Facebook: We're better at policing nudity than hate speech

May 15 2018, by Michael Liedtke

In this May 16, 2012, file photo, the Facebook logo is displayed on an iPad in Philadelphia. Facebook believes its policing system is better at scrubbing graphic violence, gratuitous nudity and terrorist propaganda from its social network than it is at removing racist, sexist and other hateful remarks. The self-assessment on Tuesday, May 15, 2018, came three weeks after Facebook tried to give a clearer explanation of the kinds of posts that it won't tolerate. (AP Photo/Matt Rourke, File)

Getting rid of racist, sexist and other hateful remarks on Facebook is

more challenging than weeding out other types of unacceptable posts because computer programs still stumble over the nuances of human language, the company revealed Tuesday.

Facebook also released statistics that quantified how pervasive fake accounts have become on its influential service, despite a long-standing policy requiring people to set up accounts under their real-life identities.

From October to December alone, Facebook disabled nearly 1.3 billion accounts—and that doesn't even count all the times the company blocked bogus profiles before they could be set up.

Had the company not shut down all those fake accounts, its audience of monthly users would have swelled beyond its current 2.2 billion and probably created more potentially offensive material for Facebook to weed out.

Facebook's self-assessment showed its screening system is far better at scrubbing graphic violence, gratuitous nudity and terrorist propaganda. Automated tools detected 86 percent to 99.5 percent of the violations Facebook identified in those categories.

For hate speech, Facebook's human reviewers and computer algorithms identified just 38 percent of the violations. The rest came after Facebook users flagged the offending content for review.

All told, Facebook took action on nearly 1.6 billion pieces of content during the six months ending in March, a tiny fraction of all the activity on its social network, according to the company.

The report marked Facebook's first breakdown on how much material it removes for violating its policies. It didn't disclose how long it takes Facebook to remove material violating its standards. The report also

doesn't cover how much inappropriate content Facebook missed.

"Even if they remove 100 million posts that are offensive, there will be one or two that have some really bad stuff and those will be the ones everyone winds up talking about on the cable-TV news," said Timothy Carone, who teaches about technology at the University of Notre Dame.

Instead of trying to determine how much offending material it didn't catch, Facebook provided an estimate on how frequently it believes users saw posts that violated its standards, including content that its screening system didn't detect. For instance, the company estimated that for every 10,000 times that people looked at content on its social network, 22 to 27 of the views may have included posts that included impermissible graphic violence.

The report also doesn't address how Facebook is tacking another vexing issue—the proliferation of fake news stories planted by Russian agents and other fabricators trying to sway elections and public opinion.

Fake accounts on Facebook have been drawing more attention because Russian agents used them to buy ads to try to influence the 2016 election in the U.S.

Even though it has been focusing on shutting down bogus accounts, Facebook has said that 3 to 4 percent of its active monthly users are fake. That means as many as 88 million fake Facebook accounts were still slipping through the cracks in the company's policing system through March.

It's not surprising that Facebook's automated programs have the greatest difficulty trying to figure out differences between permissible opinions and despicable language that crosses the line, Carone said.

"It's like trying to figure out the equivalent between screaming 'Fire!' in a crowded theater when there is none and the equivalent of saying something that is uncomfortable but qualifies as free speech," he said.

Facebook said it removed 2.5 million pieces of content deemed unacceptable hate speech during the first three months of this year, up from 1.6 million during the previous quarter. The company credited better detection, even as it said computer programs have trouble understanding context and tone of language.

Facebook took down 3.4 million pieces of graphic violence during the first three months of this year, nearly triple the 1.2 million during the previous three months. In this case, better detection was only part of the reason. Facebook said users were more aggressively posting images of violence in places like war-torn Syria.

The increased transparency comes as the Menlo Park, California, company tries to make amends for a privacy scandal triggered by loose policies that allowed a data-mining company with ties to President Donald Trump's 2016 campaign to harvest personal information on as many as 87 million users. The content screening has nothing to do with privacy protection, though, and is aimed at maintaining a family-friendly atmosphere for users and advertisers.