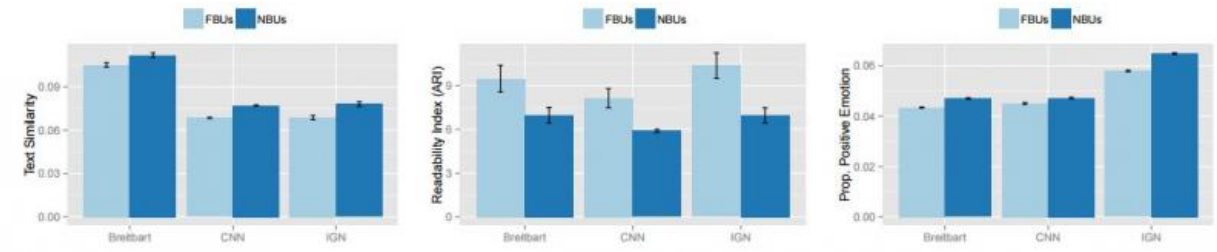
Algorithm able to identify online trolls

April 14 2015, by Bob Yirka



Users who get banned in the future (FBUs) (a) write less similarly to other users in the same thread, (b) write posts that are harder to read (i.e., have a higher readability index), and (c) express less positive emotion. Credit: arXiv:1504.00680 [cs.SI]

A trio of researchers, two from Cornell the other from Stanford has developed a computer algorithm that is capable of identifying antisocial behavior as demonstrated in website comment sections. In their paper uploaded to the preprint server *arXiv*, Justin Cheng, Cristian Danescu-Niculescu-Mizil and Jure Leskovec describe their algorithm, how they came up with it and how they plan to improve on its accuracy.

On the Internet, people who engage in antisocial ways in the comments section of web content, have come to be known as trolls, and they represent, like those who dish out spam, a constant source of annoyance—so much so that big name websites like CNN, are working with academics to find ways to identify trolls and ban them before they cause too much trouble. Visitors to web sites that are harassed or made

to feel bad often avoid websites where they feel they have been angered or annoyed. In this new effort, the researchers built their troll finding algorithm by engaging in an analysis of typical troll behavior with data provided by CNN.com, Breitbart.com and IGN.com.

In scouring the data (comparing the behavior of those that have been banned against others that have never been banned) over an 18 month period which included studying the comments of over 10,000 Future Banned Users (FBUs), the researchers discovered some patterns in troll behavior—first, they noticed that on average troll posts were less literate than non-trolls—and they tended to get less literate the more they posted to a site. They also found that fellow posters were initially patient with trolls, but reached a plateau, at which point, banning came quickly.

The researchers report that it was relatively easy to spot FBUs and to convert what they had found to something a computer could understand—starting with what they called an Automated Readability Index. After writing their algorithm and working out issues, the team reports that they were able to spot FBUs with an 80 percent accuracy rate after just ten posts. That is not high enough for web sites owners, of course, banning non-trolls by mistake 20 percent of the time could lead to driving away visitors—but it could possibly be used as a way to assist moderators.

More information: Antisocial Behavior in Online Discussion Communities, arXiv:1504.00680 [cs.SI] arxiv.org/abs/1504.00680

Abstract
User contributions in the form of posts, comments, and votes are essential to the success of online communities. However, allowing user participation also invites undesirable behavior such as trolling. In this paper, we characterize antisocial behavior in three large online discussion communities by analyzing users who were banned from these

communities. We find that such users tend to concentrate their efforts in a small number of threads, are more likely to post irrelevantly, and are more successful at garnering responses from other users. Studying the evolution of these users from the moment they join a community up to when they get banned, we find that not only do they write worse than other users over time, but they also become increasingly less tolerated by the community. Further, we discover that antisocial behavior is exacerbated when community feedback is overly harsh. Our analysis also reveals distinct groups of users with different levels of antisocial behavior that can change over time. We use these insights to identify antisocial users early on, a task of high practical importance to community maintainers.

Citation: Algorithm able to identify online trolls (2015, April 14) retrieved 5 May 2024 from https://techxplore.com/news/2015-04-algorithm-online-trolls.html