Exploring the underpinnings of shadowbanning on Twitter

20 January 2021, by Ingrid Fadelli

Credit: dole777, Unsplash

In recent years, social media platforms have been developing and implementing a variety of strategies to moderate content published by their users and ensure that it is not offensive or inappropriate. This has sparked significant debate, with some users claiming that these techniques hinder freedom of speech online.

Researchers at Inria, IRIT/ENSHEEIT and LAAS/CNRS have recently carried out a study investigating a renowned method of moderating content on social media platforms known as shadowbanning. Shadowbanning occurs when a social media site intervenes in the online activity of a specific user without their knowledge, for instance, by making their posts or comments invisible to other users. This is often achieved using decision-making algorithms or other computational techniques that are trained to identify posts or comments that could be considered inappropriate.

"As researchers, our subject of study is the understanding of interactions users can have with decision-making algorithms," Erwan Le Merrer, one of the researchers who carried out the study, told TechXplore. "These algorithms are often proposed in a black-box form, meaning that users know nothing about their functioning, but face their decisions as a consequence of the data they provide. We questioned automated moderation algorithms on social networks as an example of such decision-making algorithms."

The researchers set out to examine the underpinnings of shadowbanning on a specific social media platform: Twitter. They decided to focus on Twitter because its moderation of user content has received significant attention over the past few years.

"We relied on some open-sourced code that can detect some restrictions imposed on users and the visibility of their profiles, Tweets or interactions," the researchers explained. "We improved this code to support massive test campaigns and inspected the tweets visibility of around 2.5 million twitter users."

After compiling a dataset containing information related to the visibility of Tweets posted by users on Twitter, the researchers used it to try to understand the reasons why some users might have been subjected to shadowbanning. To do this, they analyzed the data they collected using standard mining approaches, testing two different hypotheses of why some users' visibility on Twitter might have been hindered.

The first hypothesis was that that the limitations on the visibility of some users' posts were caused by bugs or platform malfunctions. The second was that shadowbanning propagates like an epidemic across users who interact with one another.

"Since at some point, Twitter claimed that they were not using shadowbanning methods (referring to problems being bugs), we decided to leverage statistical methods to test the likelihood of such bug scenario, which should be uniformly distributed across users and hence across our data," Le

Merrer said. "We found out that several sampled populations were affected quite differently (e.g., politicians and celebrities less than bots or randomly sampled users)."

The results of the analyses show that the hypothesis that shadowbanning occurs due to bugs or errors in Twitter's system is statistically unlikely. Interestingly, they also observed that friends or followers of users who have been shadowbanned are more likely to be subjected to shadowbanning.

"To replace the unlikely bug narrative proposed by Twitter with another scenario, we devised a model that captured the frequently encountered clusters of shadowbanned users," the researchers said. "In other words, our model shows that shadowbanned users are more likely to have shadowbanned friends. This prevalence of shadowbanning around some users and their contacts is really questioning Twitter's statement about its moderation practices."

This recent study sheds some light on the dynamics and mechanisms of shadowbanning, revealing that there are often clusters of shadowbanned users who interact with one another. This could be due to decision-making algorithms learning to classify connections of shadowbanned users as other potential candidates for shadowbanning. It could also be caused by the algorithm targeting words frequently used within specific communities.

In the future, the researchers hope to conduct further investigations examining the underpinnings and limitations of machine-based systems for online content moderation and recommendation.

"We plan to pursue other investigations into algorithmic black boxes," they said. "Online services now expose their users to a large quantity of these systems (i.e., recommendation systems, credit scoring, raking of many sorts, etc.), so the choice will be difficult."

More information: Setting the record straighter on shadow banning. arXiv:2012.05101 [cs.SI]. arxiv.org/abs/2012.05101 , to be presented at INFOCOM 2021.