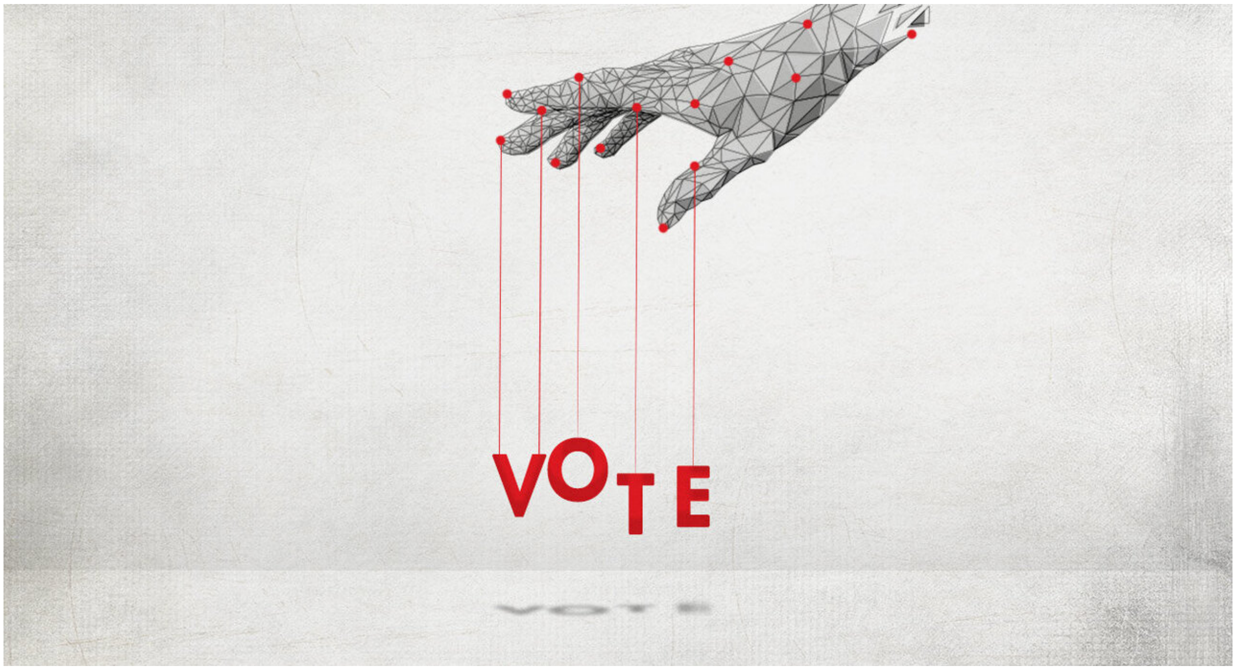
Tracking misinformation campaigns in real-time is possible, study shows

July 22 2020



A research team led by Princeton University has developed a technique for tracking online foreign misinformation campaigns in real time, which could help mitigate outside interference in the 2020 American election. Credit: Egan Jimenez, Princeton University

A research team led by Princeton University has developed a technique for tracking online foreign misinformation campaigns in real time, which could help mitigate outside interference in the 2020 American election.

The researchers developed a method for using machine learning to identify malicious internet accounts, or trolls, based on their past behavior. Featured in *Science Advances*, the model investigated past misinformation campaigns from China, Russia, and Venezuela that were waged against the United States before and after the 2016 election.

The team identified the patterns these campaigns followed by analyzing posts to Twitter and Reddit and the hyperlinks or URLs they included. After running a series of tests, they found their model was effective in identifying posts and accounts that were part of a foreign influence campaign, including those by accounts that had never been used before.

They hope that software engineers will be able to build on their work to create a real-time monitoring system for exposing foreign influence in American politics.

"What our research means is that you could estimate in real time how much of it is out there, and what they're talking about," said Jacob N. Shapiro, professor of politics and international affairs at the Princeton School of Public and International Affairs. "It's not perfect, but it would force these actors to get more creative and possibly stop their efforts. You can only imagine how much better this could be if someone puts in the engineering efforts to optimize it."

Shapiro and associate research scholar Meysam Alizadeh conducted the study with Joshua Tucker, professor of politics at New York University, and Cody Buntain, assistant professor in informatics at New Jersey Institute of Technology.

The team began with a simple question: Using only content-based features and examples of known influence campaign activity, could you look at other content and tell whether a given post was part of an influence campaign?

They chose to investigate a unit known as a "postURL pair," which is simply a post with a hyperlink. To have real influence, coordinated operations require intense human and bot-driven information sharing. The team theorized that similar posts may appear frequently across platforms over time.

They combined data on troll campaigns from Twitter and Reddit with a rich dataset on posts by politically engaged users and average users collected over many years by NYU's Center for Social Media and Politics (CSMaP). The troll data included publicly available Twitter and Reddit data from Chinese, Russian, and Venezuelan trolls totaling 8,000 accounts and 7.2 million posts from late 2015 through 2019.

"We couldn't have conducted the analysis without that baseline comparison dataset of regular, ordinary tweets," said Tucker, co-director of CSMaP. "We used it to train the model to distinguish between tweets from coordinated influence campaigns and those from ordinary users."

The team considered the characteristics of the post itself, like the timing, word count, or if the mentioned URL domain is a news website. They also looked at what they called "metacontent," or how the messaging in a post related to other information shared at that time (for example, whether a URL was in the top 25 political domains shared by trolls.)

"Meysam's insight on metacontent was key," Shapiro said. "He saw that we could use the machine to replicate the human intuition that 'something about this post just looks out of place.' Both trolls and normal people often include local news URLs in their posts, but the trolls tended to mention different users in such posts, probably because they are trying to draw their audience's attention in a new direction. Metacontent lets the algorithm find such anomalies."

The team tested their method extensively, examining performance month to month on five different prediction tasks across four influence campaigns. Across almost all of the 463 different tests, it was clear which posts were and were not part of an influence operation, meaning that content-based features can indeed help find coordinated influence campaigns on social media.

In some countries, the patterns were easier to spot than others. Venezuelan trolls only retweeted certain people and topics, making them easy to detect. Russian and Chinese trolls were better at making their content look organic, but they, too, could be found. In early 2016, for example, Russian trolls quite often linked to far-right URLs, which was unusual given the other aspects of their posts, and, in early 2017, they linked to political websites in odd ways.

Overall, Russian troll activity became harder to find as time went on. It is possible that investigative groups or others caught on to the false information, flagging the posts and forcing trolls to change their tactics or approach, though Russians also appear to have produced less in 2018 than in previous years.

While the research shows there is no stable set of characteristics that will find influence efforts, it also shows that troll content will almost always be different in detectable ways. In one set of tests, the authors show the method can find never-before-used accounts that are part of an ongoing campaign. And while social media platforms regularly delete accounts associated with foreign disinformation campaigns, the team's findings could lead to a more effective solution.

"When the platforms ban these accounts, it not only makes it hard to collect data to find similar accounts in the future, but it signals to the disinformation actor that they should avoid the behavior that led to deletion," said Buntain. "This mechanism allows [the platform] to identify these accounts, silo them away from the rest of Twitter, and make it appear to these actors as though they are continuing to share their disinformation material."

The work highlights the importance of interdisciplinary research between social and computational science, as well as the criticality of funding research data archives.

"The American people deserve to understand how much is being done by foreign countries to influence our politics," said Shapiro. "These results suggest that providing that knowledge is technically feasible. What we currently lack is the political will and funding, and that is a travesty."

The method is no panacea, the researchers cautioned. It requires that someone has already identified recent influence campaign activity to learn from. And how the different features combine to indicate questionable content changes over time and between campaigns.

More information: M. Alizadeh el al., "Content-based features predict social media influence operations," *Science Advances* (2020). advances.sciencemag.org/lookup … .1126/sciadv.abb5824

Provided by Princeton University

Citation: Tracking misinformation campaigns in real-time is possible, study shows (2020, July 22) retrieved 18 April 2024 from https://techxplore.com/news/2020-07-tracking-misinformation-campaigns-real-time.html