Instagram is using AI to stop people from posting abusive comments

Antonio Regalado

The social-media platform will flag possibly offensive comments before they're posted and ask the poster to reconsider.

The context: Online abuse has always been a complicated issue because of its scale and nuance. What counts as worthy of censorship is a perpetual debate: filter too much and it infringes on self-expression; filter too little and it creates a hostile environment. Add to that the complexity of different languages, cultures, and norms, and the challenge gets really unwieldy.

Artificial unintelligence: That's why social-media platforms like Facebook have turned to artificial intelligence to help them sort through the sheer volume of posts and comments. But language has always been particularly hard for AI to parse—simple things like double entendres, sarcasm, or even misspellings can trip up a system into mistaking their meaning. To get around that, Facebook employs thousands of content moderators to step up when its algorithms fail to make a final judgment call. Investigations have found that those jobs are often brutal and grueling, however.

Instagram's solution: Instagram, which is owned by Facebook, is now trying a new approach. Rather than rely solely on its algorithms to censor offensive material, it will draw on users' self-censorship as well. As a comment is posting, if the platform's AI model flags it as harmful, the poster will see a pop-up asking "Are you sure you want to post this?"

In early tests, Instagram found the feature encouraged many people to rescind their comments, according to yesterday's blog post announcement. It's a clever tactic to try to alleviate some of the burden on human content moderation without being too restrictive.

Training data: This isn't the first time Instagram has used AI to clean up language. In June of 2017, it also launched an offensive-comment filter using machine learning to hide the most obvious abuse. The platform has since continued to improve its machine-learning model, likely making use of the millions of data points generated by users when they reported comments in the past. The latest feature also asks users to notify Instagram if it has flagged their comment as offensive by mistake, another feedback loop that could generate more useful training data.

To have more stories like this delivered directly to your inbox, sign up for our Webby-nominated AI newsletter The Algorithm. It's free.