How Facebook is using AI to combat COVID-19 misinformation and detect ‘hateful memes’

Facebook on Monday released a new report detailing how it uses a combination of artificial intelligence and human fact-checkers and moderators to enforce its community standards. The report — called the Community Standards Enforcement Report, which usually encompasses data and findings from the prior three to six months — has a large focus on AI this time around.

That’s because Facebook is relying more on the technology to help moderate its platform during the COVID-19 pandemic, which is preventing the company from using its usual third-party moderator firms because those firms’ employees are not allowed to access sensitive Facebook data from home computers.

That said, Facebook says the data it’s compiled so far doesn’t contain any larger trends in its enforcement or in offending behavior on its platform because the pandemic hit so late in its reporting period. “This report includes data only through March 2020 so it does not reflect the full impact of the changes we made during the pandemic,” writes Guy Rosen, the company’s vice president of integrity, in a blog post. “We anticipate we’ll see the impact of those changes in our next report, and possibly beyond, and we will be transparent about them.’

Given the state of the world, Facebook’s report does contain new information about how the company is specifically combating coronavirus-related misinformation and other forms of platform abuse, like price gouging on Facebook Marketplace,

using its AI tools.

Facebook put warning labels on 50 million coronavirus-related posts last month
“During the month of April, we put warning labels on about 50 million posts related to COVID-19 on Facebook, based on around 7,500 articles by our independent fact-checking partners,” the company said in a separate blog post, penned by a group of its research scientists and software engineers, about its ongoing COVID-19 misinformation efforts published today. “Since March 1st, we’ve removed more than 2.5 million pieces of content for the sale of masks, hand sanitizers, surface disinfecting wipes and COVID-19 test kits. But these are difficult challenges, and our tools are far from perfect. Furthermore, the adversarial nature of these challenges means the work will never be done.”

Facebook says its labels are working: 95 percent of the time, someone who is warned that a piece of content contains misinformation will decide not to view it anyway. But producing those labels across its enormous platform is proving to be a challenge. For one, Facebook is discovering that a fair amount of misinformation as well as hate speech is now showing up in images and videos, not just text or article links.

“We have found that a substantial percentage of hate speech on Facebook globally occurs in photos or videos,” the company says in a separate hate speech-specific blog post about its recent moderation findings and research. “As with other content, hate speech also can be multimodal: A meme might use text and image together to attack a particular group of people, for example.”

This is a tougher challenge for AI to tackle, the company admits. Not only do AI-trained models have a harder time parsing a meme image or a video due to complexities like wordplay and language differences, but that software must also

then be trained to find duplicates or only marginally modified versions of that content as it spreads across Facebook. But this is precisely what Facebook says it's achieved with what it calls SimSearchNet, a multiyear effort across many divisions within the company to train an AI model how to recognize both copies of the original image and those that are near-duplicates and have perhaps one word in the line of text changed.

"We have found that a substantial percentage of hate speech on Facebook globally occurs in photos or videos."
"Once independent fact-checkers have determined that an image contains misleading or false claims about coronavirus, SimSearchNet, as part of our end-to-end image indexing and matching system, is able to recognize near-duplicate matches so we can apply warning labels," the company says. "This system runs on every image uploaded to Instagram and Facebook and checks against task-specific human-curated databases. This accounts for billions of images being checked per day, including against databases set up to detect COVID-19 misinformation."

Facebook uses the example of a misleading image modeled after a broadcast news graphic with a line of overlaid text reading, "COVID-19 is found in toilet paper." The image is from a known peddler of fake news called Now8News, and the graphic has since been debunked by Snopes and other fact-checking organizations. But Facebook says it had to train its AI to differentiate between the original image and a modified one that says, "COVID-19 isn't found in toilet paper."

The goal is to help reduce the spread of duplicate images while also not inadvertently labeling genuine posts or those that don't meet the bar for misinformation. This is a big problem on Facebook where many politically motivated pages and organizations or those that simply feed off partisan outrage will take photographs, screenshots, and other images and alter them to change their meaning. An AI model that knows the

difference and can label one as misinformation and the other as genuine is a meaningful step forward, especially when it can then do the same to any duplicate or near-duplicate content in the future without roping in non-offending images in the process.

Image: Facebook

“It’s extremely important that these similarity systems be as accurate as possible, because a mistake can mean taking action on content that doesn’t actually violate our policies,” the company says. “This is particularly important because for each piece of misinformation fact-checker identifies, there may be thousands or millions of copies. Using AI to detect these matches also enables our fact-checking partners to focus on catching new instances of misinformation rather than near-identical variations of content they’ve already seen.”

Facebook has also improved its hate speech moderation using many of the same techniques it’s employing toward coronavirus-related content. “AI now proactively detects 88.8 percent of the hate speech content we remove, up from 80.2 percent the previous quarter,” the company says. “In the first quarter of 2020, we took action on 9.6 million pieces of content for violating our hate speech policies – an increase of 3.9 million.”

Facebook is able to rely more on AI, thanks to some advancements in how its models understand and parse text, both as it appears in posts and accompanying links and as overlaid in images or video.

“AI now proactively detects 88.8 percent of the hate speech content we remove.”

“People sharing hate speech often try to elude detection by modifying their content. This sort of adversarial behavior ranges from intentionally misspelling words or avoiding

certain phrases to modifying images and videos," the company says. "As we improve our systems to address these challenges, it's crucial to get it right. Mistakenly classifying content as hate speech can mean preventing people from expressing themselves and engaging with others." Facebook says so-called counterspeech, or a response to hate speech that argues against it but nonetheless usually contains snippets of the offensive content, is "particularly challenging to classify correctly because it can look so similar to the hate speech itself."

Facebook's latest report includes more data from Instagram, including how much bullying content that platform removes and how much of the content is appealed and reinstated. It applied its image-matching efforts toward finding suicide and self-injury posts, raising the percentage of Instagram content that was removed before users reported it.

Suicide and self-injury enforcement on Facebook also expanded in the last quarter of 2019, when the company removed 5 million pieces of content — double the amount it had removed in the months before. A spokesperson says this spike stemmed from a change that let Facebook detect and remove lots of very old content in October and November, and the numbers dropped dramatically in 2020 as it shifted its focus back to newer material.

Facebook says its new advances — in particular, a neural network it calls XLM-R announced last November — are helping its automated moderation systems better understand text across multiple languages. Facebook says XLM-R allows it "to train efficiently on orders of magnitude more data and for a longer amount of time," and to transfer that learning across multiple languages.

But Facebook says memes are proving to be a resilient and hard-to-detect delivery mechanism for hate speech, even with its improved tools. So it built a dedicated "hateful meme"

data set containing 10,000 examples, where the meaning of the image can only be fully understood by processing both the image and the text and understanding the relationship between the two.

An example is an image of a barren desert with the text, “Look how many people love you,” overlaid on top. Facebook calls the process of detecting this with automated systems multimodal understanding, and training its AI models with this level of sophistication is part of its more cutting-edge moderation research.

Image: Facebook

“To provide researchers with a data set with clear licensing terms, we licensed assets from Getty Images. We worked with trained third-party annotators to create new memes similar to existing ones that had been shared on social media sites,” the company says. “The annotators used Getty Images’ collection of stock images to replace the original visuals while still preserving the semantic content.”

Facebook says it’s providing the data set to researchers to improve techniques for detecting this type of hate speech online. It’s also launching a challenge with a $100,000 prize for researchers to create models trained on the data set that can successfully parse these more subtle forms of speech that Facebook is seeing more often now that its systems are more proactively taking down more blatant hateful content.

https://www.theverge.com/2020/5/12/21254960/facebook-ai-moderation-covid-19-coronavirus-hateful-memes-hate-speech