "social welfare impact assessment" of such persuasive AI solutions in order to ensure ethical guidelines are respected.

"The plethora of different contexts in which false information flows online—everywhere from an election in India to a major sporting event in South Africa—makes it tricky for AI to operate on its own, absent human knowledge. But in the coming months and years it will take hordes of people across the world to effectively vet the massive amounts of content in the countless circumstances that will arise" [2]. Human expertise and artificial intelligence expertise complement one another very strongly.

"It is clear that more than one approach, technique or tool is needed to navigate effectively in a post-truth world. End users need to have greater awareness of the existence of fake news, alternative facts, and misinformation. Organizations need to develop information policies to deal with internal issues while countries need to develop laws that enforce concrete consequences for the deliberate dissemination of fake news. Last but not least, we need to deploy an effective set of tools to help us detect fake news" [1].

The possibilities for deception seem endless and while today's AI systems may not be ready to parse complicated claims independently or to make sophisticated decisions about truth, they can still be used to combat the onslaught of fake news. We don't need to get it perfect and it is better to start using it and let the AI learn and improve over time.

5. References

[1] Kimiz Dalkir and Rebecca Katz. (Eds). *Navigating fake news, alternative facts and misinformation in a post-truth world.* Hershey, PA: IGI Global.

[2] Samuel Woolley. *The Reality Game: How the Next Wave of Technology Will Break the Truth.* New York, NY: Hachette Book Group.

[3] Hannah Murphy. The new AI tools spreading fake news in politics and business. *The Financial Times Ltd.* Available online at: https://www.ft.com/content/55a39e92-8357-11ea-b872-8db45d5f6714.

[4] George Dvorsky. Hackers have already started to weaponized Artificial Intelligence. *Gizmodo online.* Available at: https://gizmodo.com/hackers-have-already-started-to-weaponize-artificial-in-1797688425.

[5] N. Conroy, Victoria Rubin, and Y Chen. Automatic deception detection: methods for finding fake news. In, *Proceedings, ASIST 2015*. November 6-10, 2015. St. Louis, MO, USA.

[6] A. Hanselowski, B Schiller, B., F. Caspelherr, D Chaudhuri, D, C Meyer & I. Gurevych, I. A retrospective analysis of the fake news challenge stance detection task. *arXiv preprint arXiv:1806.05180*. Available at: https://arxiv.org/pdf/1806.05180.pdf.

[7] H. Rashkin, e. Choi, J. Jang., S. Volkova, & Y. Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* (pp. 2931-2937).

[8] S. Sharma, & A. Jain, A. Role of sentiment analysis in social media security and analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), e1366.

[9] K. Shu, S. Wang, D. Lee, D., & H. Liu. Mining disinformation and fake news: concepts, methods, and recent advancements. In *Disinformation, Misinformation, and Fake News in Social Media*. Berlin, Germany: Springer. (pp. 1-19).

[10] Fuhr, Norbert, et al. "An information nutritional label for online documents." *ACM SIGIR Forum*, 51(3). New York, NY: ACM.

[11] E. Strickland. AI-human partnerships tackle" fake news": Machine learning can get you only so far-then human judgment is required. *IEEE Spectrum*, *55*(9), 12-13.

[12] Díaz Ferreyra, N.; Aïmeur, E.; Hage, H.; Heisel, M. and van Hoogstraten, C. (2020). Persuasion Meets AI: Ethical Considerations for the Design of Social Engineering Countermeasures. In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KMIS*.