CSC 555 - Report R1

The Impact of Country-Specific Content Moderation Policies on X's User Engagement: A Comparative Analysis of Retweet Behavior for Blocked or Flagged Content.

Ashwattha Phatak, Ajay Chundi, Maverick Middleton

RELEVANT LITERATURE

1. Elmas, Tuğrulcan, Rebekah Overdorf, and Karl Aberer. "A dataset of state-censored tweets." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 15. 2021. https://doi.org/10.48550/arXiv.2101.05919
2. Sanderson, Zeve, et al. "Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform." Harvard Kennedy School Misinformation Review (2021). https://doi.org/10.37016/mr-2020-77

The first paper studies 583,000 censored tweets and performs exploratory data analysis using the temporal data, country the tweet is censored in, and the country of origin. The data and code associated with the paper were obtained because they were published to assist studies of government censorship, hate speech detection, and the effect of censorship on social media users.

The second paper studies the difference in reach of Tweets simply flagged by Twitter as misleading (a "soft intervention,") and Tweets that were blocked from being able to be retweeted, liked, or replied to (a "hard intervention.") The account being studied in this case is that of former President Donald J. Trump in the time frame of November 1, 2020 (2 days before the 2020 Presidential Election) through January 8, 2021 (the day that his account was suspended.) The data and code relating to this study was obtained.

WHAT PROBLEM ARE YOU ADDRESSING?

Social media platforms regularly censor and flag posts that violate certain policies. The policies in question may be policies of the platform itself, or certain governments across the world. Social media platforms have strict policies when it comes to posts engaging in misinformation or hate speech not only to protect their users, but also to ensure that advertisers do not leave the platform. Additionally, these platforms also have to flag or censor certain posts to adhere to the rules of the country in which the platform is active. Not adhering to the rules of world governments and judicial systems can result in the platform being entirely banned in a certain country, costing the platform millions of users and advertising revenue from that area.

The entire purpose of censoring or flagging a post on social media - the reason why so many countries have lately paid special attention to this issue - is to severely limit its reach and the number of people that

see the post. Social media platforms also want to crank down on posts with misinformation or hate speech for this reason, so that the posts reach will be limited and the number of people that believe the information will be minimized. Otherwise, there may be a case when posts spreading wild misinformation and hate speech take over a platform, resulting in advertisers pulling out and serious investigations into a platform's management.
So can it really be factually stated that censoring or flagging certain posts actually results in reduced spread? And because policies on content moderation and censorship vary so much throughout the world, we will provide comparative analysis across countries to see how moderation policies impact retweet behavior using visualizations and ratings of country-specific moderation policies based on their alignment with public sentiment, helping users determine if a country's policy is too strict or too lenient.
And even when a platform does limit the reach of the post in an extremely strict way that prevents any interaction with the post, does that post really stop spreading? What prevents the users from discussing it without interacting with the post directly - whether it's on the same platform, or another platform? How does the reach of a flagged or censored post differ when the user is a person with a massive following vs if the user is a common person? In the case of smaller accounts, even if the exact posts by smaller accounts are censored and limited, are there still widespread discussions across platforms regarding the same topic as what the original users were posting about?

WHY IS THIS PROBLEM IMPORTANT?

This project addresses the growing tension between content moderation and public discourse on platforms like Twitter. By analyzing how country-specific moderation policies affect public engagement, such as retweet behavior for flagged or blocked content, we can assess the effectiveness of these policies and their alignment with public sentiment. This is crucial for improving transparency and trust in digital content regulation. The platform will help social media companies refine moderation algorithms, assist policy makers in evaluating regulations, and support academics studying digital governance, censorship, and user engagement. The analysis will focus on retweet behavior in relation to country-specific moderation policies.

The hypothesis we hope to verify is the following :
Countries with moderation policies more aligned with public views will see fewer retweets for flagged content, whereas those with less-aligned policies will witness continued engagement with blocked or flagged content.

HOW WILL WE ADDRESS THIS PROBLEM?

To address this problem, we intend on analyzing relevant sets of tweets. Our main focus will be analyzing censored tweets, included in the dataset of the first article. We will start by downloading the archive, and extracting all necessary files to put into relevant batches. Using provided python scripts, we can mine the data and parse it into .csv files to serve as our main source of data for the study.

After acquiring the extracted dataset, we can focus on cleaning and preprocessing it to be easier to work with, identifying only relevant fields of information. We can clean up the dataset by additionally only keeping data that has been tagged. Removing any unhelpful metadata and fields can help us focus on the

more prominent data. Tweet data such as the country of origin, timestamp, tweet content (text, hashtags, potentially images), retweet counts, and user would all be relevant to work with.

The next step would be to identify any relevant trends and patterns within each country, to evaluate how censorship impacts the reach of tweets. We can segment the data by country, with listings of any relevant moderation policies. We can evaluate the spread of censored topics through the number of similar tweets and their retweets. Doing so would serve as a key method for testing our hypothesis and seeing how successful censorship can be at stopping the spread of flagged content. While the process may be altered by iterating through and extracting the contents of additional datasets for censored tweets, the analysis for supporting the hypothesis will remain the same.

WHAT ARE SOME ALTERNATIVES AND HOW CAN WE JUSTIFY OUR APPROACH?

In terms of alternatives we have multiple options that may provide additional insight to varying degrees. Instead of relying on the larger scale of varying tweets and their trends, we could instead focus on tweets within a specific time-frame relative to a political controversy. While it would not provide an overhead view of tweets and general censorship, it could be an opportunity to dive into additional context with a smaller dataset. Additionally, another option for handling the data could be extracting tags from raw tweets. Machine learning could be utilized to accomplish this, but within our current scope we believe it would be more efficient to simply focus on overall tagged data.

With our approach, it can be justified through how it can give a more quantifiable understanding of moderation policies and the impacts on the flow of information. The overall trend analysis can show how the behavior of masses of users can shift from data not being accessible due to censorship. Through utilizing existing datasets taken from twitter, the information is able to provide a more concise form of validation versus working with simulations. The scalability of this project can shift with the datasets available, allowing future analysis with given access to more information.

EVALUATION OF THE APPROACH

The approach will be evaluated on its ability to provide statistically significant insights into how country-specific moderation policies affect public engagement. The findings will be validated through multiple statistical tests (ANOVA, for example) and visual comparisons across countries. If the hypotheses hold, this method will offer a comprehensive understanding of the relationship between moderation actions and public behavior.