New algorithm can distinguish cyberbullies from normal Twitter users with 90% accuracy

16 September 2019

Credit: CC0 Public Domain

A team of researchers, including faculty at Binghamton University, have developed machine learning algorithms which can successfully identify bullies and aggressors on Twitter with 90 percent accuracy.

Effective tools for detecting harmful actions on social media are scarce, as this type of behavior is often ambiguous in nature and/or exhibited via seemingly superficial comments and criticisms. Aiming to address this gap, a research team featuring Binghamton University computer scientist Jeremy Blackburn analyzed the behavioral patterns exhibited by abusive Twitter users and their differences from other Twitter users.

"We built crawlers—programs that collect data from Twitter via variety of mechanisms," said Blackburn. "We gathered tweets of Twitter users, their profiles, as well as (social) network-related things, like who they follow and who follows them."

The researchers then performed natural language processing and sentiment analysis on the tweets themselves, as well as a variety of social network analyses on the connections between users. The researchers developed algorithms to automatically classify two specific types of offensive online behavior, i.e., cyberbullying and cyberaggression. The algorithms were able to identify abusive users on Twitter with 90 percent accuracy. These are users who engage in harassing behavior, e.g. those who send death threats or make racist remarks to users.

"In a nutshell, the algorithms 'learn' how to tell the difference between bullies and typical users by weighing certain features as they are shown more examples," said Blackburn.

While this research can help mitigate cyberbullying, it is only a first step, said Blackburn.

"One of the biggest issues with cyber safety problems is the damage being done is to humans, and is very difficult to 'undo,'" Said Blackburn. "For example, our research indicates that machine learning can be used to automatically detect users that are cyberbullies, and thus could help Twitter and other social media platforms remove problematic users. However, such a system is ultimately reactive: it does not inherently prevent bullying actions, it just identifies them taking place at scale. And the unfortunate truth is that even if bullying accounts are deleted, even if all their previous attacks are deleted, the victims still saw and were potentially affected by them."

Blackburn and his team are currently exploring pro-active mitigation techniques to deal with harassment campaigns.

The study, "Detecting Cyberbullying and Cyberaggression in Social Media," was published in *Transactions on the Web*.

More information: Detecting Cyberbullying and Cyberaggression in Social Media, arXiv:1907.08873 [cs.SI] arxiv.org/abs/1907.08873

Provided by Binghamton University