Facebook Proactively Removed 80% of Posts Pulled for Hate Speech in Q2, Q3

|
7 hours ago
Facebook released the fourth edition of its Community Standards Enforcement Report Wednesday, covering the second and third quarters of 2019.

CEO Mark Zuckerberg said during a press call discussing the report, “While we err on the side of free expression, we do have community standards defining what is acceptable on our platform and what isn’t. This is a tiny fraction of content on our platform, but this is some of the worst content out there.”

Vice president of integrity Guy Rosen said in a Newsroom post that the report now includes metrics across 10 policies on Facebook and four on Instagram.

The Facebook policies covered are: adult nudity and sexual activity; bullying and harassment; child nudity and sexual exploitation of children; fake accounts; hate speech; regulated goods (drugs and firearms); spam; suicide and self-injury; terrorist propaganda; and violent and graphic content.

And for Instagram, policies covered are: child nudity and sexual exploitation of children; regulated goods (drugs and firearms); suicide and self-injury; and terrorist propaganda.

Instagram head of product Vishal Shah said during the press call that future reports will include data on additional policy areas.

Metrics detailed by Facebook include how often content that violates its policies was viewed, how much content action was taken on, how much of that content was detected before someone reported it, how many appeals there were on content after taking action and how much of that content was restored after the company initially took action.

Rosen said of the first metric, which Facebook refers to as prevalence, “Think of this as an air quality test to determine the amount of pollutants in the air. We focus on how much content is seen.”

Rosen said data on appeals and restores was not available for Instagram, as appeals were not added to that platform until the second quarter, but such data will be included in future reports.

He also pointed out that metrics may vary between Facebook and Instagram, as the latter does not offer links or reshares in feed, pages or groups.

Rosen was also asked about how the figures being reported by Facebook Wednesday accounted for Facebook Stories and Instagram Stories, and whether the fact that content in that form disappears within 24 hours affected enforcement, and he replied that the same user reporting capabilities and proactive detection systems apply to Stories, and the approach is by and large the same.

Rosen said Facebook’s rate of removing hate speech proactively is up to 80%, compared with 68% in its last Community Standards Enforcement Report, attributing the increase to investments and advances in detection techniques such as text and image matching, as well as machine learning classifiers that examine language and reactions and comments to posts.

On automatic removals, he wrote, “We only do this in select instances, and it has only been possible because our automated systems have been trained on hundreds of thousands, if not

millions, of different examples of violating content and common attacks. In all other cases when our systems proactively detect potential hate speech, the content is still sent to our review teams to make a final determination."

"Defining hate speech involves a lot of linguistic nuance," Zuckerberg said during the press call.

Facebook said its rate of detecting and removing content associated with al-Qaida, Isis (Islamic State) and their affiliates remains above 99%, and for all terrorist organizations, that figure is 98.5% on Facebook and 92.2% on Instagram.

Facebook vp of global policy management Monika Bickert said during the press call that the company expanded its report to include enforcement against all terrorist organizations, adding, "We are always evolving our tactics because we know bad actors will continue to change theirs."

Continue Reading

Pages: 1 2

https://www.adweek.com/digital/facebook-proactively-removed-80-of-posts-pulled-for-hate-speech-in-q2-q3/