Principles: Privacy, Human oversight and determination, Responsibility and accountability
Values: Human Rights
Stakeholders: Media, Private sector, Technical community

Improving the detection of misinformation with AI

Mark is a developer currently working on an algorithm that will be able to moderate the superfluous amount of hate speech, violence, and misinformation on a renown social media platform. He understands that while human moderators work extremely hard to maintain the site, they are subject to human error and also undergo a great deal of trauma confronting the extreme content they filter through day to day.

The AI is able to detect 88.8% of the hate speech that is removed thanks to its advancements in how its models understand and sort text. Though, a significant issue on the platform is that there is a fair amount of misinformation not only showing up on text or article links, but also on images and videos, which is difficult for the AI to detect. Hate speech and misinformation are now multimodal.

A new AI system put in place used to detect misinformation regarding Covid-19 called the SimSearchNet has been able to recognize both copies of the original image and the near duplicates where a few words have been modified.

This same system can also be used to detect hate speech in a very similar process. Independent fact-checkers are still required as part of the end-to-end system in order to determine hate speech content. The entire operation from human fact-checking to the SimSearchNet accounts for billions of images being checked on a daily basis.

The implementation of putting in place more AI systems to conduct content moderation has proven to be extremely helpful for purging the platform of its most heinous content. However, human fact-checkers are still part of this equation and are being exposed to extreme content that takes a significant toll on their mental health. If the site decides to continue using third parties to fact-check, they should ensure more psychological support, compensation, and better working conditions for their human fact-checkers.

- Social media platforms in the private sector need to strike a balance of using AI and human moderators. Using AI where possible will be cost-effective and reduce the amount of potentially psychologically harmful content humans will be subjected to. At the same time, human input can provide the moderation process a more comprehensive result with full cultural and emotional context.

- Developers in the technical community should refine and employ the use of image recognition technology. Computer vision allows for the rapid analysis and identification of visual content and is conducted with almost perfect accuracy, thereby alleviating the workload of human moderators.

Human oversight & determination, respect and protection of human dignity, responsibility & accountability.

Know more about this case:

- “Here’s how we’re using AI to help detect misinformation”, Facebook, https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation?utm_source=hootsuite&utm_medium=&utm_term=&utm_content=&utm_campaign=

Additional resources:

- “Hateful memes challenge and data set”, Facebook, https://ai.facebook.com/tools/hatefulmemes/