The Blackbox of Social Media Content Moderation

A first look into a novel Twitter dataset

Mael Kubli [†], Emma Hoes[‡] and Natalia Umansky[§]

December 11, 2023

Abstract

Out of all social media platforms, Twitter was known for sharing data with academia and industry. Yet, its content moderation practices remained opaque. In response, Twitter established the 'Twitter Moderation Research Consortium' (TMRC) in September 2022, allowing researchers to access data on accounts deleted for violating its Platform Manipulation and Spam Policy. This paper investigates Twitter's moderation decisions using TMRC14- and TMRC15-datasets, which include data from deleted accounts from 2016 to October 2022. Using network analysis and quantitative text analysis, we examine the alignment of Twitter's suspension decisions with its community guidelines. Our findings suggest that Twitter was generally accurate in its suspensions, staying well within the policies established in the community guidelines. However, they also highlight the need for more transparency and suggest avenues to understand the intricacies of content moderation, underscoring the importance of broader initiatives like TMRC.

***Keywords*—-** Twitter, Content Moderation, Twitter Moderation Research Consortium, Platform Manipulation and Spam Policy, Computational Methods

[†]PhD Candidate, University of Zurich
[‡]Postdoctoral Research Fellow, University of Zurich
[§]Postdoctoral Research Fellow, University of Zurich

1 Introduction

In the past couple of years, content moderation practices of social media platforms such as Facebook and Twitter have become an increasingly contested (political) issue (Alizadeh et al., 2022; Gillespie, 2018; Klonick, 2017, 2020; Siapera & Viejo-Otero, 2021; Stewart, 2021; Gillespie et al., 2023). This is in part because these platforms yield considerable political power by shaping what type of content millions of users are exposed to, and thus whose voices can be heard (Barrett & Kreiss, 2019; Caplan & Gillespie, 2020; Douek, 2021). While there are differences between platform's decisions to moderate some content but not other, much overlap between platforms' so-called 'community guidelines' exists. Community guidelines refer to the rules and standards that govern user behavior on social media platforms. Generally speaking, such guidelines are put in place to ensure that users are able to engage in online interactions that are safe, respectful, and inclusive. This is because it is feared that social media platforms can be leveraged to spread disinformation, incite violence, and promote hate speech, leading to concerns about the impact of these platforms on democracy and free speech.

One of the most common forms of content moderation is the suspension or removal of user accounts. Social media platforms have the power to suspend or ban user accounts for a variety of reasons, including violating community guidelines, spamming, or engaging in abusive or harmful behavior. However, the process of account suspension is often opaque and difficult to understand, with little information provided about why and how platforms decide to remove or block accounts (Roberts, 2019; Urman & Makhortykh, 2023). Indeed, [...] "there are few good data available about how platforms make content moderation decisions [...] which leads to confusion among users and makes it more difficult to have an informed public debate about how to regulate Internet content in a way that protects freedom of expression and other legitimate interests " (Suzor et al., 2019, 1527).

In this paper, we turn to Twitter's community guidelines to investigate the extent to which Twitter's decision to suspend numerous accounts is indeed in line with their reported content moderation practices. We do so by relying on a novel dataset which was published in September 2022 under the 'Twitter Moderation Research Consortium' (TMRC). This was done with a means to "prioritize transparency by sharing more data on more issues to those who are studying content moderation", and share "data [...] about the networks we remove and technical information about the presumptive country of origin and information operations." [1] Broadly speaking, these datasets contain Twitter data about numerous accounts which have been deleted because they in some way violated Twitter's Platform Manipulation and Spam Policy.[2] Specifically, we compare the TMRC-14 and TMRC-15 datasets — which contain Tweets and user information belonging to accounts deleted from 2016 to October 2022 — to Twitter's community

[1] https://blog.twitter.com/en_us/topics/company/2022/twitter-moderation-research-consortium-open-researchers

[2] https://help.twitter.com/en/rules-and-policies/platform-manipulation

guidelines, and inductively analyse the users that were suspended, the content they shared on Twitter before their suspension, and the extent to which their suspensions are in line with Twitter's community guidelines. Even though Twitter (now 'X') has changed since and is still undergoing change, examining its (some of which are former) guidelines can help us gain insights into the platform's development and response to various challenges.

This paper is structured as follows: Firstly, we provide a brief overview of the community guidelines commonly found across various platforms, with a particular focus on Twitter's salient and distinctive policies. Specifically, we delve into Twitter's Platform Manipulation and Spam Policy, which serves as the primary reason for suspending the accounts listed in the TMRC14/15 datasets. Subsequently, we provide a detailed description of the datasets and outline the preprocessing steps undertaken to extract the relevant information from their original nested structure. Finally, employing quantitative text analysis and network analysis tools, we perform an inductive analysis of the TMRC14/15 datasets, comparing them to Twitter's content moderation guidelines to assess the level of adherence and transparency in Twitter's moderation practices as outlined in their reported guidelines.

2 Platform Community Guidelines

Facebook, Twitter, Instagram, TikTok, and YouTube are some of the most popular social media platforms that engage in content moderation to ensure that users comply with rules and policies. While each platform has its unique set of guidelines, they share some common content moderation processes and values (Hovyadinov, 2019; York & Zuckerman, 2019; Hallinan et al., 2022). We refer to these as community guidelines, defined as the rules and standards that commonly govern user behavior on social media platforms.

Broadly speaking, social media platforms rely on several processes when flagging (i.e., marking content such as Tweets or videos that violate community guidelines) and removing content (Urman & Makhortykh, 2023). Firstly, these platforms leverage Artificial Intelligence (AI) and Machine Learning (ML) algorithms to detect and flag potentially inappropriate content. [3] Such content may include hate speech, violence, nudity, and other violations of community guidelines. Most platforms also rely on users to report inappropriate content to their moderators, who then review the flagged content to determine whether it violates the platform's policies or not. Thirdly, these platforms employ human moderators who continuously perform small scale analyses to identify harmful content. In most cases, however, these content moderation decisions may be reverted. Users can choose to appeal the moderation decisions adopted by platforms like Facebook, Twitter, Instagram, TikTok, and Youtube, which may result in the

[3] See https://developers.facebook.com/docs/threat-exchange and https://www.interpol.int/en/Crimes/Crimes-against-children/International-Child-Sexual-Exploitation-database for examples of databases that platforms check posted content against.

content being reinstated if it had been erroneously flagged or removed.

Across popular social media platforms, the common values that drive content moderation practices are user safety, fostering a sense of community, promoting authenticity, and striving for diversity and inclusion. User safety, however, normally takes precedence as platforms actively work to prevent harassment, cyberbullying, and other harmful behaviors. By doing so, they aim to cultivate a supportive and inclusive digital community. Additionally, these platforms are committed to promoting authentic engagement and combating the dissemination of false information, propaganda, and fake accounts. Lastly, they endeavor to foster diversity and inclusion, ensuring that all voices are heard, irrespective of race, ethnicity, gender, or sexual orientation. Yet, in practice, it is unclear whether the reported content moderation guidelines are applied, and to what extent users may be affected by opaque and unaccounted decisions that fall outside the reported community guidelines (Gillespie, 2018). In this paper, we turn to Twitter as a case study to explore this further.

2.1 Twitter's Rules & Policies

In order to ultimately be able to explore the extent to which account suspensions on Twitter are in line with their guidelines, we now turn to a discussion of Twitter's rules. Considering Twitter's policies have been subject to change —especially following Elon Musk's acquisition of Twitter— and since our dataset contains account suspensions from September and October 2022, we rely on Twitter's active guidelines during that same period. [4]. To facilitate straightforward bench-marking of our data against Twitter's community guidelines, we present the guidelines separated by topic (content of a Tweet) and behavior (pattern of Tweets), respectively. All topics and behaviors are considered harmful and thus subject to content moderation practices and the suspension of accounts specifically.

2.1.1 Harmful Content

1. **Hate speech and harassment**: Twitter prohibits hate speech and harassment against individuals or groups based on race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. This includes targeted abuse, such as sending abusive messages or creating abusive accounts, as well as using slurs, epithets, or other offensive languages. Among it is also content that promotes or glorifies violence, hatred, or discrimination against individuals or groups based on race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. Examples of hateful conduct include using slurs or derogatory language, making demeaning or dehumanising remarks about a group, or promoting the superiority of one group over another. If Twitter becomes aware of any hateful

[4] See here: https://web.archive.org/web/20220901055609/https://help.twitter.com/en/rules-and-policies/twitter-rules

conduct on its platform, it may take various actions, including removing the offending content, limiting the account's functionality, or suspending the account responsible.

2. **Violence and threats**: Twitter does not allow threats of violence or the promotion of terrorism or violent extremism and the dissemination of manifestos. This includes direct or indirect threats of violence, as well as the glorification or celebration of violence.

3. **Child sexual exploitation**: Twitter has a zero-tolerance policy when it comes to this type of content. Twitter prohibits any content that promotes or glorifies child sexual exploitation, including child pornography and any material that sexualises minors. Twitter also works with law enforcement to report any instances of child sexual exploitation that it becomes aware of on its platform. In addition, if a tweet or account is found to be promoting or glorifying child sexual exploitation, Twitter will take swift action, including reporting it to the appropriate authorities and suspending the account.

4. **Abuse/harassment**: Abuse and harassment are prohibited on Twitter. Targeted harassment includes directing abuse towards a particular person or group of people. It can take many forms, such as sending threats, engaging in doxxing (sharing personal or identifying information) or encouraging others to harass someone. Additionally, wishing or hoping for someone to experience physical harm is also considered a violation of Twitter's policies on abusive behavior. Suppose Twitter becomes aware of abuse or harassment on its platform. In that case, it may take various actions, including removing the offending content, suspending the account responsible, or reporting it to the relevant authorities.

5. **Suicide and or Self-Harm**: Twitter prohibits any content that promotes or encourages self-harm, suicide, or eating disorders. This includes graphic descriptions or images that may encourage or glamorise self-harm or suicide. If Twitter becomes aware of any content that violates its policies on suicide and self-harm, it may take various actions, including removing the content, suspending the account responsibly, or providing resources and support to individuals who may be at risk.

6. **Spam**: Twitter prohibits spam, including using bots, fake accounts, or other automated tools to amplify or manipulate conversations on the platform artificially. This also includes attempts to artificially inflate metrics such as likes, retweets, and followers. An example of this are Tweets that (only) contain repeated mentions of specific user accounts.

7. **Misinformation / Manipulated Media**: Twitter takes action against misinformation that can cause harm, including misinformation about public health, civic integrity, and manipulated media. This includes false or misleading information about elections, COVID-19, or other public health crises, as well as deep fakes or other manipulated media designed to deceive people.

8. **Intellectual property**: Twitter respects intellectual property rights and requires users to do the same. This includes copyrights, trademarks, and other forms of intellectual property

9. **Adult content**: Twitter has policies against adult content, including pornography and sexual services. This also includes excessively graphic or violent content, such as depictions of extreme violence or gore.

2.1.2 Harmful Behavior

1. **Coordinated dissemination of harmful content**: Twitter works to identify and remove networks of accounts engaged in coordinated attempts to deceive or manipulate others on its platform, or disseminate any harmful content (as listed above, e.g., violence, harassment) in an organized fashion. This includes content related to civic processes, such as elections, political campaigns, and democratic institutions. Twitter prohibits any content that seeks to manipulate or interfere with these processes, such as *coordinated* attempts to spread false information about candidates or election procedures.

2. **Manipulation**: Twitter prohibits spam, including using bots, fake accounts, or other automated tools to amplify or manipulate conversations on the platform artificially. This also includes attempts to artificially inflate metrics such as likes, retweets, and followers or duplicate content. Examples of behavior that may result in removal or permanent suspension include: 1) Using automation or scripting to post duplicate content; 2) Operating one account or multiple accounts where most of the content promotes duplicate content resulting in spammy, inauthentic engagement; 3) Repeated participation in copypasta and duplicate Tweet efforts to promote content that is in violation of other Twitter Rules.

3. **Misleading and Deceptive Identities**: Twitter prohibits the creation and use of misleading and deceptive identities on its platform. This includes impersonation, where a user falsely claims to be someone else or creates a fake account to deceive others. Additionally, Twitter prohibits any attempt to manipulate the platform, such as creating multiple accounts to amplify a particular message or artificially inflate a user's influence. Suppose Twitter becomes aware of any accounts or behavior that violates its policies on misleading and deceptive identities. In that case, it may take various actions, including removing the content, suspending the account responsibly, or providing resources and support to users who may have been misled or harmed by the content. Twitter also works to identify and remove networks of accounts engaged in coordinated attempts to deceive or manipulate others on its platform.

3 Data & Methods

We rely on Twitters TMRC data-set, which was made available through Twitter on September 22^{nd} 2022[5]. The data is only available for academic use through an application form. The data-set covers two samples containing a total of 21 moderation events from August and September 2022. The first sample is called the TMRC14 which covers fifteen moderation events, while the second is called TMRC15 covering six events. The released data is by no means complete, since the data-sets provided are only about a select list of moderation events and do not include moderation's of single accounts. Furthermore it is unknown how many other moderation events of multiple users there are in the same time frame as covered by the data-set. Given the numbering of the TMRC folders, there is a high chance that there are at least thirteen more collections including many more moderation events. Nevertheless, this is — to the best of our knowledge — the first time a social media platform released some detailed data from moderated accounts. Although the number of moderated accounts are only very few and spread over many different regions which complicates a comprehensive analysis, in this paper we make a first attempt at doing so.

In total, both events contain the complete user data of 19,425 suspended users responsible for 23,528,985 million tweets. Furthermore, the data includes some information regarding the different moderation events including the suspension reasoning, stated origin of accounts and presumed origin (IP localization). Moderation events, in this context, refer to actions taken by Twitter in response to activities that violate its community standards and policies, especially pertaining to information operations. They encompass a range of activities such as detecting, investigating, and ultimately removing accounts involved in malicious activities. Figure 1 depicts the different countries of origin of the suspended accounts moderated in all 21 events.

3.1 TMRC14 & TMRC15

The two data-sets are similar in the information reported in them but different in size and time of creation. The first data-set contains a total of 17,408 accounts from 15 moderation events, while the second contains 2,017 accounts from 6 moderation events. In total, this data-set contains 23.50 million unique tweets of which 6,668,886 are original tweets. Across both data-sets, Twitter consistently attributes all moderation events to violations of the platform's manipulation and spam policy. However, in three events within the TMRC14 data-set and two events within the TMRC15 data-set, secondary reasons were provided. In the first set of cases, the moderation also materialized due to coordinated and baseless reporting of other accounts. In the latter cases, account suspensions were additionally enforced due to fabricated

[5] https://blog.twitter.com/en_us/topics/company/2022/twitter-moderation-research-consortium-open-researchers

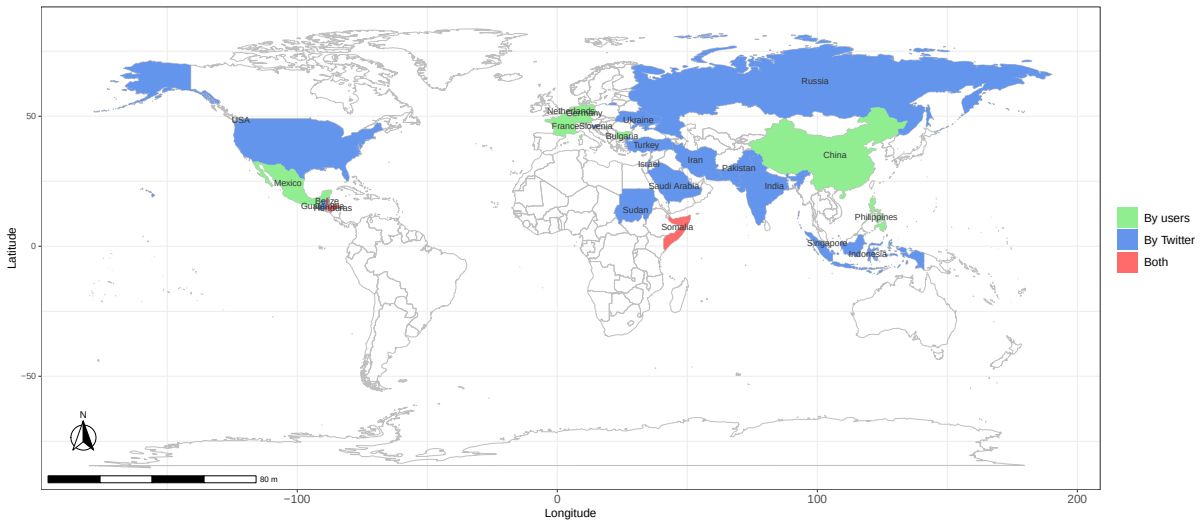

Figure 1: Map displaying both the user-reported locations and Twitter reported location of suspended accounts grouped by moderation events most often reported locations.

engagements with posts from other users. All moderation events were executed either after Twitter itself found the network of malicious accounts, or was informed about them by industry partners, researchers or intelligence companies.

3.2 Data Preparation

The data as provided by Twitter is a structured tree of folders and sub-folders containing json-files for each user, including a user, tweets, followers, and following files for each user as well as a folder containing the shared media files of each user. We ingested this data into as SQL database behind a secure firewall to be used only by the people granted access to the data by Twitter. The database is structured as shown in Figure 2.

This structuring of the data allows us to conduct our analysis of the data efficiently.The ingesting of the data follows a two-step process. In a first step, we process the files downloaded from Twitter by unzipping all files in all the moderation events in both the TMRC14 and TMRC 15 data-set. Then we ingest the information for a new moderation event into the moderations table, after which we ingest the users of said moderation event into the database one by one into the users table as well as the tweets of these users in the tweets table. This is straightforward since each user and tweet has a unique ID. For the followers and following tables we had to check for each user if the follower of the following user of said user is already in the table to keep these IDs unique. To keep the relationship of the users to their followers and following, we use a relationship table for each of the two tables. One thing we had to look

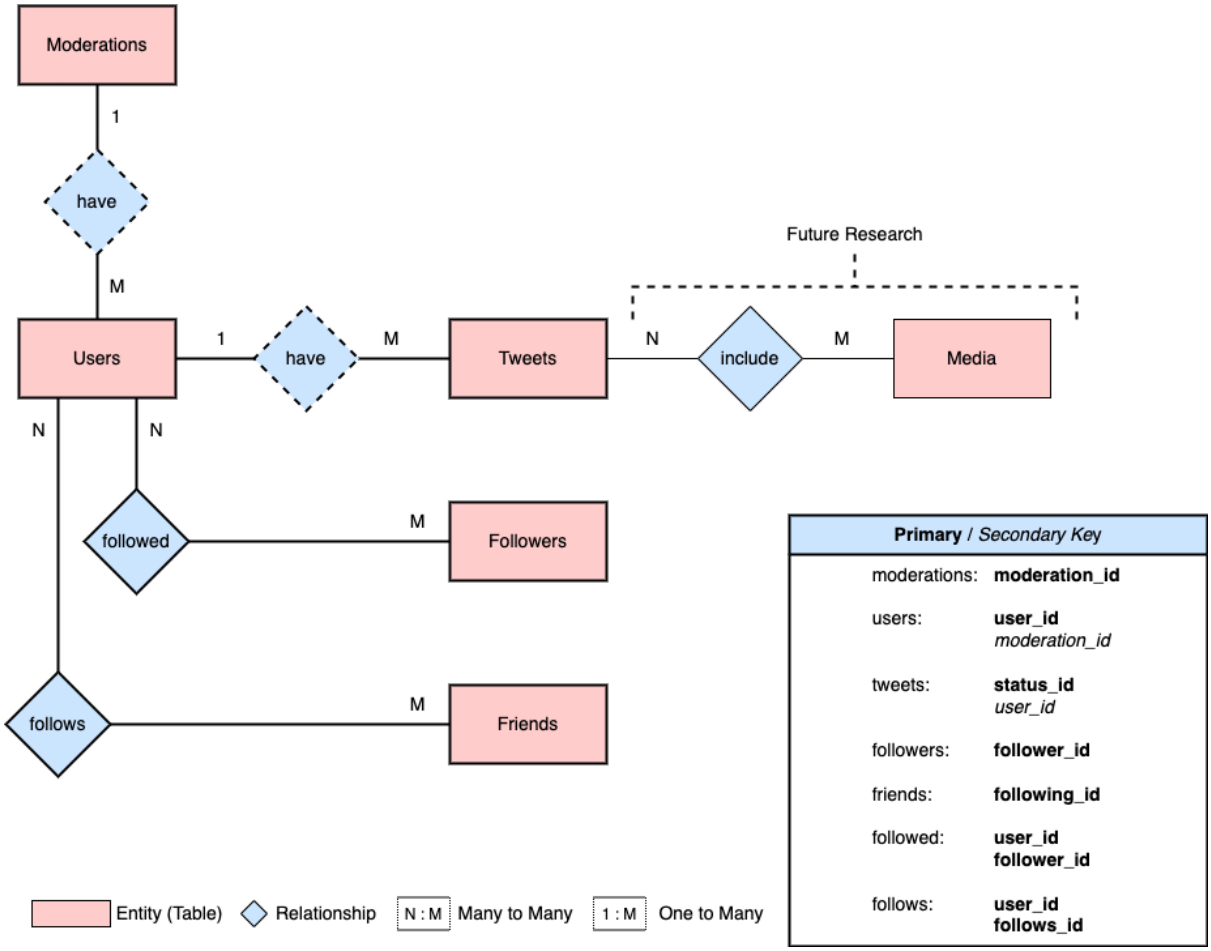

Figure 2: Entity relationship model of TMRC data for research useage.

out for is users with just a folder but no additional data in their folder. This includes also the users' account data itself. Hence, it looks like the reported data is not complete in all cases. In total, we end up with just 18,759 of 19,425 users in the database, since the missing 666 users had no information at all in their respective zip files.

The modelling of the database is expansible for the media files and if new moderation events are published we can easily update the tables in the future.

3.3 Methods

We rely on descriptive analytics, network analysis and topic modelling for our evaluation of the data. This allows us to produce a first insight into the moderation events.

First, we look at user metrics and aggregations of the tweets posted by the suspended users accounts. This includes correlation estimates between the account lifetime and tweet frequency as well as distributions of the different variables over all users and as facets between different geographical areas. Second, we computed follower networks of the suspended users only, as well as with other followers which occur more than 25 times in the full data-set as well as in the small multiples. Lastly, we use BERTopic to explore

what kind of content suspended accounts posted the most and what kind of topics got them suspended in specific events.

BERTopic is a method for generating topic representations from a collection of documents. It does so by converting each document into an embedding representation using a pre-trained language model. Subsequently, it reduces the dimensionality of the embeddings to improve the clustering process. Finally, it extracts topic representations from the clusters of documents using a class-based variation of the term frequency-inverse document frequency (TF-IDF) algorithm (Grootendorst, 2022). For our analysis, we use the standard pipeline as available through the library BERTopicopic in Python with one major change (Grootendorst, 2022). Instead of using the pre-trained Bert model from Grootendorst (2022) we use Barbieri et al. (2022) pre-trained xlm-t multilingual transformer model, which was optimised for tweets. This allows our model to better converge around the topics most often used in tweets, which are short texts mostly consisting of just one or two sentences. This embedding is used to cluster semantically similar documents.

In addition to this, we set the minimum document number for a topic to 2000 tweets, the number of neighbors for the cluster analysis to 100 and the minimum cluster size to 1000. Furthermore, we remove all stop-words and all terms from the TF-IDF, which occur in more than 90 % of tweets in the corpus. All other parameters are kept at their standard value.

To measure the content posted by the accounts, we only use the original tweets and quoted tweets of the accounts without the retweets posted in the last 12 months before the moderation event. This results in a substantial reduction of tweets with just 2,115,356 tweets left for the topic model. This has two advantages. First and foremost, it reduces the probability of seeing content unrelated to the moderation events. Second, it reduces the size of the model considerably.

The model results in 46 distinct topics which contain 82.4 % of all Tweets. The other 17.6 % of the tweets were not assignable to any of the 46 topics. We then looked at each topic assigning a title to each topic if possible. This is a manual process where we use the top 10 features per topic and the top 50 tweets per document to get an understanding of the substance of each topic.

We were able to detect 32 topics with enough substance to classify their content as relevant for the moderation of the collection of moderation events. At the same time, we also assigned each topic a moderation reason given through the community guidelines of Twitter as mentioned in section 2.1

In the last step, we employ Social Network Analysis (SNA), a research methodology to examine the patterns and structures of interrelationships between entities. This method is often applied to study the relational dynamics within social structures using networks and graph theory. Our analysis focused on a "friends follower network", a specific form of social network where nodes represent individuals and edges correspond to "follower" relationships between them. This type of network provided us with the potential

to understand the intricate dynamics and influence patterns among individuals within the network.

To visualize and interpret this complex web of relationships, we applied the Fruchterman-Reingold algorithm, a force-directed layout algorithm used in network visualizations. This algorithm employs a physical metaphor to position nodes (i.e., individuals) in the network, wherein nodes behave as if they are repelling each other like charged particles, while edges behave like springs holding nodes together (Fruchterman & Reingold, 1991). The equilibrium state achieved via this method allows for a balanced and interpretable representation of the network structure, resulting in distinct clustering, which facilitates further analysis of community and influence structures within the network.

4 Results

In this study, we comprehensively analyze the user's behavior, content and network characteristics on Twitter for all moderation events. Our analysis focuses on three distinct levels: user-level statistics, content-level statistics and network analysis.

4.1 Content analysis

It is important to note that several factors limit our analysis. First, the dataset we used is most likely selective and may not be representative of the entire moderation efforts by Twitter. Additionally, we only analyze data from two periods in 2022, which may not reflect the long-term trends and patterns in user behavior on the platform. Furthermore, our analysis only focuses on moderated users, and we need to know the completeness of the moderation data. Therefore, our findings may not accurately reflect the overall content production of moderated users on Twitter. Lastly, the datasets are from August and September 2022, which means the results are different from the moderation behavior today since the data published is from before the overtaking of Twitter by Elon Musk.

Despite these limitations, our study provides valuable insights into the behaviors and characteristics of moderated users on Twitter. By analyzing user activity, content production, and network dynamics, we shed light on the complex interactions between users in online communities. Our findings can inform the development of targeted interventions to promote healthy and productive online discourse, particularly for moderated users on Twitter.

At the user level, we examine the behaviors and characteristics of aggregate users over the different moderation areas by geographical separation into countries and country groups. This first level of analysis should make differences visible in the moderation of accounts over different areas and give a first insight into the characteristics of moderated users on Twitter.

How long-lived are moderated accounts on average and are there differences between languages and regions? In Figure 3 we can see the distribution of the account lifespan over different regions available

in the data. We are able to show that there are large differences within regions which are clearly skewed towards shorter lifetimes. On average for all regions together, the mean lifespan is 59.915 ($\pm$0.995) weeks. The interesting finding with the lifespan is the fact that — for the USA and Russia/Ukraine — the lifespan is the shortest on average. We further see that the distribution concentrates around more than one maximum for several regions. For example, in the Guatemalan case, we see three areas around which most account centres and not one. This shows that Twitter cannot detect all moderation-worthy networks of users within the same amount of time but sometimes misses them for a more extended period. At times it takes third parties to report them to Twitter. This is also clearly stated in the description of the moderation datasets, where they acknowledge moderation through reports by third parties.

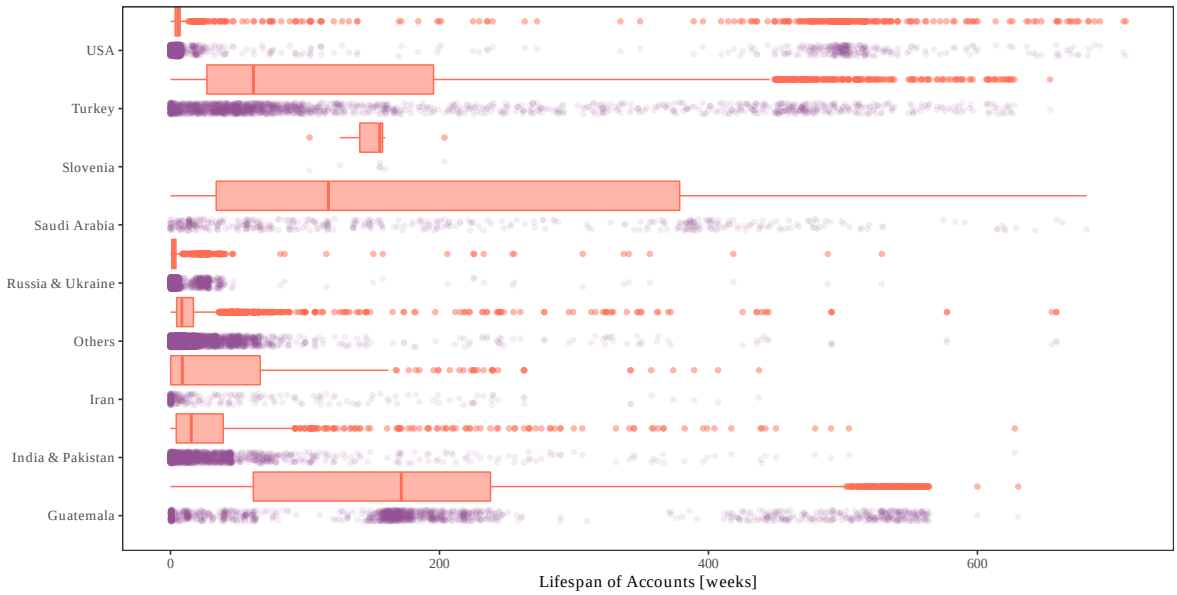

Figure 3: Figure shows the lifetime of users' accounts over different geographical areas.

We also display the distribution of followers and friends over the regions for the moderated users. Our analysis reveals that the number of followers for moderated users on Twitter is relatively high. This may indicate that moderated users are more visible and influential than non-moderated users on the platform. Additionally, we find that the number of friends of moderated users is high on average. Interestingly, this suggests that moderated users are actively engaged in the Twitter community and may have a broad network of connections.

However, our analysis also reveals bot-like accounts among moderated users. Specifically, we observed that for multiple moderation areas, there were accounts with zero or very few friends who follow many people. This pattern may indicate that these accounts are not authentic and may be used for spamming or other malicious activities.

At the content level, we analyze all original tweets posted in the last 12 months before the moderation using a BERTopic model to find common topics used by moderated accounts and compare them with

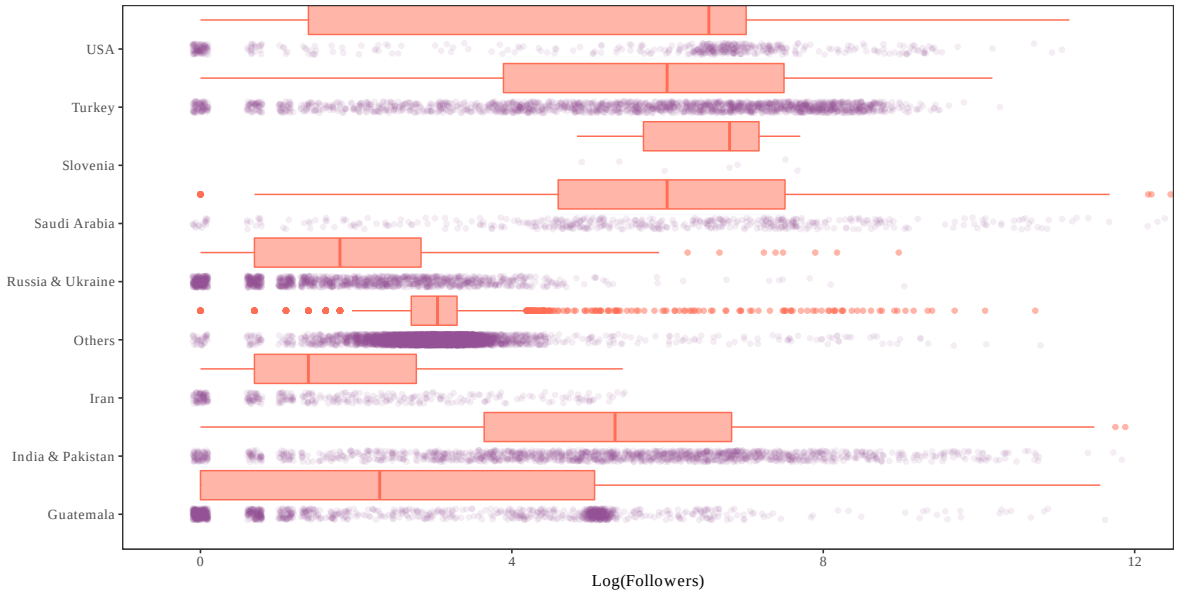

Figure 4: Figure shows the distribution of followers of users' accounts over different geographical areas.

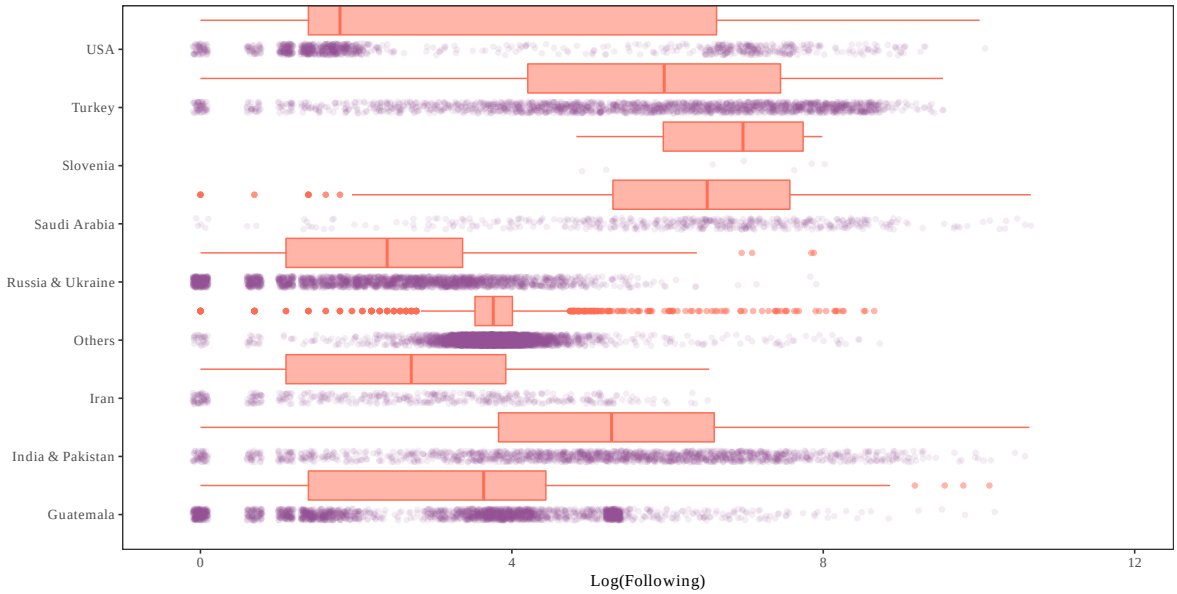

Figure 5: Figure shows the distribution of friends of users' accounts over different geographical areas.

Twitter's community guidelines. This allows us to show which content most likely resulted in the concurrent moderation of accounts for most moderation events.

We show in Figure 7 and Figure 8 that the identified topics of interest are clustering well towards the moderation events. Thus, we can be fairly confident that the identified topics represent the moderation reason behind the events. Unfortunately, some moderation events are not clearly attributed to one of the identified topics. This means we cannot make any assumptions about the moderation reason for using this method. Nevertheless, our approach works for the majority of events, giving us enough insight to draw conclusions on twitters reasoning for moderation while revealing some bias in their process from

what they share in the two datasets.

Let us look at the distribution of the moderation reason. We can see that for several moderation events we can identify the reasoning for suspending the accounts, while this is not possible for others. In section 2.1, we discuss all possible rules and policies that could lead to the suspension of accounts, grouped into different topics of harmful content and harmful behavior.

First, we examine the different topics identified in the moderation dataset which users discussed in the last 12 months before the moderation event. Our model identified 46 different topics from all original tweets posted in the 12 months leading up to the moderation event. Of these, 32 topics are of interest (1,511,453 Tweets), as they most likely violate a policy or rule of Twitter. Surprisingly, the topics also center around the different moderation events.

Figure 6 and Figure 7 display 16 different topics each, showing the share of the topic in the different moderation events. As we can see, most topics clearly align with just one event, with very little noise. Only five of the topics are shared with more than one moderation event, where the most likely event has less than 75 % of the topics tweets. One of them is the "Unclear" topic, where we do not know the actual topic as the most likely documents do not center around a clear issue or topic. The other four topics include, initially, the subject of Chinese misinformation, a theme shared between the two moderation events named TMRC 15 APAC 2 and TMRC 15 APAC 3. Both these events, which occurred in September 2022, involved accounts Twitter identified as originating from China, thereby linking them to the dissemination of Chinese misinformation. Hence, it is very plausible that both events center around Chinese misinformation in the US. Another case is the topic of the 2021 Taliban Offensive, which is relevant in several moderation events, as well as two topics about automated spamming of other accounts with friendly greetings and religious greetings. All other topics are clearly associated with just one moderation event. There are several important topics found, such as the topic Chinese Propaganda in Figure 6, targeting US accounts, which heavily centers around the moderation event TMRC 15 APAC 3 or the topic Propaganda for Imran Khan centered around the moderation event TMRC 14 APAC 3.

Next, we examine the policy behind the moderation of these topics and events. We can show six types of harmful content shared in these topics that got users suspended in the moderation events. These are Hate Speech and harassment, Misinformation, Violence and threats, Spam and Abuse/Harassment, as well as several events where the topic is not giving away any reason for moderation in our analysis. What is interesting is the fact that the reasons center somewhat around different moderation regions. In Figure 8, we can see that most of the moderation events occur due to the use of hate speech and harassment happened in the Asia Pacific region (APAC), while Misinformation seems to be the policy most responsible for moderation events in Europe (EUR). The same is true for moderation due to the dissemination of violence and threats, which happens to be the most common suspension reason in the

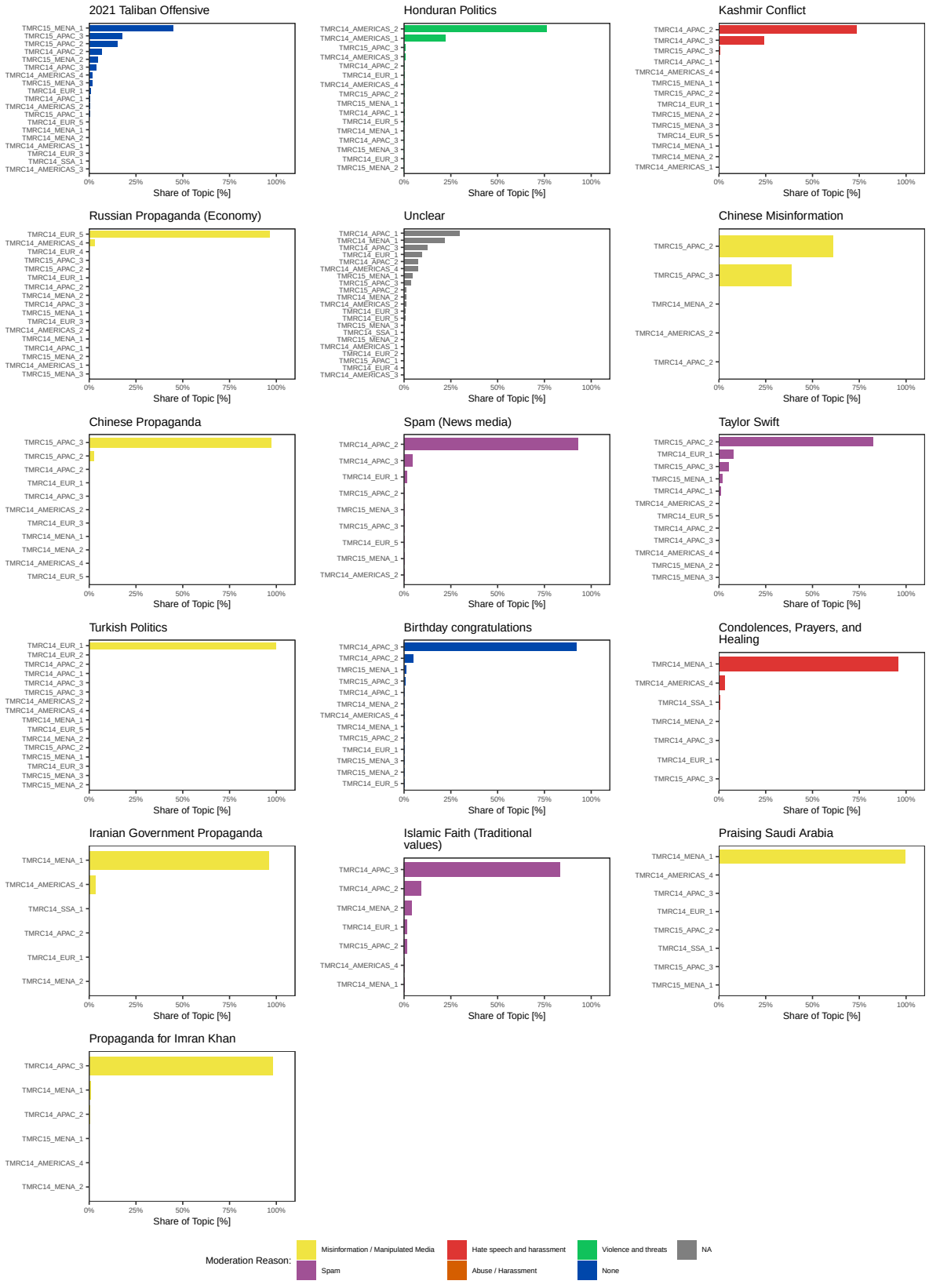

Figure 6: Figure displays the share of all possible moderation reasons faceted by the topic from the BERTopic model, where we were able to identify a coherent topic

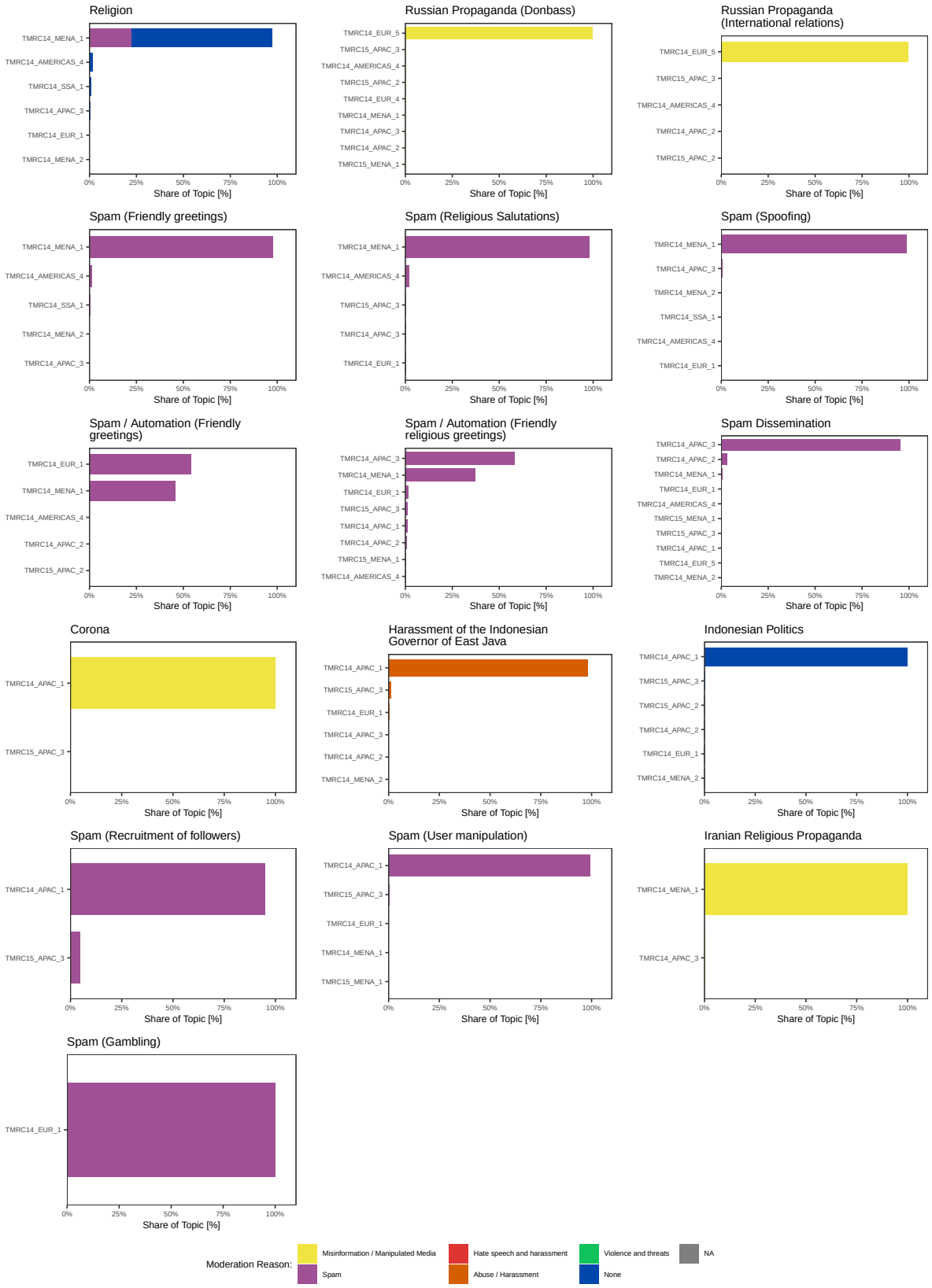

Figure 7: Figure displays the share of all possible moderation reasons faceted by the topic from the BERTopic model, where we were able to identify a coherent topic

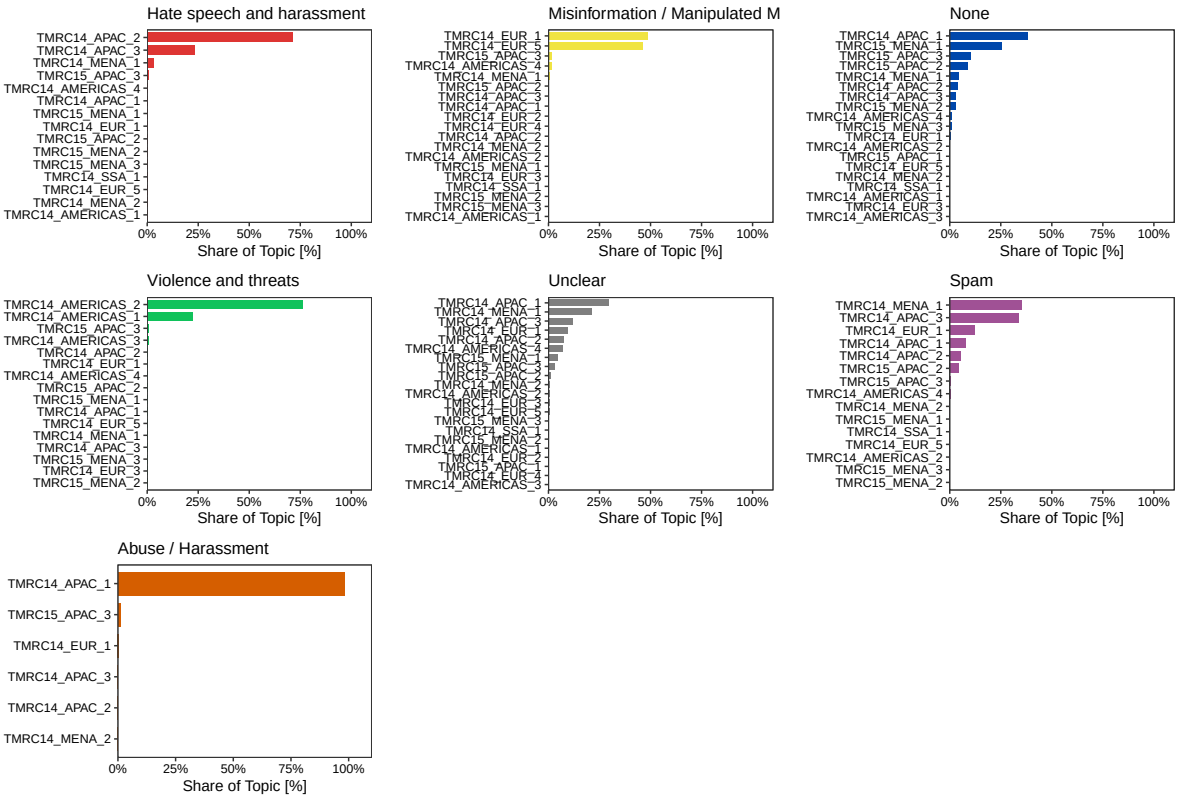

Figure 8: Figure displaying the share of Moderation Reason for all different Moderation Events in the Dataset

Americas (AMERICAS) Region.

4.2 Network analysis

Social Network Analysis (SNA) provides valuable insights into the connections and interactions between users on social media platforms. In this section, we explore the SNA results of the moderated accounts on Twitter and analyze the network characteristics to detect harmful behavior, particularly in relation to Twitter's moderation policies.

We first examine the network clusters formed by the moderated accounts, focusing on whether these clusters correlate with color coding indicating the accounts' geographical regions. The formation of distinct clusters based on region would suggest that the moderated accounts were not only acting simultaneously within their regions but also following each other, indicating coordinated behavior.

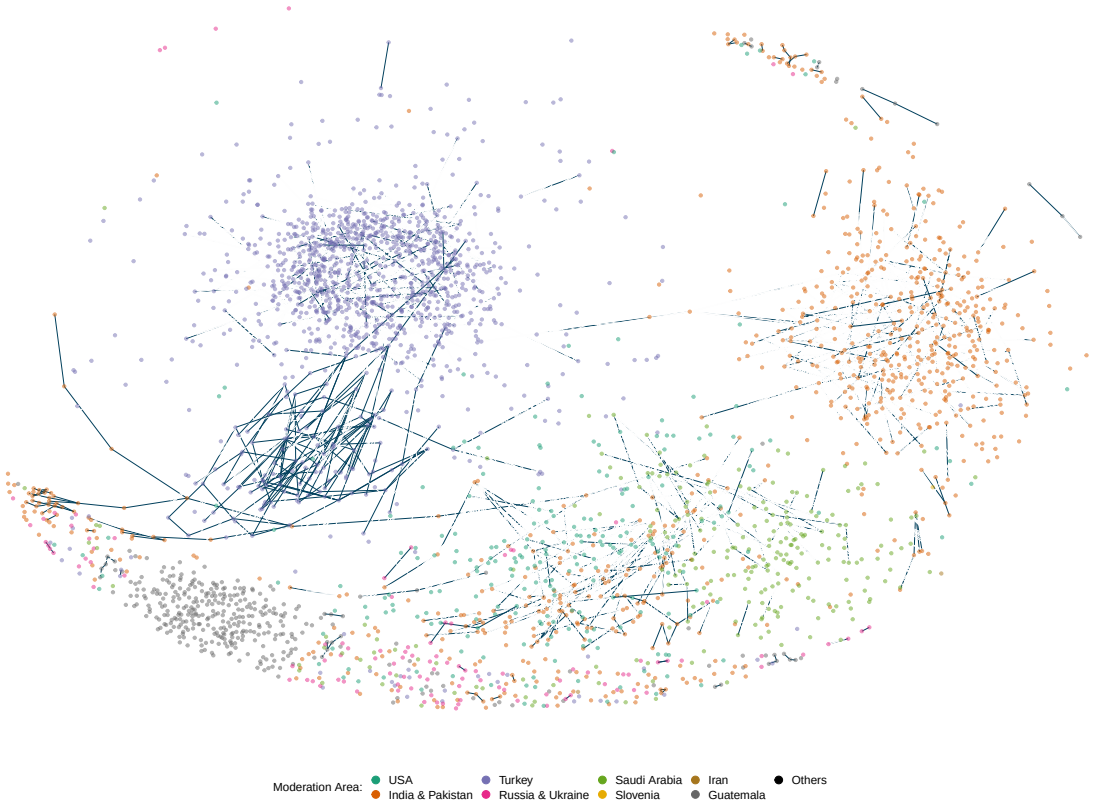

Figure 9: Network of moderated accounts color coded by location. Nodes are moderated users and edges are following relations.

Our SNA results displayed in Figure 9 reveal that the network clusters do indeed show a strong correlation with the color coding representing different geographical regions. This finding suggests that the moderated accounts were not acting in isolation but were part of coordinated efforts within their respective regions. The formation of region-specific clusters indicates the existence of coordinated dissemination of harmful content, which is in line with Twitter's moderation policy on identifying and removing networks of accounts engaged in organized attempts to deceive or manipulate others. Moreover, the fact that these accounts were following each other within their clusters strengthens the evidence of coordinated behavior. Coordinated networks of accounts with mutual connections can be indicative of attempts to amplify certain content artificially or engage in spammy, inauthentic engagement, both of which violate Twitter's Platform Manipulation and Spam Policy. The presence of coordinated networks in different regions also raises concerns about the use of misleading and deceptive identities, which is another policy violation according to Twitter's guidelines. The coordinated behavior within clusters indicates that these accounts may have been part of campaigns attempting to manipulate the spread of information on Twitter, potentially using fake accounts or impersonating others to deceive users.

To gain further insights into the potential impact of these moderated accounts and the harmful content they may have spread, we extend the analysis to include not only the moderated accounts but

also their followers. This expanded network analysis allows us to assess the reach and influence of the accounts spreading harmful content.

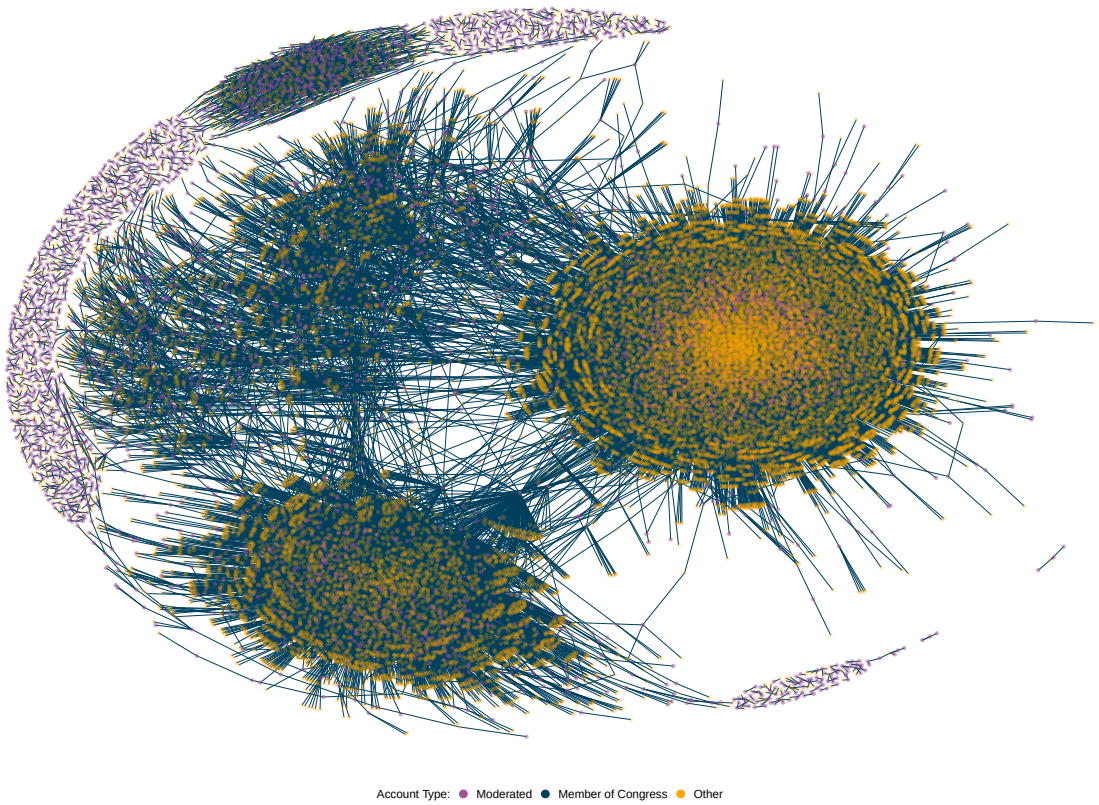

Figure 10: Full network of moderated accounts and their followers. Nodes are users and edges are following relations.

Our findings shown in Figure 10 indicate that accounts spreading harmful content had a large potential to spread information online, as they had extensive networks of followers. This suggests that these accounts had significant reach and influence within the Twitter community, which can exacerbate the impact of harmful content. The combination of coordinated behavior within clusters and the extensive reach of accounts spreading harmful content highlights the potential dangers associated with such accounts on social media platforms. The ability to manipulate and amplify information through coordinated efforts and the presence of deceptive identities can contribute to the spread of disinformation, harassment, and violence, which are all major concerns addressed in Twitter's moderation policies.

However, it is important to acknowledge the limitations of our analysis. As mentioned earlier, the dataset used in this study is likely selective and may not be fully representative of all moderation efforts by Twitter. Additionally, our analysis only covers data from two periods in 2022, and moderation behaviors may have changed over time. Furthermore, our focus on moderated accounts may not reflect the overall content production of all users on Twitter, and there might be other accounts engaging in harmful behavior that have not been moderated. Despite these limitations, our social network analysis provides

valuable insights into the behaviors of moderated users on Twitter. The identification of region-specific clusters and coordinated behavior within these clusters suggests the existence of harmful activities that violate Twitter's moderation policies. The extensive reach and influence of accounts spreading harmful content also underscore the potential impact of such behavior on the Twitter community. In conclusion, social network analysis is a powerful tool for understanding the dynamics of user behavior on social media platforms.

5 Conclusion & Avenues for Future Research

In this study, we relied on several computational tools of descriptive analytics, topic modeling, and social network analysis. Our primary objective was to delve into the alignment between Twitter's account suspension decisions and their officially reported content moderation practices. To achieve this, we leveraged a novel dataset made available in September 2022 by the 'Twitter Moderation Research Consortium' (TMRC).

In the realm of social media research, transparency in content deletion practices remains a challenge, with most platforms shrouding such actions in secrecy. The work of (Gillespie et al., 2023) underlines the growing concern around content moderation, emphasizing the need for more comprehensive studies beyond U.S.-based platforms and high-profile incidents. However, the TMRC dataset provided us with a unique opportunity to shed some first light on this intricate aspect of Twitter's operations. To the best of our knowledge, we are the first to undertake a study of this kind, and thus provide initial insights into the application of Twitter's reported content moderation practices. Nevertheless, we recognize the presence of several considerable limitations in the dataset as previously mentioned, which merit careful consideration and do not allow us to generalize our findings to all content moderation practices by Twitter, let alone other social media platforms. Still, by highlighting some opaque and unaccounted decisions leading to account suspensions, we aspired to contribute meaningfully to the ongoing discourse surrounding the impact of platform policies on user experiences.

Our study yielded several key insights. Most notably, in most cases, the reasons for suspending Twitter accounts align with Twitter's community guidelines. However, we found instances where these reasons remain opaque, underscoring the need for greater transparency in moderation decisions. Our analysis further revealed geographical disparities in the reasons for moderation, and through the use of Social Network Analysis (SNA), we discovered potential harmful behavior in the form of coordinated content dissemination and deceptive identities among moderated accounts.

Taken together, we find that in the majority of cases the reasons for the suspension of Twitter accounts are in line with Twitter's community guidelines. Nevertheless, there are instances in which these reasons indeed remain opaque, suggesting further need to understand how and why social media

platforms such as Twitter decide to suspend user accounts. Interestingly, we showed that the most common reasons for moderation differ between various regions across the globe. Finally, we used SNA to explore the networks of moderated accounts on Twitter and detected potential harmful behavior through coordinated dissemination of content and the presence of misleading identities. Our findings contribute to the understanding of the complexities surrounding content moderation practices and their implications for the safety and well-being of online communities. Moving forward, it is essential for social media platforms to continue investing in transparency and data sharing initiatives, like Twitter's Moderation Research Consortium, to enable further research and better inform efforts to address harmful behaviors on their platforms.

Acknowledgments

This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement nr. 883121).

References

Alizadeh, M., Gilardi, F., Hoes, E., Klüser, K. J., Kubli, M., & Marchal, N. (2022). Content moderation as a political issue: The twitter discourse around trump's ban. *Journal of Quantitative Description: Digital Media*, *2*.

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022). Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 258–266).

Barrett, B., & Kreiss, D. (2019). Platform transience: changes in facebook's policies, procedures, and affordances in global electoral politics. *Internet Policy Review*, *8*(4), 1–22.

Caplan, R., & Gillespie, T. (2020). Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media+ Society*, *6*(2), 2056305120936636.

Douek, E. (2021). Governing online speech: From" posts-as-trumps" to proportionality and probability. *Colum. L. Rev.*, *121*, 759.

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, *21*(11), 1129–1164.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros Fernandez, A., ... Myers West, S. (2023). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Gillespie, T. & Aufderheide, P. & Carmi, E. & Gerrard, Y. & Gorwa, R. & Matamoros-Fernández, A. & Roberts, ST & Sinnreich, A. & Myers West, S.(2020). Expanding the debate about content moderation: scholarly research agendas for the coming policy debates. Internet Policy Review*, *9*(4).

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hallinan, B., Scharlach, R., & Shifman, L. (2022). Beyond neutrality: Conceptualizing platform values. *Communication Theory*, *32*(2), 201–222.

Hovyadinov, S. (2019). Toward a more meaningful transparency: Examining twitter, google, and facebook's transparency reporting and removal practices in russia. *Google, and Facebook's Transparency Reporting and Removal Practices in Russia (November 30, 2019)*.

Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, *131*, 1598.

Klonick, K. (2020). The facebook oversight board: Creating an independent institution to adjudicate online free expression. *Yale Law Journal*, *129*(2418).

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Roberts, S. T. (2019). *Behind the screen*. Yale University Press.

Siapera, E., & Viejo-Otero, P. (2021). Governing hate: Facebook and digital racism. *Television & New Media*, *22*(2), 112–130.

Stewart, E. (2021). Detecting fake news: Two problems for content moderation. *Philosophy & technology*, *34*(4), 923–940.

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, 18.

Urman, A., & Makhortykh, M. (2023). How transparent are transparency reports? comparative analysis of transparency reporting across online platforms. *Telecommunications Policy*, 102477.

York, J. C., & Zuckerman, E. (2019). Moderating the public sphere. *Human rights in the age of platforms*, *137*, 143.

A Topic Modelling with BERTopic

A.1 Preprocessing and Parameter Configuration

Preprocessing involved constructing a TF-IDF matrix X with terms in less than 90% of the documents:

$$X_{i,j} = \text{tf}_{i,j} \times \log\left(\frac{N}{\text{df}_j}\right)$$

A.2 Model Customization and UMAP

We employed the xlm-t multilingual transformer model from Barbieri et al. (2022) to obtain embeddings, then used UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction proposed by McInnes et al. (2018). UMAP is a nonlinear dimensionality reduction technique that efficiently captures both the local and global structure of the data, making it highly effective for clustering in topic modeling.

Unlike some other dimension reduction techniques, UMAP does not solely focus on preserving global structures. It constructs a topological representation of the high-dimensional data structure and aims to maintain both the macroscopic relationships and the microscopic relationships between data points.

The UMAP method is guided by minimizing the cross-entropy loss between the high-dimensional data F and the low-dimensional representation Y:

$$L(F, Y) = \sum_{i \neq j} P_{i|j} \log\left(\frac{P_{i|j}}{Q_{i|j}}\right)$$

where $P_{i|j}$ represents pairwise similarities in the original space, capturing the probability that point j is selected as a neighbor of point i, and $Q_{i|j}$ represents pairwise similarities in the low-dimensional space.

Several key factors drove the choice of UMAP:

- **Preservation of Structure**: UMAP preserves both local and distant relationships, allowing for nuanced discovery of topics in the tweets, which may contain both clear clusters and subtle connections.
- **Computational Efficiency**: UMAP is known for its scalability and computational efficiency, making it suitable for handling large datasets, like our collection of tweets.
- **Flexibility**: The algorithm provides a balance between preserving local and global structures, making it adaptable to the unique characteristics of tweet data, where topics can be tightly or loosely related.

In the context of our analysis, the application of UMAP provided a robust low-dimensional representation

of the data, contributing to the effective detection and interpretation of the underlying topics, their relationships, and the specific content that led to user suspension.

B Network Analysis

B.1 B.1. Follower Network Construction

The follower network was modeled as a directed graph $G = (V, E)$, where V represents individual Twitter accounts and E represents the "follower" relationship. Edges were included only if occurring more than a threshold $T = 25$ times, resulting in the adjacency matrix A.

B.2 Fruchterman-Reingold Algorithm Implementation

The Fruchterman & Reingold (1991) algorithm is a force-directed method particularly suitable for complex social network visualization. By modeling nodes as repelling charges and edges as springs, the algorithm metaphorically captures the inherent structure of social relations.

The algorithm's energy function:

$$E = \sum_{i \neq j} \left(\frac{k^2}{d_{ij}} - A_{ij} \frac{d_{ij}^2}{k^2} \right)$$

is minimized to determine the layout of the graph, where d_{ij} represents the distance between nodes i and j, and k is a constant.

The natural metaphors used in the algorithm, such as charges and springs, align well with the relational dynamics within social structures. This allows the visualization to represent the closeness or distance between individuals (or Twitter accounts in our case) in an intuitive and meaningful way.

Moreover, the energy-minimizing property of the Fruchterman-Reingold algorithm ensures a stable and balanced layout that highlights the clustering and interconnectedness within the network. This makes it easier to identify community structures, influential nodes, and other key characteristics essential in understanding the complex patterns of follower relationships in Twitter's suspended user accounts (Fruchterman & Reingold, 1991).