Research shows that anyone can become an internet troll

February 7 2017, by Taylor Kubota

Credit: Wikipedia

Internet trolls, by definition, are disruptive, combative and often unpleasant with their offensive or provocative online posts designed to disturb and upset.

The common assumption is that people who troll are different from the rest of us, allowing us to dismiss them and their behavior. But research from Stanford University and Cornell University, published as part of the upcoming 2017 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017), suggests otherwise. The research offers evidence that, under the right circumstances, anyone can become a troll.

"We wanted to understand why trolling is so prevalent today," said Justin Cheng, a computer science researcher at Stanford and lead author of the paper. "While the common knowledge is that trolls are particularly sociopathic individuals that occasionally appear in conversations, is it really just these people who are trolling others?"

Taking inspiration from social psychology research methods, Cheng investigated whether trolling behavior is an innate characteristic or if situational factors can influence people to act like trolls. Through a combination of experimentation, data analysis and machine learning, the researchers honed in on simple factors that make the average person more likely to troll.

Becoming a troll

Following previous research on antisocial behavior, the researchers decided to focus on how mood and context affect what people write on a discussion forum. They set up a two-part experiment with 667 subjects recruited through a crowdsourcing platform.

In the first part of the experiment, the participants were given a test,

which was either very easy or very difficult. After taking the tests, all subjects filled out a questionnaire that evaluated various facets of their mood, including anger, fatigue, depression and tension. As expected, the people who completed the difficult test were in a worse mood than those who had the easy test.

All participants were then instructed to read an article and engage in its comment section. They had to leave at least one comment, but could leave multiple comments and up-votes and down-votes and could reply to other comments. All participants saw the same article on the same platform, created solely for the experiment, but some participants were given a forum with three troll posts at the top of the comment section. Others saw three neutral posts.

Two independent experts evaluated whether the posts left by subjects qualified as trolling, defined generally in this research by a combination of posting guidelines taken from several discussion forums. For example, personal attacks and cursing were indicative of troll posts.

About 35 percent of people who completed the easy test and saw neutral posts then posted troll comments of their own. That percentage jumped to 50 percent if the subject either took the hard test or saw trolling comments. People exposed to both the difficult test and the troll posts trolled approximately 68 percent of the time.

The spread of trolling

To relate these experimental insights to the real world, the researchers also analyzed anonymized data from CNN's comment section from throughout 2012. This data consisted of 1,158,947 users, 200,576 discussions and 26,552,104 posts. This included banned users and posts that were deleted by moderators. In this part of the research, the team defined troll posts as those that were flagged by members of the

community for abuse.

It wasn't possible to directly evaluate the mood of the commenters, but the team looked at the time stamp of posts because previous research has shown that time of day and day of week correspond with mood. Incidents of down-votes and flagged posts lined up closely with established patterns of negative mood. Such incidents tend to increase late at night and early in the week, which is also when people are most likely to be in a bad mood.

The researchers investigated the effects of mood further and found that people were more likely to produce a flagged post if they had recently been flagged or if they had taken part in a separate discussion that merely included flagged posts written by others. These findings held true no matter what article was associated with the discussion.

"It's a spiral of negativity," explained Jure Leskovec, associate professor of computer science at Stanford and senior author of the study. "Just one person waking up cranky can create a spark and, because of discussion context and voting, these sparks can spiral out into cascades of bad behavior. Bad conversations lead to bad conversations. People who get down-voted come back more, comment more and comment even worse."

Predicting bad behavior

As a final step in their research, the team created a machine-learning algorithm tasked with predicting whether the next post an author wrote would be flagged.

The information fed to the algorithm included the time stamp of the author's last post, whether the last post was flagged, whether the previous post in the discussion was flagged, the author's overall history of writing flagged posts and the anonymized user ID of the author.

The findings showed that the flag status of the previous post in the discussion was the strongest predictor of whether the next post would be flagged. Mood-related features, such as timing and previous flagging of the commenter, were far less predictive. The user's history and user ID, although somewhat predictive, were still significantly less informative than discussion context. This implies that, while some people may be consistently more prone to trolling, the context in which we post is more likely to lead to trolling.

Troll prevention

Between the real-life, large-scale data analysis, the experiment and the predictive task, the findings were strong and consistent. The researchers suggest that conversation context and mood can lead to trolling. They believe this could inform the creation of better online discussion spaces.

"Understanding what actually determines somebody to behave antisocially is essential if we want to improve the quality of online discussions," said Cristian Danescu-Niculescu-Mizil, assistant professor of information science at Cornell University and co-author of the paper. "Insight into the underlying causal mechanisms could inform the design of systems that encourage a more civil online discussion and could help moderators mitigate trolling more effectively."

Interventions to prevent trolling could include discussion forums that recommend a cooling-off period to commenters who have just had a post flagged, systems that automatically alert moderators to a post that's likely to be a troll post or "shadow banning," which is the practice of hiding troll posts from non-troll users without notifying the troll.

The researchers believe studies like this are only the beginning of work that's been needed for some time, since the Internet is far from being the worldwide village of cordial debate and discussion people once thought

it would become.

"At the end of the day, what this research is really suggesting is that it's us who are causing these breakdowns in discussion," said Michael Bernstein, assistant professor of computer science at Stanford and co-author of the paper. "A lot of news sites have removed their comments systems because they think it's counter to actual debate and discussion. Understanding our own best and worst selves here is key to bringing those back."

More information: Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. files.clr3.com/papers/2017_anyone.pdf

Provided by Stanford University