ADMM Cybersecurity and Information Centre of Excellence

UPDATE ON THE INFORMATION DOMAIN

Issue 03/25 (March)

AI-Powered Content Moderation in Combating Web Disinformation

Introduction

1. In today's digital age, web disinformation has become one of the most pressing challenges in recent years. As disinformation continues to undermine trust, incite conflict, and mislead the public, effective tools and strategies are required to maintain trust and safety in digital spaces. AI-powered content moderation is a tool to combat web disinformation.

2. AI-powered content moderation uses artificial intelligence (AI) and natural language processing (NLP) to evaluate, interpret, and categorise potentially harmful online information. Using machine learning algorithms, these systems can automatically recognise and report harmful content, such as hate speech, misinformation, and inappropriate language, across different digital platforms. This technology provides a more efficient and rapid solution for monitoring large amounts of user-generated content, which would be difficult for human moderators to accomplish alone. The use of AI-powered content moderation tools helps to reduce the time taken to detect and remove inappropriate content, thereby decreasing the spread of disinformation to create a safer digital environment.

Effectiveness of AI-Powered Content Moderation

Scale and Speed

3. AI-powered content moderation is highly effective in handling large amounts of data quickly and efficiently. For example, once a piece of content is uploaded online, AI systems can start processing it in real time, scanning through vast amounts of data across multiple platforms. This enables AI to rapidly flag harmful or inappropriate content for removal, ensuring large quantities of new posts are monitored and screened without delay. AI thus serves as a vital barrier against the rapid spread of misleading narratives.

4. For example, YouTube has experienced significant challenges in moderating hate speech and violent extremism on its site due to the massive number of video footage produced every minute. As part of removing inappropriate content on its platform, YouTube launched a content management policy in 2017 that included the use of machine learning tools to identify potentially hazardous content to augment existing human flagging of extremist content for human review. In 2022, through the use of an AI-driven approach, YouTube "took down 5.6 million videos that violated the platform's community guidelines, such as hate speech, harassment, child endangerment, and violent or graphic materials." Without AI, content moderation on this scale would be nearly impossible.

5. YouTube's use of machine learning algorithms to detect harmful content exemplifies how AI systems are continuously learning from the vast amount of data, adapting to new patterns and improving. These autonomous systems can assess patterns in content and behaviour to detect possible hate speech more effectively than traditional methods such as manual moderation. As YouTube grows, the ability to expand content moderation while maintaining quality becomes increasingly important. AI's ability to handle an increasing volume of contents means that YouTube can retain its content moderation standards even as the platform grows globally.

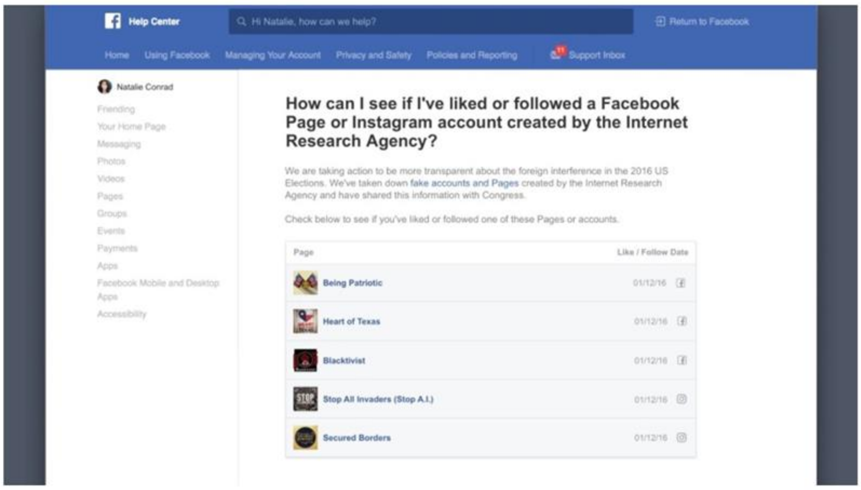
Excelling at identifying emerging trends and spotting recurring patterns of deception

6. AI can identify patterns of misinformation and disinformation more effectively than human moderators. AI excels at recognising

patterns, including identifying the characteristics of misinformation and disinformation. It can assess a variety of factors, including the spread of particular narratives and inconsistencies in the content itself. For instance, it may detect when a video claims to be from a reliable source, but the facts given are disputed by reliable sources in the same content, thus signalling potential misinformation. Over time, AI models learn to recognise emerging trends in misinformation and disinformation, even if the content is new or unique. This allows for the early detection of misinformation and disinformation before it spreads. AI can also analyse user behaviour patterns to identify coordinated attempts to propagate false information.

7. For example, following the 2016 US Presidential Election, the Russian Internet Research Agency (IRA) used a network of fake accounts and bots to spread political disinformation on platforms such as Facebook and Instagram. These accounts spread false narratives, incited political division, and promoted contentious content. This included posts that exaggerated political scandals, misrepresented candidates' positions, and conspiracy theories such as election fraud. Facebook employed AI tools to determine the scope of the IRA's activities (Figure 1). Its AI systems were able to identify fake accounts as well as automated bot activity used to amplify content related to the election.

Figure 1: Facebook created a new tool for users to check if they had liked or followed a Facebook page or Instagram account created by the IRA.



How can I see if I've liked or followed a Facebook Page or Instagram account created by the Internet Research Agency?

We are taking action to be more transparent about the foreign interference in the 2016 US Elections. We've taken down fake accounts and Pages created by the Internet Research Agency and have shared this information with Congress.

Check below to see if you've liked or followed one of these Pages or accounts.

Page	Like / Follow Date
Being Patriotic	01/12/16
Heart of Texas	01/12/16
Blacktivist	01/12/16
Stop All Invaders (Stop A.I.)	01/12/16
Secured Borders	01/12/16

8. The IRA employed bots to magnify disinformation by automating content sharing. Detecting bot activity manually is extremely difficult on a large scale. AI-powered solutions designed to detect and flag automated activities such as repetitive material and non-human interaction patterns provides a useful tool. AI algorithms can effectively differentiate between real human participation and automated amplification, thus allowing platforms such as Facebook to detect and prevent coordinated disinformation campaigns in real time.

AI's limited understanding of deeper context

9. While AI is an effective tool for moderating content at scale and speed, its lack of contextual awareness is problematic. AI algorithms are mostly based on patterns and pre-defined rules; thus, they are unable to understand the full context in the same way humans can. Moreover, if the training data contains biases, the AI may misinterpret content based on the patterns learned from the training data. For example, if AI is predominantly trained on Western cultural norms, it may classify content from other cultures as inappropriate simply because it does not fall into the AI's predefined categories.

10. In an AP News article published in 2019, it was reported that "a new study shows that leading AI models are 1.5 times more likely to flag tweets written by African Americans as "offensive" compared to other tweets." This demonstrated that AI systems are unable to fully comprehend the cultural context or intent behind a tweet. It was further mentioned that "Algorithms and content moderators who graded the test data that teaches these algorithms how to do their job, don't usually know the context of the comments they're reviewing". This highlighted the detachment of AI systems from the real-world context in which the content is being generated. Without context, AI can only rely on patterns or surface-level content analysis, and fail to determine the underlying tone or social dynamics that influence how a comment is perceived by its intended audience.

11. AI-powered content moderation is also substantially weaker when it comes to images and videos. This gap in AI capabilities has raised concerns as content that requires nuanced interpretation, such

as photographs and videos frequently fall through the cracks. The limitations of AI to deal with such content, combined with the pressure imposed on human moderators to oversee the massive volume of flagged material, contributed to a lawsuit filed by Facebook moderators against Meta. It has been reported in a news article that "The moderators from Kenya and other African countries were tasked from 2019 to 2023 with checking posts emanating from Africa and in their own languages but were paid eight times less than their counterparts". Additionally, "The images and videos including necrophilia, bestiality and self-harm caused some moderators to faint, vomit, scream and run away from their desks", underscoring the significant emotional and psychological toll that content moderation may have on employees.

12. In this case, AI overlook the emotional or psychological impact of the images, making it an inadequate alternative for human judgement in situations involving highly disturbing content, such as depictions of violence, where moderators must assess not only the content but also its potential psychological harm. Images communicate more than what is visible to the eye; visuals can evoke emotions and thoughts. Our brains will interpret these images according to our beliefs, values and inherent biases. AI does not have the ability to fully understand these dimensions, thus making human judgement vital in sensitive or disturbing situations. The psychological burden that human moderators faced is a direct result of AI's limits in moderating certain types of content.

Conclusion

13. The rapid development of AI presents challenges for AI-powered content moderation when it comes to misinformation and disinformation. Malicious actors are constantly refining their tactics to exploits AI weaknesses. For example, they may utilise AI-generated accounts, and targeted AI algorithms to spread misinformation and disinformation more efficiently. AI moderation systems thus need to stay ahead of these evolving tactics, which requires constant updates to the models. AI systems themselves can also become targets of AI manipulation, resulting in an ongoing struggle for change.

14. While AI may detect harmful content, it cannot address the root causes or provide nuanced solutions such as offering corrections. Therefore, human intervention is still required to ensure that mistakes or misinterpretations are addressed appropriately, particularly when context play an important role in defining what is considered detrimental. For AI-powered content moderation to be effective and widely trusted, platforms must increase transparency and provide channels for users whose content has been identified for removal to appeal. AI systems should also be trained on more diverse, representative datasets and constantly enhanced to keep up with the complexities of online misinformation and disinformation. It is also crucial to strike a balance between AI-driven automation and human oversight. Creating more context-aware and culturally appropriate AI models will ultimately enhance country’s ability to effectively counter misinformation and disinformation. AI-powered content moderation can help combat misinformation and disinformation; however, it lacks in keeping up with AI-generated content without continual development and a collaborative strategy that leverages the strengths of both AI and human moderators.

. . . . .

CONTACT DETAILS

All reports can be retrieved from our website at www.acice-asean.org/resource/.

For any queries and/or clarifications, please contact ACICE at ACICE@defence.gov.sg.

Prepared by:
ADMM Cybersecurity and Information Centre of Excellence

REFERENCES

News Articles

1. 5 types of AI content moderation and how they work [Link: https://www.techtarget.com/searchContentManagement/tip/Types-of-AI-content-moderation-and-how-they-work]

2. Natural Language Processing: A Comprehensive Guide to its Applications and More [Link: https://datasciencedojo.com/blog/natural-language-processing-applications/]

3. The Four Rs of Responsibility, Part 1: Removing harmful content [Link: https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/]

4. The role of AI in improving content moderation in social media [Link: https://chekkee.com/the-role-of-ai-in-improving-content-moderation-in-social-media/]

5. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior [Link: https://www.nature.com/articles/s41467-022-35576-9]

6. Facebook tool shows if users were duped by Russian propaganda [Link: https://www.irishtimes.com/business/technology/facebook-tool-shows-if-users-were-duped-by-russian-propaganda-1.3309497]

7. The Unseen Perils of AI: How Lack of Contextual Intelligence Entrenches Biases [Link: https://medium.com/@willabraczinskas/the-unseen-perils-of-ai-how-lack-of-contextual-intelligence-entrenches-biases-289f5ffba3cb]

8. The algorithms that detect hate speech online are biased against black people [Link: https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter]

9. Facebook’s parent company Meta and moderators suing it for $1.6 billion in Kenya agree to mediation [Link: https://apnews.com/article/kenya-facebook-content-moderators-meta-lawsuit-sama-5dca81fa5df9aa87886366945818dfa9]

10. More than 140 Kenya Facebook moderators diagnosed with severe PTSD [Link: https://www.inkl.com/news/more-than-140-kenya-facebook-moderators-sue-after-diagnoses-of-severe-ptsd]

11. The Growing Role Of AI In Content Moderation [Link: https://www.forbes.com/councils/forbestechcouncil/2022/06/14/the-growing-role-of-ai-in-content-moderation/]