The double-down is real: Correcting online falsehoods might make matters worse

May 20 2021, by Peter Dizikes

Not only is misinformation increasing online, but attempting to correct it politely on Twitter can have negative consequences, leading to even less-accurate tweets and more toxicity from the people being corrected, according to a new study co-authored by a group of MIT scholars. Credit: Christine Daniloff, MIT

So you thought the problem of false information on social media could

not be any worse? Allow us to respectfully offer evidence to the contrary.

Not only is misinformation increasing online, but attempting to correct it politely on Twitter can have negative consequences, leading to even less-accurate tweets and more toxicity from the people being corrected, according to a new study co-authored by a group of MIT scholars.

The study was centered around a Twitter field experiment in which a research team offered polite corrections, complete with links to solid evidence, in replies to flagrantly false tweets about politics.

"What we found was not encouraging," says Mohsen Mosleh, a research affiliate at the MIT Sloan School of Management, lecturer at University of Exeter Business School, and a co-author of a new paper detailing the study's results. "After a user was corrected … they retweeted news that was significantly lower in quality and higher in partisan slant, and their retweets contained more toxic language."

The paper, "Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment," has been published online in CHI '21: Proceedings of the 2021 Conference on Human Factors in Computing Systems.

The paper's authors are Mosleh; Cameron Martel, a Ph.D. candidate at MIT Sloan; Dean Eckles, the Mitsubishi Career Development Associate Professor at MIT Sloan; and David G. Rand, the Erwin H. Schell Professor at MIT Sloan.

From attention to embarrassment?

To conduct the experiment, the researchers first identified 2,000 Twitter users, with a mix of political persuasions, who had tweeted out any one of 11 frequently repeated false news articles. All of those articles had been debunked by the website Snopes.com. Examples of these pieces of misinformation include the incorrect assertion that Ukraine donated more money than any other nation to the Clinton Foundation, and the false claim that Donald Trump, as a landlord, once evicted a disabled combat veteran for owning a therapy dog.

The research team then created a series of Twitter bot accounts, all of which existed for at least three months and gained at least 1,000 followers, and appeared to be genuine human accounts. Upon finding any of the 11 false claims being tweeted out, the bots would then send a reply message along the lines of, "I'm uncertain about this article—it might not be true. I found a link on Snopes that says this headline is false." That reply would also link to the correct information.

Among other findings, the researchers observed that the accuracy of news sources the Twitter users retweeted promptly declined by roughly 1 percent in the next 24 hours after being corrected. Similarly, evaluating over 7,000 retweets with links to political content made by the Twitter accounts in the same 24 hours, the scholars found an upturn by over 1 percent in the partisan lean of content, and an increase of about 3 percent in the "toxicity" of the retweets, based on an analysis of the language being used.

In all these areas—accuracy, partisan lean, and the language being used—there was a distinction between retweets and the primary tweets written by the Twitter users. Retweets, specifically, degraded in quality, while tweets original to the accounts being studied did not.

"Our observation that the effect only happens to retweets suggests that the effect is operating through the channel of attention," says Rand,

noting that on Twitter, people seem to spend a relatively long time crafting primary tweets, and little time making decisions about retweets.

He adds: "We might have expected that being corrected would shift one's attention to accuracy. But instead, it seems that getting publicly corrected by another user shifted people's attention away from accuracy—perhaps to other social factors such as embarrassment." The effects were slightly larger when people were being corrected by an account identified with the same political party as them, suggesting that the negative response was not driven by partisan animosity.

Ready for prime time

As Rand observes, the current result seemingly does not follow some of the previous findings that he and other colleagues have made, such as a study published in *Nature* in March showing that neutral, nonconfrontational reminders about the concept of accuracy can increase the quality of the news people share on social media.

"The difference between these results and our prior work on subtle accuracy nudges highlights how complicated the relevant psychology is," Rand says.

As the current paper notes, there is a big difference between privately reading online reminders and having the accuracy of one's own tweet publicly questioned. And as Rand notes, when it comes to issuing corrections, "it is possible for users to post about the importance of accuracy in general without debunking or attacking specific posts, and this should help to prime accuracy and increase the quality of news shared by others."

At least, it is possible that highly argumentative corrections could produce even worse results. Rand suggests the style of corrections and

the nature of the source material used in corrections could both be the subject of additional research.

"Future work should explore how to word corrections in order to maximize their impact, and how the source of the correction affects its impact," he says.

More information: Mohsen Mosleh et al, Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment, *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021). DOI: 10.1145/3411764.3445642

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology