TechXplore

Elon Musk could roll back social media moderation, just as we're learning how it can stop misinformation

May 12 2022, by Harith Alani, Grégoire Burel and Tracie Farrell

Credit: Pixabay/CC0 Public Domain

The US$44 billion (£36 billion) purchase of Twitter by "free speech absolutist" Elon Musk has many people worried. The concern is the site will start moderating content less and spreading misinformation more, especially after his announcement that he would reverse the former U.S. president Donald Trump's ban.

There's good reason for the concern. Research shows the sharing of unreliable information can negatively affect the civility of conversations, perceptions of key social and political issues, and people's behavior.

Research also suggests that simply publishing accurate information to counter the false stuff in the hope that the truth will win out isn't enough. Other types of moderation are also needed. For example, our work on social media misinformation during COVID showed it spread much more effectively than related fact-check articles.

This implies some sort of moderation is always going to be needed to boost the spread of accurate information and enable factual content to prevail. And while moderation is hugely challenging and not always successful at stopping misinformation, we're learning more about what works as social media firms increase their efforts.

During the pandemic, huge amounts of misinformation was shared, and unreliable false messages were amplified across all major platforms. The role of vaccine-related misinformation on vaccine hesitancy, particularly, intensified the pressure on social media companies to do more moderation.

Facebook-owner Meta worked with factcheckers from more than 80 organizations during the pandemic to verify and report misinformation, before removing or reducing the distribution of posts. Meta claims to have removed more than 3,000 accounts, pages and groups and 20 million pieces of content for breaking rules about COVID-19 and vaccine-related misinformation.

Removal tends to be reserved for content that violates certain platform rules, such as showing prisoners of war or sharing fake and dangerous content. Labeling is for drawing attention to potentially unreliable content. Rules followed by platforms for each case are not set in stone and not very transparent.

Twitter has published policies to highlight its approach to reduce misinformation, for example with regards to COVID or manipulated media. However, when such policies are enforced, and how strongly, is difficult to determine and seem to vary significantly from one context to another.

Why moderation is so hard

But clearly, if the goal of moderating misinformation was to reduce the spread of false claims, social media companies' efforts were not entirely effective in reducing the amount of misinformation about COVID-19.

At the knowledge media institute at the Open University, we have been studying how both misinformation and corresponding fact checks spread on Twitter since 2016. Our research on COVID found that fact checks during the pandemic appeared relatively quickly after the appearance of misinformation. But the relationship between appearances of fact checks and the spread of misinformation in the study was less clear.

The study indicated that misinformation was twice as prevalent as the corresponding fact checks. In addition, misinformation about conspiracy theories was persistent, which meshes with previous research arguing that truthfulness is only one reason why people share information online and that fact checks are not always convincing.

So how can we improve moderation? Social media sites face numerous challenges. Users banned from one platform can still come back with a new account, or resurrect their profile on another platform. Spreaders of misinformation use tactics to avoid detection, for example by using euphemisms or visuals to avoid detection.

Automated approaches using machine learning and artificial intelligence are not sophisticated enough to detect misinformation very accurately. They often suffer from biases, lack of appropriate training, over-reliance on the English language, and difficulty handling misinformation in images, video or audio.

Different approaches

But we also know some techniques can be effective. For example, research has shown using simple prompts to encourage users to think about accuracy before sharing can reduce people's intention to share misinformation online (in laboratory settings, at least). Twitter has previously said it has found that labeling content as misleading or fabricated can slow the spread of some misinformation.

More recently, Twitter announced a new approach, introducing measures to address misinformation related to the Russian invasion of Ukraine. These including adding labels to tweets sharing links to Russian state-affiliated media websites. It also reduced the circulation of this content as well as improving its vigilance of hacked accounts.

> Today, we're adding labels to Tweets that share links to Russian state-affiliated media websites and are taking steps to significantly reduce the circulation of this content on Twitter.
>
> We'll roll out these labels to other state-affiliated media outlets in the coming weeks. pic.twitter.com/57Dycmn8lx
>
> — Yoel Roth (@yoyoel) February 28, 2022

Twitter is employing people as curators to write notes giving context or notes on Twitter trends, relating to the war to explain why things are trending. Twitter claims to have removed 100,000 accounts since the Ukraine war started that were in "violation of its platform manipulation strategy." It also says it has also labeled or removed 50,000 pieces of Ukraine war-related content.

In some as-yet unpublished research, we performed the same analysis we did for COVID-19, this time on over 3,400 claims about the Russian invasion of Ukraine, then monitoring tweets related to that misinformation about the Ukraine invasion, and tweets with factchecks attached. We started to observe different patterns.

We did notice a change in the spread of misinformation, in that the false claims appear not to be spreading as widely, and being removed more quickly, compared to previous scenarios. It's early days but one possible explanation is that the latest measures have had some effect.

If Twitter has found a useful set of interventions, becoming bolder and more effective in curating and labeling content, this could serve as a model for other social media platforms. It could at least offer a glimpse into the type of actions needed to boost fact-checking and curb misinformation. But it also makes Musk's purchase of the site and the implication that he will reduce moderation even more worrying.

Provided by The Conversation

Citation: Elon Musk could roll back social media moderation, just as we're learning how it can stop misinformation (2022, May 12) retrieved 25 April 2024 from https://techxplore.com/news/2022-05-elon-musk-social-media-moderation.html