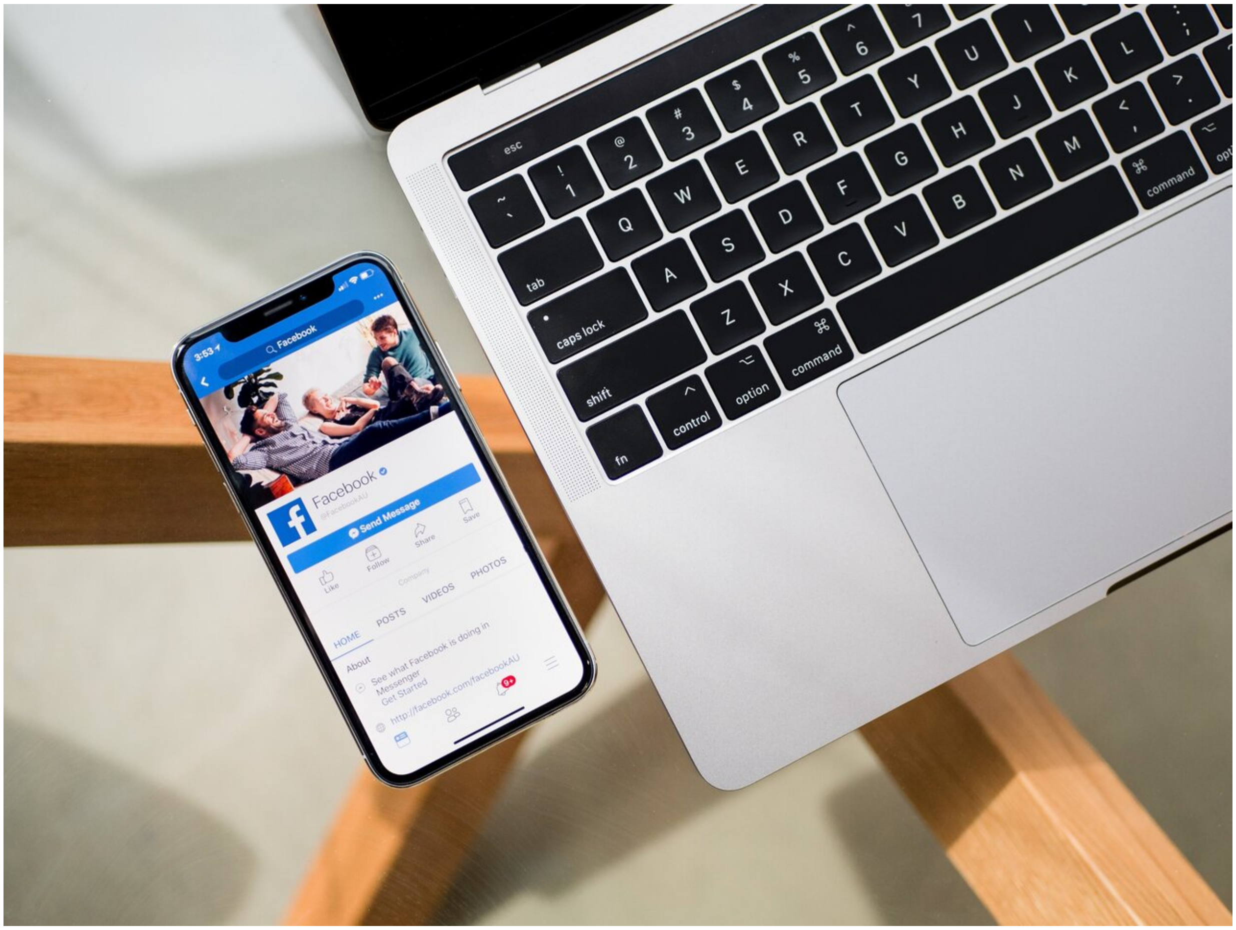
Facebook is using AI to stem dangerous and false posts

November 14 2020, by Peter Grad



Credit: Unsplash/CC0 Public Domain

Facebook has come under withering criticism this past year from folks who say the company is not doing enough to stem hate speech, online harassment and the spread of false news stories.

To be fair, the task of policing the activities of 1.62 billion daily users generating 4 petabytes of data, including 350 million photos, per day is no small task. It's not easy being the world's largest social platform.

Still, the company has been criticized for allowing scores of hate-fueled groups to spread offensive and threatening posts and for allowing ultra-rightwing conspiracy-theory groups such as QAnon to freely spread false political allegations. Academic and governmental analyses of the 2016 presidential election uncovered evidence of massive interference by domestic and foreign actors, and it appears similar efforts were undertaken in the 2020 election as well.

Facebook employs 15,000 content moderators to review reports of misbehavior ranging from political subterfuge to harassment to terroristic threats to child exploitation. They have generally tackled reports chronologically, frequently allowing more serious allegations to go unaddressed for days while lesser issues were reviewed.

On Friday, Facebook announced that it will bring machine learning into he moderating process. It will utilize algorithms to detect the most severe issues and assign them to human moderators. Software moderators will continue to handle lower-level abuse such as copyright infringement and spam.

Facebook says it will evaluate problematic posts according to three criteria: virality, severity and likelihood they are violating rules. An obscenity-laced post threatening violence at the site of racial unrest, for example, would be given top priority, either removed automatically by machine or assigned to a moderator for immediate evaluation and action.

"All content violations ... still receive some substantial human review," said Ryan Barnes, a product manager on Facebook's Community Integrity team. "We'll be using this system to better prioritize content. We expect to use more automation when violating content is less severe, especially if the content isn't viral, or being … quickly shared by a large number of people."

Facebook has been accused of mishandling accounts during recent high-profile disturbances. In one instance, the company was sued after deadly shootings by vigilantes who descended on Kenosha, Wisconsin, following protests against police officers who gravely wounded a black man after firing four shots into his back during an arrest. The suit alleges Facebook failed to remove the pages of hate groups involved in the vigilante shootings.

During the pandemic of the past year, a study by a non-profit organization found 3.8 billion views on Facebook of misleading content related to COVID-19.

Sometimes, criticism is prompted by overly cautious Facebook moderators. Last June, The Guardian newspaper complained that readers attempting to circulate a historic photo it published were blocked and issued warnings by Facebook. The image of nearly naked Aboriginal men in chains in Western Australia, taken in the 1890s, was published in response to a denial by Australian Prime Minister Scott Morrison that his country never engaged in slavery. Morrison retracted his comments following publication of the article and photo. Facebook subsequently apologized for incorrectly categorizing the photo as inappropriate nudity.

Facebook officials say applying machine learning is part of a continuing effort to halt the spread of dangerous, offensive and misleading information while ensuring legitimate posts are not censored.

An example of the challenges Facebook confronts was the virtual overnight creation of a massive protest group contesting the 2020 election count. A Facebook group demanding a recount garnered 400,000 members within just a few days. Facebook has not blocked the page.

While there is nothing illegal about requesting a recount, a tidal wave of misinformation regarding alleged voting abuses—charges that have been categorically dismissed by officials in all 50 states and by Republicans as well as Democrats this past week —is a troubling reminder of the potential of false information to shape political views.

"The system is about marrying AI and human reviewers to make less total mistakes," said Chris Palow, a member of Facebook's Integrity team. "The AI is never going to be perfect."

More information: www.theverge.com/2020/11/13/21 … cebook-ai-moderation

Citation: Facebook is using AI to stem dangerous and false posts (2020, November 14) retrieved 20 April 2024 from https://techxplore.com/news/2020-11-facebook-ai-stem-dangerous-false.html