I. PROJECT TITLE AND TEAM

Title/Question:

Factors Influencing How Quickly Content Is Removed from Social Media

Team:

Ryan Winston (RWinston1)

II. PROJECT SUMMARY

Over the past 20 years, social media has become an integral part of many individual's lives, as it has redefined the way that people communicate with each other, consume information, and engage with the world around them. Given social media's immense impact on the lives and identities of its users, it is important for the public to have a full understanding of the platforms they regularly use. In recent years, there have been significant efforts taken by the general public, organizations, and governments to understand how social media companies contribute to the general welfare of the public.

This paper endeavors to explore how social media companies censor content, regardless of whether the platform is justified doing so. Specifically, the aim of this project is to investigate what factors influence how quickly content is censored on social media platforms; some of these factors include the social media platform, the manner in which the content was detected, the content form, and the grounds on which the content was removed, among others.

This is a deeply important issue to explore, as having an understanding of the speed at which content in violation of the platforms' guidelines is removed, and why, enables users to be more informed about how they use these platforms. For example, a social media company that is slow to take down sexual content and images from their platform, might not be a platform suitable for children. Similarly, a site that is quick to identify and remove content that qualifies as misinformation may be a more trusted news source. These are only a few examples of the countless hypothetical scenarios in which the results of this study might improve users' interaction with social media.

III. RELATED WORK

Similar Studies:

Currently, there are not many related studies pertaining to social media censorship speeds. The only study that is directly related to the focus of this paper was done by researchers at Rice University and the University of New Mexico, who exclusively analyzed Weibo, a Chinese social media platform similar to Twitter. They found that ~30% of deletions on Weibo occurred within 5-10 minutes from the time the content was posted, and 90% within 24 hours. The researchers also investigated the mechanisms that Weibo used to identify content they later deleted, but that is not the primary concern of this project. Given that this paper was published in 2013, it is important to reevaluate the findings of this study to not only see if there have been any significant changes in the past decade, but also to expand this study to other platforms that are more popular such as X, TikTok, and Facebook, for example.

Zhu, Tao, David Phipps, Adam Pridgen, Jedidiah R. Crandall, and Dan S. Wallach. "The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions." Arxiv. Last modified March 4, 2013. https://arxiv.org/abs/1303.0597.

Boyd, Jade. "Study shows just how fast censorship can occur in social media." Rice University. Last modified March 13, 2013. https://news2.rice.edu/2013/03/13/study-shows-just-how-fast-censorship-can-occur-in-social-media-2/.

Additional Information Sources:

The source below details some of the difficulties working on TikTok's platform, as the process by which content is deemed unacceptable is automated. This has made it hard to unpack or completely understand how decisions are made. While this project does not aim to resolve the bugs and errors noted in the article, it does provide useful background information and a brief history of TikTok's automated content moderation process.

Ohlheiser, A.W. "Welcome to TikTok's endless cycle of censorship and mistakes." MIT Technology Review. Last modified June 13, 2021. https://www.technologyreview.com/2021/07/13/1028401/tiktok-censorship-mistakes-glitches-apologies-endless-cycle/.

The EU has more stringent standards regarding online hate speech and harmful content, in which they expect flagged content to be reviewed and/or taken down within 24 hours after it was flagged or posted.

Laub, Zachary. " Hate Speech on Social Media: Global Comparisons." Council on Foreign Relations. Last modified June 7, 2019. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons.

IV. PROJECT PREPARATION AND PREREQUISITES

This study will rely on the information contained within the DSA Transparency Database. The database contains information regarding content that was removed from various platforms, as well as why the content was removed. Specifically, the database contains information regarding how they will censor the content, the basis for their decision, the reasoning for their decision, how the decision was made, the category that the content falls under, the content type, the date and time at which the content was released, and the date and time at which the content was removed. There are several other variables included in the data that pertain to specific decision outcomes, but the aforementioned variables are those most relevant to this project.

This project will analyze content posted to TikTok and X, as they are among the most popular social media platforms in the world. Given that the DSA Transparency Database releases new data daily, there is no need to conduct independent data collection.

V. EVALUATION

I plan to evaluate the data by running a LASSO regression to see which are the most consequential/significant factors in explaining how quickly content is removed from

X and TikTok. By conducting this process on two different platforms, I would have the ability to present and directly compare the weights on all the factors. This could help to elucidate some of the differences between a largely automated decision process, in the case of TikTok, and a more manual one, in the case of X. Additionally, I could narrow in on specific variables regarding the nature of the content to potentially reveal more about the platforms' ability to ensure the safety of users, or the trustworthiness of its information.

VI. ETHICS

Although the data is related to censored content, which sparks the debate around whether or not it was justly taken down, that is not within the scope of this project's investigation. Instead, there may be ethical implications from this project if the findings show that there is a very long or short period of time between when content is posted and censored. If the time is short, some might come to the conclusion that these platforms can wait longer to remove content, if it means eliminating the frequency of falsely flagging content. If that time is long, some might conclude that they should speed up the removal process, even if it means wrongly deleting some content. There is no correct way to handle this issue, but this project might provide more information regarding how these platforms should act.