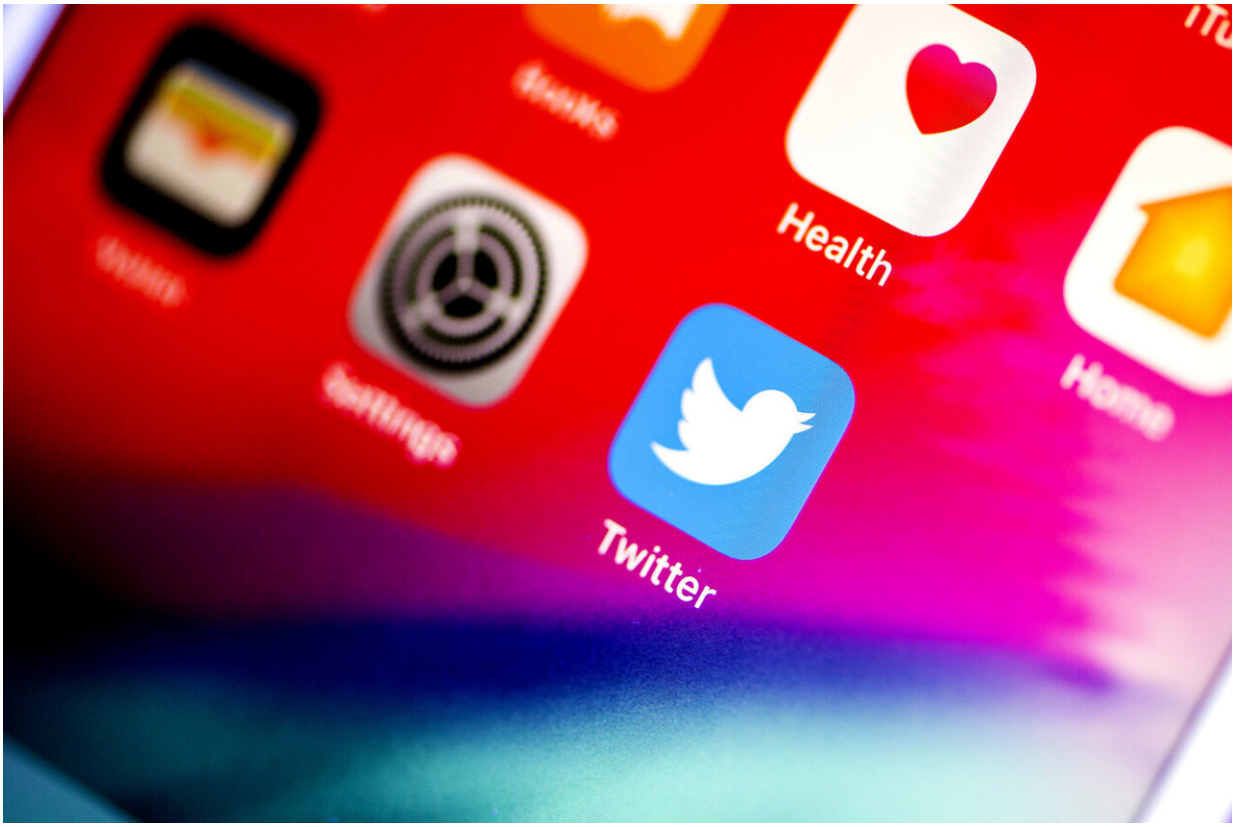
Should Facebook and Twitter review your posts before they're published?

April 20 2021, by Ian Thomsen



Social media channels are accused of enabling hate speech and misinformation that contributes to violence, including the Jan. 6 attack on the US Capitol. Credit: Matthew Modoono/Northeastern University

The day is coming when your posts to social media may travel through

checkpoints before the messages go public.

All of your posts to Facebook, Twitter, and other social platforms may be instantly examined by an artificial intelligence filter that roots out hate speech and misinformation. Some posts that have been flagged by artificial intelligence may then be reviewed by a human supervisor.

Such is the recommendation of Usama Fayyad, executive director for Northeastern's Institute of Experiential Artificial Intelligence, in response to the increasingly urgent desire for oversight of social media.

Fayyad believes social media filters are needed because the platforms have grown and scaled faster than they can be regulated—with the result that social channels are now being accused of enabling hate speech and misinformation that contributes to violence, including the Jan. 6 attack on the US Capitol.

"Social media needs to pass through some hurdle, mostly algorithmic and automated, but correctly augmented with the right human intervention," says Fayyad, a leader in artificial intelligence for three decades who founded Open Insights, Yahoo! Research Labs, Data Mining at Microsoft, and the Machine Learning Systems group at NASA's Jet Propulsion Laboratory. "The problem of misinformation is difficult but not impossible.

"We know that some stuff fits in this zone of uncertainty where we still need human judgment," he says "With a relevant feedback loop that appropriately leverages human judgment, the more we deal with these issues through human intervention, the more the system learns—and the smarter the technology gets."

Facebook has been developing artificial intelligence tools to root out "bad activity," said Mark Zuckerberg, chief executive of the social

media giant, in Congressional testimony as long ago as 2018. But will the biggest social channels—including Facebook, Twitter, and Instagram—commit to comprehensive changes to snuff out hate speech?

Fayyad is proposing a social media equivalent to the seven-second delay that TV networks use when covering sports and other live events. If a performer uses profanity, the delay enables censors to silence the foul language before it can be broadcast into homes.

"The social media companies have achieved their mission of transforming how we communicate," Fayyad says. "We've got to remember that all these companies are fairly new, and that we have not had the time to think about the issues as a society.

"This is an environment where something like social media can emerge and within two or three years become ubiquitous," Fayyad says. "We have not seen that in the history of humanity before. We need a new way of thinking around regulation and what it means and how quickly to evolve it."

The platforms' nascent efforts to develop filters are promising, says Christo Wilson, associate professor of computer science at Northeastern. There are a number of ways to sift out bad behavior: Reddit reminds users that they are responsible for community guidelines, notes Wilson, which reduces conflict, hate speech, and misinformation.

Wilson adds that the big social media companies have shown an ability to sniff out copyright violations and terrorist propaganda.

"The platforms are actually pretty good at catching that stuff, because they have very strong incentives to do so," says Wilson, who also directs Northeastern's bachelors program in cybersecurity. "So it's not like they can't do content moderation."

The Northeastern professors say that the goal of limiting bad actors will require a minimum of three steps:

Government regulation

There is bipartisan support for laws to curb the power of the biggest platforms.

Wilson argues that social media companies cannot be asked to set standards, police themselves, or take full responsibility for acting in the best interests of society when those measures may affect profits.

"At some point they're going to have to decide what's beyond the pale, and how it should be remedied by private actors," Wilson says of government regulators. "And I'm sure, immediately, someone will sue and take [the regulations] to the Supreme Court. Maybe it will fail, but we still have to try."

Fayyad believes comprehensive regulations will include heavy government fines for platforms that enable hate speech and misinformation. Such penalties will create additional incentives to invest in the necessary content moderation and technology.

Development of artificial intelligence

The evolution of search engines provides a helpful example.

"This is exactly how the popular search engines, including Google, emerged," Fayyad says. "They started out with a basic algorithm, and then they needed a lot of feedback with literally tens of thousands of editorial people reviewing search results. In the year 2000, it probably looked impossible: The search engines were nowhere near as good, the

problem was too hard, the web was growing fast.

"It's pretty amazing how far search technology has come, just by incorporating that feedback loop and that ability to capture learnings and continuously improve over time," Fayyad says. "So I am hopeful that technology can help. Smart monitoring is not an impossible problem."

Include humans in the loop

Regulations will help define the role of human oversight whenever a post is flagged by artificial intelligence.

"Whether you refuse to host it entirely, or it goes for human review, there's a bunch of normative issues there," Wilson says. "And then there will be issues with human review as well. But if we accept that social networks and social media are going to be huge, we have to have moderation systems that scale—and there's no denying that AI and machine learning is going to be part of that."

Wilson says the role of human review may hinge upon a new perspective on the platforms and their role in society.

"When the platforms describe themselves, they don't talk about content moderation," Wilson says. "They talk about being an open community. And that sets an expectation that people can do whatever the hell they want. So they need to more strongly acknowledge their role in shaping discourse."

Wilson believes the harm of bad intentions can be further limited if the big platforms are broken down into smaller channels that no longer will be able to influence large populations.

"The ability to spread [dangerous] messages and really impact the

mainstream is very much linked to the centralization of social media platforms," Wilson says. "This is one reason I also favor antitrust remedies: Smaller platforms have a little bit more latitude to do bad or to not moderate well, but there's less collateral damage."

Fayyad says most users won't notice the new measures.

"Ninety to 95 percent of the posts hopefully should not be blocked and should have no intervention," says Fayyad, who hopes that the relatively small percentage of harmful content will make it easier to catch.

Developing the technology will be expensive, Fayyad acknowledges. But he believes it will ultimately pay for itself.

"You create economic value," Fayyad says. "People will tend to use the social media that they trust and that they know is going to be safe. And that's going to be worth something."

Provided by Northeastern University