Warnings may reduce hate speech on Twitter, new study finds

November 22 2021

Credit: CC0 Public Domain

Warning Twitter users about potential adverse consequences of their use of hate speech can decrease their subsequent posting of hateful language for a week, finds a new study by New York University's Center for Social Media and Politics.

"Debates over the effectiveness of social media account suspensions and bans on abusive users abound, but we know little about the impact of either warning a user of suspending an account or of outright suspensions in order to reduce hate speech," explains Mustafa Mikdat Yildirim, an NYU doctoral candidate and the lead author of the paper, which appears in the journal *Perspectives on Politics*. "Even though the impact of warnings is temporary, the research nonetheless provides a potential path forward for platforms seeking to reduce the use of hateful language by users."

In the aftermath of decisions by Twitter and other social media platforms to suspend large numbers of accounts, in particular those of former President Donald Trump following the Jan. 6, 2021 attack on the U.S. Capitol, many have asked about the effectiveness of measures aimed at curbing hate speech and other messages that may incite violence.

In the *Perspectives on Politics* paper, the researchers examined one approach—issuing *warnings* of possible suspensions resulting from the use of hate speech—to determine its efficacy in diminishing future use of this type of language.

To do so, the paper's authors designed a series of experiments aimed at instilling the possible consequences of the use of hate and related speech.

"To effectively convey a warning message to its target, the message needs to make the target aware of the consequences of their behavior and also make them believe that these consequences will be administered," they write.

In constructing their experiments, the authors focused on the *followers* of users whose accounts had been suspended for posting tweets that used hateful language in order to find a group of users for whom they could create credible warning messages. The researchers reasoned that the followers of those who had been suspended and who also used hateful language might consider themselves potential "suspension candidates" once they learned someone they followed had been suspended—and therefore be potentially willing to moderate their behavior following a warning.

To identify such candidates, the team downloaded more than 600,000 tweets on July 21, 2020 that were posted in the week prior and that contained at least one word from hateful language dictionaries used in previous research. During the period, Twitter was flooded by hateful tweets against both the Asian and Black communities due to the coronavirus pandemic and Black Lives Matter protests.

From this group of users of hateful language, the researchers obtained a sample of approximately 4,300 followers of users who had been suspended by Twitter during this period (i.e., "suspension candidates").

These followers were divided into six treatment groups and one control group. The researchers tweeted one of six possible warning messages to these users, all prefaced with this sentence: "The user [@account] you follow was suspended, and I suspect that this was because of hateful language." It was followed by different types of warnings, ranging from "If you continue to use hate speech, you might get suspended temporarily" to "If you continue to use hate speech, you might lose your

posts, friends and followers, and not get your account back." The control group did not receive any messages.

Overall, the users who received these warning messages reduced the ratio of tweets containing hateful language by up to 10 percent a week later (there was no significant reduction among those in the control group). And, in cases in which the messaging to users was more politely phrased ("I understand that you have every right to express yourself but please keep in mind that using hate speech can get you suspended."), the decline reached 15 to 20 percent. (Based on previous scholarship, the authors concluded that respectful and polite language would be more likely to be seen as legitimate.) However, the impact of the warnings dissipated a month later.

The paper's other authors were Joshua A. Tucker and Jonathan Nagler, professors in NYU's Department of Politics, and Richard Bonneau, a professor in NYU's Department of Biology and Courant Institute of Mathematical Sciences. Tucker, Nagler, and Bonneau are co-directors of the NYU Center for Social Media and Politics, where Yildirim conducts research as a Ph.D. candidate.

Provided by New York University

Citation: Warnings may reduce hate speech on Twitter, new study finds (2021, November 22) retrieved 3 May 2024 from https://phys.org/news/2021-11-speech-twitter.html