Researchers are creating science-backed tools to improve social media content moderation policies

November 7 2023, by Julia Cohen

Credit: Pixabay/CC0 Public Domain

Flagging, demotion, and deletion of content; temporary or permanent suspension of users—these are some of the interventions used to keep social media platforms safe, trustworthy, and free from harmful content. But what is the best way for these interventions to be implemented?

Luca Luceri, a research scientist at USC's Information Sciences Institute (ISI), is part of a team that is using science to guide social media regulations.

Luceri is working on CARISMA (CAll for Regulation Support In Social MediA), an interdisciplinary research project with the goal of "establishing a clear, traceable, and replicable methodological framework for assessing policies that effectively mitigate the harms of online actors responsible for abusive and illicit behavior."

But in order to assess social media content moderation policies, first, they must understand them. Luceri explained, "Content moderation strategies change frequently. They're not clearly communicated or transparent. There are no guidelines about the potential interventions, for example, how many times you have to do a certain action to be suspended temporarily or permanently."

He has recently co-authored two papers for CARISMA. "These papers are the first attempt to better understand how moderation policy strategies work, if they work, and what kind of misbehavior they are capable of identifying and moderating," he said.

The 'when,' 'how,' and 'what' of suspended accounts

Luceri worked alongside Francesco Pierri, a former postdoctoral researcher at ISI who is now Assistant Professor of Data Science at Politecnico di Milano, to co-author the *EPJ Data Science* paper "How Does Twitter Account Moderation Work? Dynamics of Account Creation and Suspension on Twitter During Major Geopolitical Events."

Prior research shows that there is a spike in creation and suspension of Twitter accounts in connection with major geopolitical events. Because of this, said Luceri, "We wanted to look at how Twitter dealt with new

accounts that were created in correspondence with major geopolitical events." The team chose two global political events: Russia's invasion of Ukraine and the 2022 French Presidential election.

They analyzed over 270M tweets in multiple languages to show that the increase in activity on Twitter is accompanied by peaks in account creation and abusive behavior, exposing legitimate users to spam campaigns and harmful speech.

The results?

1. Timing. They found that Twitter is more proactive in content moderation of recently created Twitter accounts compared to those with a longer lifespan.
2. Behavior. They observed that, compared to legitimate accounts, suspended accounts show an excessive use of replies, excessive toxic language, and a higher level of activity in general. Additionally, suspended accounts interact more with legitimate users, as opposed to other suspicious accounts.
3. Content. They found that the suspended accounts frequently shared harmful and spam messages.

These findings help shed light on patterns of platform abuse and subsequent moderation during major events, and they are the type of insights the CARISMA team is looking for when reverse engineering social media platforms' content moderation policies.

It's all connected

In a second CARISMA paper, "The Interconnected Nature of Online Harm and Moderation: Investigating the Cross-Platform Spread of Harmful Content between YouTube and Twitter," Luceri and his co-authors studied how one platform could leverage another platform's

moderation actions. This paper appears in *Proceedings of the 34th ACM Conference on Hypertext and Social Media*.

The team analyzed "moderated YouTube videos" that were shared on Twitter. This refers to YouTube videos that were deemed problematic by YouTube's content moderation policy and were eventually removed from YouTube.

Using a large-scale dataset of 600M tweets related to the 2020 U.S. election, they sought out YouTube videos that were removed. Once they knew a video had been removed from YouTube by YouTube moderators, they looked at the behavioral characteristics, interactions, and performance of the video when it was shared on Twitter.

The results? Removed YouTube videos, when they are shared on Twitter prior to being removed, show different interaction and behavioral characteristics than non-removed (acceptable) YouTube videos.

1. They spread differently. "If we look at the diffusion of videos in the first week they are on Twitter, moderated [removed] videos have many more Tweets linking to them than videos that were non-moderated [non-removed]. So, the way the moderated video diffuses is much faster," said Luceri.

2. User behavior is different. The researchers observed that users sharing removed YouTube videos tend to passively retweet content rather than create original tweets. While users spreading non-removed videos were much more involved in creating original content.

3. Users themselves are different. The researchers observed that the users who share removed YouTube videos in relation to the 2020 U.S. election were far-right politically and backed Trump during

the 2020 US election. While the political leanings of users who spread non-removed YouTube videos were less extreme and more diverse. Additionally, they found that the users spreading removed YouTube videos are not necessarily bots, meaning that research in this field should not only target bots and trolls but instead consider the role of online crowds and more complex social structures on different social media platforms.

And a more general finding from the research team is that they have proven that harmful content originating in a source platform (i.e., YouTube) significantly pollutes discussion on a target platform (i.e., Twitter).

"This work," Luceri says, "highlights the need for cross-platform moderation strategies, but also shows that it can be practically valuable. Knowing that a certain piece of content was considered to be inappropriate or harmful on one platform, can inform operational strategies on another platform."

A content moderation simulator

The CARISMA team is using results from papers like these and others to create a methodological framework where they can experiment with content moderation strategies.

"We are building a simulator that mimics social networks, interactions, and diffusion of harmful content, such as misinformation or hateful and toxic content," said Luceri. "What we want to do with this framework is not only mimic the information ecosystems, but we want to understand the potential impact of policy instruments."

He offered examples of how they might experiment in the simulator. "What would be the follow-on effects if a certain piece of

misinformation content was removed; versus what if the user was temporarily suspended; versus what if the user was permanently suspended. What would be the effect after one hour? After seven days? Or if we don't remove it at all?"

He continued, "What happens if we remove accounts that violate certain policies and how does that compare to what would happen if, instead, we provided those users with some nudges that tend to improve the quality of information they share?"

In the end, the simulator and the CARISMA project on the whole will provide quantitative evidence of the impact and effect of policy instruments that might be beneficial to mitigate harmful behaviors on social media.

"The hope is that this tool might be used by policymakers and regulators to evaluate the efficiency and the efficacy of policy instruments in a clear, traceable, and replicable way," said Luceri.

"The Interconnected Nature of Online Harm and Moderation: Investigating the Cross-Platform Spread of Harmful Content between YouTube and Twitter" was presented at ACM HyperText 2023, where it was nominated for the Best Paper Award.

More information: Francesco Pierri et al, How does Twitter account moderation work? Dynamics of account creation and suspension on Twitter during major geopolitical events, *EPJ Data Science* (2023). DOI: 10.1140/epjds/s13688-023-00420-7

Valerio La Gatta et al, The Interconnected Nature of Online Harm and Moderation, *Proceedings of the 34th ACM Conference on Hypertext and Social Media* (2023). DOI: 10.1145/3603163.3609058