AI isn't smart enough yet to save us from fake news: Facebook users (and their bias) are key

23 September 2019, by Gianluca Demartini

On its own, human judgement can be subjective and skewed towards personal biases.

The information we encounter online everyday can be misleading, incomplete or fabricated.

Being exposed to "fake news" on social media platforms such as Facebook and Twitter can influence our thoughts and decisions. We've already seen misinformation interfere with elections in the United States.

Facebook founder Mark Zuckerberg has repeatedly proposed artificial intelligence (AI) as the solution to the fake news dilemma.

However, the issue likely requires high levels of human involvement, as many experts agree that AI technologies need further advancement.

I and two colleagues have received funding from Facebook to independently carry out research on a "human-in-the-loop" AI approach that might help bridge the gap.

Human-in-the-loop refers to the involvement of humans (users or moderators) to support AI in doing its job. For example, by creating training data or manually validating the decisions made by AI.

Our approach combines AI's ability to process large amounts of data with humans' ability to understand digital content. This is a targeted solution to fake news on Facebook, given its massive scale and subjective interpretation.

The dataset we're compiling can be used to train AI. But we also want all social media users to be more aware of their own biases, when it comes to what they dub fake news.

Humans have biases, but also unique knowledge

To eradicate fake news, asking Facebook employees to make controversial editorial decisions is problematic, as our research found. This is because the way people perceive content depends on their cultural background, political ideas, biases, and stereotypes.

Facebook has employed thousands of people for content moderation. These moderators spend eight to ten hours a day looking at explicit and violent material such as pornography, terrorism, and beheadings, to decide which content is acceptable for users to see.

Consider them cyber janitors who clean our social media by removing inappropriate content. They play an integral role in shaping what we interact with.

A similar approach could be adapted to fake news, by asking Facebook's moderators which articles should be removed and which should be allowed.

AI systems could do this automatically at a large scale by learning what fake news is from manually

annotated examples. But even when AI can detect "forbidden" content, human moderators are needed to flag content that is controversial or subjective.

A famous example is the Napalm Girl image.

The Pulitzer Prize-winning photograph shows children and soldiers escaping from a napalm bomb explosion during the Vietnam War. The image was posted on Facebook in 2016 and removed because it showed a naked nine-year-old girl, contravening Facebook's official community standards.

Significant community protest followed, as the iconic image had obvious historical value, and Facebook allowed the photo back on its platform.

Using the best of brains and bots

In the context of verifying information, human judgment can be subjective and skewed based on a person's background and implicit bias.

In our research we aim to collect multiple "truth labels" for the same news item from a few thousand moderators. These labels indicate the "fakeness" level of a news article.

Rather than simply collect the most popular labels, we also want to record moderators' backgrounds and their specific judgments to track and explain ambiguity and controversy in the responses.

We'll compile results to generate a high-quality dataset, which may help us explain cases with high levels of disagreement among moderators.

Currently, Facebook content is treated as binary—it either complies with the standards or it doesn't.

The dataset we compile can be used train AI to better identify fake news by teaching it which news is controversial and which news is plain fake. The data can also help evaluate how effective current AI is in fake news detection.

Power to the people

While benchmarks to evaluate AI systems that can detect fake news are significant, we want to go a step further.

Instead of only asking AI or experts to make decisions about what news is fake, we should teach social media users how to identify such items for themselves. We think an approach aimed at fostering information credibility literacy is possible.

In our ongoing research, we're collecting a vast range of user responses to identify credible news content.

While this can help us build AI training programs, it also lets us study the development of human moderator skills in recognizing credible content, as they perform fake news identification tasks.

Thus, our research can help design online tasks or games aimed at training social media users to recognize trustworthy information.

Other avenues

The issue of fake news is being tackled in different ways across online platforms.

It's quite often removed through a bottom-up approach, where users report inappropriate content, which is then reviewed and removed by the platform's employees..

The approach Facebook is taking is to demote unreliable content rather than remove it.

In each case, the need for people to make decisions on content suitability remains. The work of both users and moderators is crucial, as humans are needed to interpret guidelines and decide on the value of digital content, especially if it's controversial.

In doing so, they must try to look beyond cultural differences, biases and borders.

Provided by The Conversation