PHYS.ORG

Twitter tweak steps up fight against trolls

May 16 2018

Twitter said Tuesday it was stepping up its long-running battle against online trolls, trying to find offenders by looking at "behavioral signals."

The new approach looks at behavioral patterns of users in addition to the content of the tweets, allowing Twitter to find and mute online bullies and trolls.

Even if the offending tweets are not a violation of Twitter policy, they may be hidden from users if they are deemed to "distort" the conversation, Twitter said.

The announcement is the latest "safety" initiative by Twitter, which is seeking to filter out offensive speech while remaining an open platform.

Twitter already uses artificial intelligence and machine learning in this effort but the latest initiative aims to do more by focusing on the actions of certain users in addition to the content.

"Our ultimate goal is to encourage more free and open conversation," chief executive Jack Dorsey said.

"To do that we need to significantly reduce the ability to game and skew our systems. Looking at behavior, not content, is the best way to do that."

A Twitter blog post said the move aims at "troll-like behavior" which targets certain users and tweets with derisive responses.

"Some troll-like behavior is fun, good and humorous. What we're talking about today are troll-like behaviors that distort and detract from the public conversation on Twitter," said the blog from Twitter executives Del Harvey and David Gasca.

"Some of these accounts and tweets violate our policies, and, in those cases, we take action on them. Others don't but are behaving in ways that distort the conversation."

Harvey and Gasca said the challenge has been to address "disruptive behaviors that do not violate our policies but negatively impact the health of the conversation."

The new approach does not wait for people who use Twitter to report potential issues.

"There are many new signals we're taking in, most of which are not visible externally," the blog post said.

"Just a few examples include if an account has not confirmed their email address, if the same person signs up for multiple accounts simultaneously, accounts that repeatedly tweet and mention accounts that don't follow them, or behavior that might indicate a coordinated attack."

In some cases, if the content is not a violation of Twitter policies, it will not be deleted but only shown when a user clicks on "show more replies."

"The result is that people contributing to the healthy conversation will be more visible in conversations and search," Harvey and Gasca wrote.

Twitter said its tests of this approach shows a four percent drop in abuse reports from search and eight percent fewer abuse reports from conversations.

Citation: Twitter tweak steps up fight against trolls (2018, May 16) retrieved 18 April 2024 from https://phys.org/news/2018-05-twitter-tweak-trolls.html