Assessing the toxicity of Reddit comments

August 18 2022

Credit: CC0 Public Domain

New research, published in *PeerJ Computer Science*, which analyzes over 87 million posts and 2.205 billion comments on Reddit from more than 1.2 million unique users, examines changes in the online behavior of users who publish in multiple communities on Reddit by measuring "toxicity."

User behavior toxicity analysis showed that 16.11% of users publish toxic posts, and 13.28% of users publish toxic comments. 30.68% of users publishing posts, and 81.67% of users publishing comments, exhibit changes in their toxicity across different communities—or subreddits—indicating that users adapt their behavior to the communities' norms.

The study suggests that one way to limit the spread of toxicity is by limiting the communities in which users can participate. The researchers found a positive correlation between the increase in the number of communities and the increase in toxicity but cannot guarantee that this is the only reason behind the increase in toxic content.

Various types of content can be shared and published on social media platforms, enabling users to communicate with each other in various ways. The growth of social media platforms has unfortunately led to an explosion of malicious content such as harassment, profanity, and cyberbullying. Various reasons may motivate users of social media platforms to spread harmful content. It has been shown that publishing toxic content (i.e., malicious behavior) spreads—the malicious behavior of non-malicious users can influence non-malicious users and make them misbehave, negatively impacting online communities.

"One challenge with studying online toxicity is the multitude of forms it takes, including hate speech, harassment, and cyberbullying. Toxic content often contains insults, threats, and offensive language, which, in turn, contaminate online platforms. Several online platforms have implemented prevention mechanisms, but these efforts are not scalable enough to curtail the rapid growth of toxic content on online platforms. These challenges call for developing effective automatic or semiautomatic solutions to detect toxicity from a large stream of content on online platforms," say the authors, Ph.D. (ABD) Hind Almerekhi, Dr. Haewoon Kwak and Professor Bernard J. Jansen.

"Monitoring the change in users' toxicity can be an early detection method for toxicity in online communities. The proposed methodology can identify when users exhibit a change by calculating the toxicity percentage in posts and comments. This change, combined with the toxicity level our system detects in users' posts, can be used efficiently to stop toxicity dissemination."

The research team, with the aid of crowdsourcing, built a labeled dataset of 10,083 Reddit comments, then used the dataset to train and fine-tune a Bidirectional Encoder Representations from Transformers (BERT) neural network model. The model predicted the toxicity levels of 87,376,912 posts from 577,835 users and 2,205,581,786 comments from 890,913 users on Reddit over 16 years, from 2005 to 2020.

This study utilized the toxicity levels of user content to identify toxicity changes by the user within the same community, across multiple communities, and over time. For the toxicity detection performance, the fine-tuned BERT model achieved a 91.27% classification accuracy and an Area Under the Receiver Operating Characteristic Curve (AUC) score of 0.963 and outperformed several baseline machine learning and neural network models.

More information: Hind Almerekhi et al, Investigating toxicity changes of cross-community redditors from 2 billion posts and comments, *PeerJ Computer Science* (2022). DOI: 10.7717/peerj-cs.1059

Provided by PeerJ