Content Moderation As a Political Issue: The Twitter Discourse Around Trump's Ban*

Meysam Alizadeh[1], Fabrizio Gilardi[1], Emma Hoes[1], K. Jonathan Klüser[1], Mael Kubli[1], and Nahema Marchal[1]

[1]Department of Political Science, University of Zurich

December 16, 2021
Work in progress

Abstract

Content moderation—the regulation of the material that users create and disseminate online—is an important activity for all social media platforms. While routine, this practice raises significant questions linked to democratic accountability and civil liberties. Following the decision of many platforms to ban Donald J. Trump in the aftermath of the attack on the U.S. Capitol in January 2021, content moderation has increasingly become a politically contested issue. This paper studies that process with a focus on the public discourse on Twitter. The analysis includes over 3 million tweets and retweets posted by over 1 million users between January 2020 and April 2021. We find that while content moderation was already widely discussed in 2020, public interest in the issue really peaked in January 2021. Our analysis also shows the US Twitter discourse on this topic to be largely driven by non-elite users and polarized along ideological lines, with right-leaning and left-leaning actors emphasizing different dimensions of the issue. These findings highlight relevant elements of the ongoing process of political contestation surrounding this issue, and provide a descriptive foundation to understand the politics of content moderation.

*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). We thank Fabio Melliger and Paula Moser for excellent research assistance.

1 Introduction

On January 6, 2021 thousands of protesters stormed the Capitol Building in Washington, DC, in an attempt to contest the results of the 2020 US presidential election. Days after the event, Twitter took unprecedented steps to permanently ban the then president Donald J. Trump from the platform after he had voiced support for the protesters, citing a "risk of further incitement of violence" (Twitter, 2021). Shortly after that, other platforms including Facebook followed suit (Rosen and Bickert, 2021). This was widely seen as a pivotal moment in how social media companies enforce their own terms of services and rules in general, and how they treat politicians and other public figures in particular (Wiener, 2021). Facebook had previously shied away from banning President Trump's content, for example, citing public interest and newsworthiness, despite him repeatedly breaking the platforms' terms of services.

Platforms have long laid claim to neutrality when it comes to adjudicating the content circulating on their sites (Gillespie, 2018). Facebook CEO Mark Zuckerberg repeatedly said that the company should not be seen or indeed become an "arbiter of truth".[1] However, content moderation decisions are part and parcel of social media's business operations—and are far from neutral (Gillespie, 2020, 2018; Gerrard and Thornham, 2020). Over time, most social media platforms have had to develop and refine sets of rules and guidelines about what should and should not be allowed on their sites. These encompass everything from nudity to misleading information, harassment and hate speech, as well as plain old spam, and must balance the need for specificity with the fact that they are implemented in diverse markets. Importantly, making and implementing decisions on the kinds of content that is allowed online is not limited to large internet corporations; instead, the question emerges in virtually any context in which user-generated content is permitted. For example, a niche website popular among crafting enthusiasts, known as "the Facebook of knitting," found itself at the center of controversy after it decided to prohibit users from submitting content such as patterns to knit pro-Trump sweatshirts or

[1] Mark Zuckerberg, Facebook Status Update, Facebook (Nov. 18, 2016) (https://perma.cc/AQ2F-S6EM).

Confederate flags.[2]

While the term "content moderation" suggests a menial, relatively harmless activity, it is an increasingly contested practice, linked to fundamental political questions such as freedom of expression. Consequently, it has recently been subject to growing scrutiny by legislators and civil society groups, while Facebook set up a quasi-independent para-judicial institution to adjudicate its decisions (Klonick, 2020). The increased salience of content moderation on the political agenda led to calls for policy action, such as revisions of Section 230 of the United States Communications Decency Act, which allows social media companies to freely moderate user-generated content while protecting them from liability. This paper aims to understand the politics of content moderation by focusing on how the issue has been framed by different sets of actors over time. Problem definition is an essential component of any policy process, but it is particularly important in an area such as content moderation, in which no established consensus exists regarding the nature of the problem, nor the appropriateness of potential solutions. Furthermore, no such consensus is in sight, or is even possible, to a greater degree than in many other policy areas. As Douek (2021, 769) argues, "[o]nline speech governance is a wicked problem with unenviable and perhaps impossible trade-offs. There is no end-state of content moderation with stable rules or regulatory forms; it will always be a matter of contestation, iteration, and technological evolution."

Specifically, in this paper we trace this ongoing "issue definition" process by analyzing the discourse surrounding content moderation on Twitter between January 2020 and April 2021, a period leading up to and following Donald Trump's ban, which contributed decisively to politicizing the issue among the public in the US as well as internationally. We collected and classified about 3 million tweets making statements on online content moderation, posted by over 1 million unique users. Our analysis proceeds in four steps. First, we find that content moderation became a salient topic after Twitter first fact-checked Donald Trump's tweets in June 2020, and reached a peak in January 2021. Second, we show that the retweet network is structured in two big, separated clusters of

[2]The New Yorker, March 22, 2021 (https://www.newyorker.com/magazine/2021/03/29/how-politics-tested-ravelry-and-the-crafting-community).

left-leaning and right-leaning users. Third, our analysis reveals that tweets cluster within five categories, which were distributed to varying degrees across different kinds of users: there are more general discussions about deplatforming (1) and "censorship" by social media platforms (2). Moreover, users voice opinions about Twitter's decision to suspend accounts in general (3) and to ban Donald Trump in particular (4). Lastly, users debate the repeal or rescue of Section 230 (5). Fourth, user engagements varies across topics as well as, to a smaller extent, different kinds of users. These results highlight relevant aspects of the Twitter discourse on content moderation and provide a basis for a better understanding of the politics of content moderation.

2 Content Moderation as a Political Issue

Content moderation refers to the "organized practice of screening user-generated content" (Roberts, 2019, 12) to determine whether it violates a site or platform's terms of service, as well as the institutional mechanisms put in place to adjudicate and enforce these rules (Grimmelmann, 2015). Other definitions describe content moderation as a mean to "structure participation in an online community" (Grimmelmann, 2015, 47) by setting the standards and norms by which users must abide when interacting with each other. Content moderation is generally performed by a combination of trained human moderators and algorithmic systems designed to parse through large quantities of data to identify potentially problematic material (Gorwa et al., 2020). This process can conceptually be disentangled along two dimensions. Firstly, this review can take place at different stages of the content life cycle, meaning that posts are being screened either before ("ex-ante") or after ("ex-post") they have been posted online (Klonick, 2018). The second dimension differentiates between a primarily reactive or proactive approach to content moderation. For the most part, moderators review content only after it has been flagged and brought to their attention, either algorithmically or by users themselves (Klonick, 2018). Yet, mounting pressure from governments and public groups in recent years has nudged platforms to adopt a more proactive approach to the detection and removal of content that violates their own private rules or public laws (Keller et al., 2020; Gillespie,

2020).

How platforms set and enforce rules and standards about the use of their sites has become a major global regulatory and political issue. In Europe, France and Germany were among the first to introduce legislation[3] requiring platforms to identify and remove illegal hate speech within imparted time limits under threat of hefty pecuniary penalties—raising concerns that legislation extended beyond its original scope with potentially negative consequences (Human Rights Watch, 2018; Article 19, 2019). In recent years, Twitter and Facebook have seen a surge in direct government requests—especially from India, Turkey, and Pakistan—to take down content by journalists and publishers (Dang and Culliford, 2021), while other governments are considering banning platforms from moderating content altogether (Nicas, 2021). Despite these efforts to regulate companies' practices, however, there is no consensus on how content should be moderated. Instead, solutions are contested, in part because the issues at stake are themselves subject to political contestation.

On the surface, the lack of consensus regarding appropriate regulatory frameworks for content moderation is surprising. The issue can be perceived as largely technical, an engineering problem that can be solved using modern AI approaches, which is a framing that fits well with Silicon Valley's worldview. However, at a deeper level, content moderation is an issue that raises fundamental questions about democratic accountability and civil liberties. Who should wield the power to adjudicate between permissible and prohibited content? Is the ensuing amplification or suppression of voices online a much welcomed objective, or a deplorable side effect? Under what conditions are government oversight and regulation warranted—and is it even feasible? Specifically, some scholars and civil society groups have raised concerns about the opacity and lack of transparency surrounding content take-downs (Gillespie, 2018; Kaye, 2019; Suzor, 2019), and questioned whether platforms have too much power in devising the set of rules and standards that govern online speech. Others lament platforms' tendency to suppress speech beyond what is required by law at the risk of infringing on users' rights to expression (Keller,

[3]the Network Enforcement Act and "Avia" law

2018; York, 2021). A growing scholarship, finally, points out that platforms' governance regimes only reflect and promote a narrow set of values at the expense of others (Klonick, 2018; Hallinan et al., 2021; Leurs and Zimmer, 2017).

Such fundamental questions about content moderation cannot be resolved purely with technical tools, and different answers generally entail distinct policy responses. In other words, content moderation is a political issue, which has become increasingly contested. How political actors frame this issue, and how successful they are in doing so, matters. A key insight from the agenda setting literature is that not every problem automatically becomes *political*. In other words, as Baumgartner and Jones (2010, 27) put it, "social conditions do not automatically generate policy actions." To rise to the political agenda, a given issue must first be construed as politically salient and specific arguments put forward as to how and why it might warrant policy intervention. Therefore, how political actors frame content moderation may impact the kinds of solutions proposed. For example, if content moderation is primarily framed as a violation of free speech, policy-makers might be more hesitant to implement strict regulation on platforms' rules around hate speech, misinformation and sensitive content. If, in contrast, content moderation is understood as an indispensable tool to fight online harms and create a safe environment for users, regulators might instead be inclined to push platforms towards greater moderation, even at the risk of suppressing legitimate voices. These are just two broad ways in which content moderation can be framed, and many more ways may be thinkable as we are just starting to understand the politics of content moderation.

In a first step to understand how content moderation has become a politically contested issue, this paper studies the Twitter discourse around Trump's ban. Specifically, we consider the following elements. First, how did the salience of content moderation on Twitter change over time, and which kinds of users were most active on Twitter about this issue? Second, was the discourse polarized along the ideological lines common to U.S. politics? Third, in what context was content moderation discussed, i.e., in conjunction with which topics? Fourth, which kinds of actors, and which topics, received more user engagements? Taken together, these questions address relevant aspects of the politicization

of content moderation, as expressed in Twitter discourse.

3 Data and Methods

We collected tweets posted between January 2020 and April 2021 using Twitter's academic API and the R package RTwitterV2.[4] Our search query included several keywords related to content moderation, such as "deplatforming" and "social media ban" as well as stopwords such as "coronavirus" and "travel" to remove tweets that were not related to content moderation (full list in Appendix S1.1). We exclude all tweets that were not written in the English language. Using this procedure, our initial dataset consists of 6,476,231 tweets posted by 2,065,154 unique users. About one in five tweets in our dataset is an original tweet, while the rest are retweets. Only about one third of users in our dataset posted an original tweet, while the others only retweeted other users (see Appendix S1.2). Despite our stopwords, a significant number of tweets were not related to content moderation, and therefore had to be removed. To do so, we developed classifiers that achieved degrees of accuracy between 90% and 95%. The classification procedures are described in detail in Appendix S2. Our final corpus consists of 562,899 original tweets and 2,568,245 retweets by 1,177,186 unique users.

We construct the retweet network to identify various communities of users who posted about content moderation, and to examine the extent to which users were polarized. In the retweet network, we represent each Twitter account as a node and include a directed edge between nodes if at least one of them retweeted the other one. We only consider nodes whose degree is equal to or greater than 10. The resulting graph consists of 28,971 nodes and 184,967 edges.[5] We use the *Louvain* community detection algorithm to identify various communities of users in the retweet network. To characterize the detected communities, as a further step we analyze their top hashtags and n-grams, their most influential users (according to their centrality measures), and these users' bio descriptions. In line with previous research on US Twitter users, we expect to find clusters structured

[4] https://github.com/MaelKubli/RTwitterV2

[5] To determine the visual layout of this network, we used the Force Atlas 2 layout in Gephi (https://gephi.org/).

around the US political spectrum, i.e., one large left cluster pitted against a right one.

To identify themes in Twitter users' activity on content moderation, we use Latent Dirichlet Allocation (LDA) by estimating probabilistic models of the topic for each tweet (Blei et al., 2003). While noisy, these models allows us to identify key content moderation sub-topics that various users discussed over time. We train an LDA model on our entire sample of relevant tweets (excluding the retweets). Each individual tweet is considered a random mix over K topics, which are in turn represented as a distribution over one or two word phrases, commonly referred to as tokens. Standard LDA approaches begin by choosing the number of topics for a given corpus k. By qualitatively assessing the semantic coherence of topic models estimated for a range of 3 to 100 individual topics, we found that a model where K is set to 6 classifies tweets most sensibly (see Figure S1 in SI for the plot of computed coherence scores).[6] Thereafter, the authors and research assistants manually labelled each topic after having reviewed associated top terms and sample tweets. Finally, each tweet is assigned to its highest probability topic according to the LDA model.

4 The Twitter Discourse on Content Moderation

We present four sets of results on the Twitter discourse on content moderation. First, we show the distribution of tweets on content moderation over time, which reflects the salience of the issue on Twitter. Second, we demonstrate, based on the retweet network, that the discourse on content moderation is polarized along the left-right ideological dimension. Third, we identify six main topical clusters within the tweets, ranging from general discussions of deplatforming and platform "censorship" in general, to discussions of Twitter's decision to block users, of the repeal or rescue of Section 230 of the US Communications Decency Act, and of the seemingly "concerted" move to ban Donald Trump from social media. Fourth, we show how user engagements varied across ideological communities as well as the six topics.

[6] We used the *Gensim* library of Python for this task with its parameters set as chunksize = 10,000, passes = 50, iterations = 100, alpha = asymmetric, and the default values for the rest of parameters.

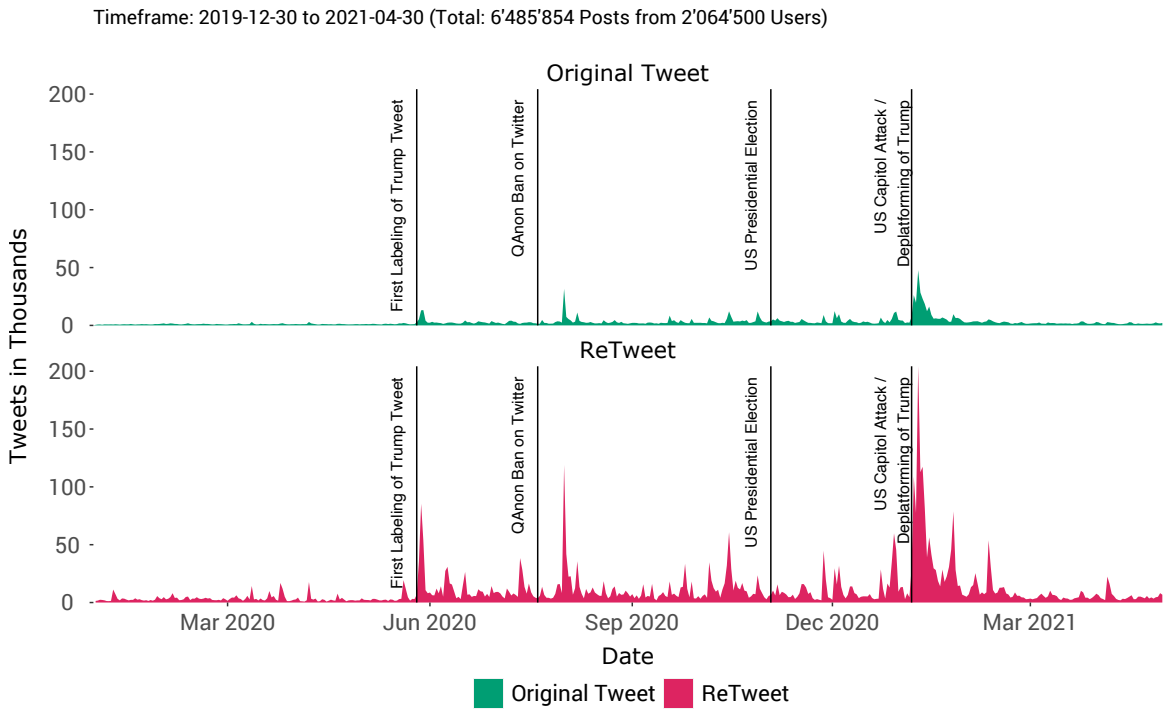


Figure 1: Frequency of original tweets and retweets on content moderation.

4.1 Salience of Content Moderation

Figure 1 shows the distribution of tweets on content moderation over time. The salience of content moderation was very low until Twitter started fact-checking Donald Trump's tweets in June 2020. After that, the issue only received a moderate amount of attention, as can be seen in the number of retweets. The peak was reached, unsurprisingly, when Donald Trump was initially banned from Twitter for incitement of violence in the context of the assault on the Capitol on January 6, 2021.

We highlight the role of different categories of users by displaying their tweeting activity over time in Figure 2. Specifically, we consider the following groups: members of the 116th[7] and 117th US Congress[8], news outlets[9] and political journalists[10], prominent think tanks[11], and individual experts[12] who are active in the area of content moderation. We

[7] https://www.sbh4all.org/wp-content/uploads/2019/04/116th-Congress-Twitter-Handles.pdf
[8] https://triagecancer.org/congressional-social-media
[9] See a list of 220 national news organizations in the appendix of Eady et al. (2020).
[10] See the appendix section of Alizadeh et al. (2020).
[11] https://guides.library.harvard.edu/hks/think_tank_search/US.
[12] See Table S9 in SI for a complete list of individuals.

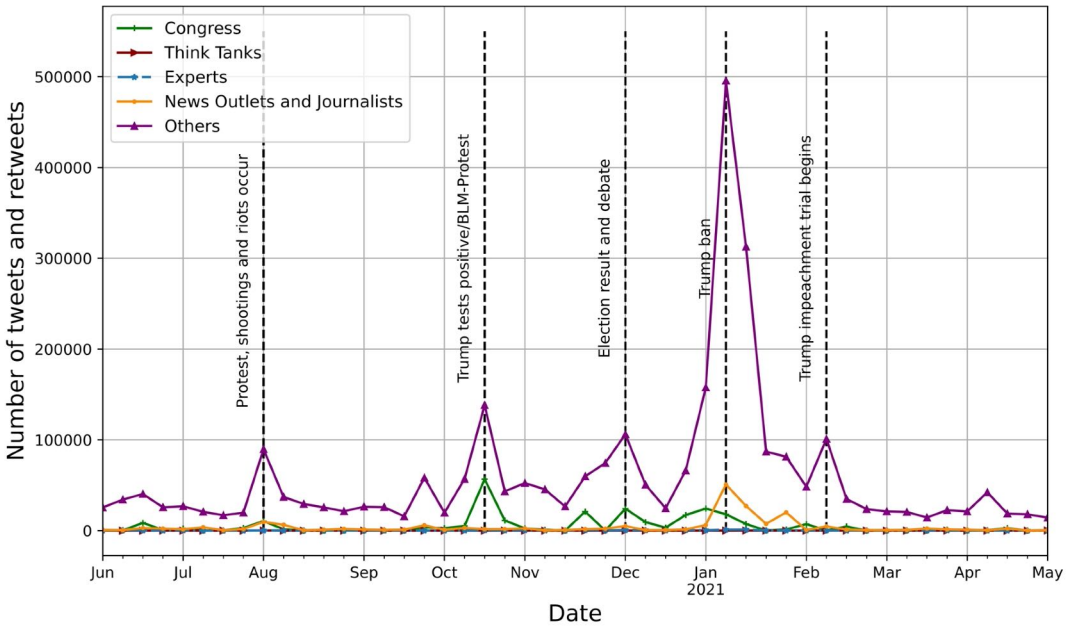
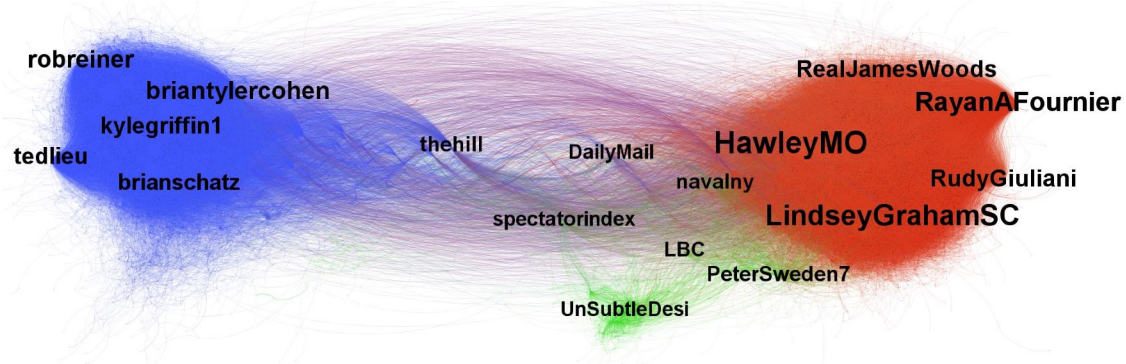


Figure 2: **Timeline of the number of tweets and retweets associated with political elites.** Plots show the number of tweets produced by each group, plus the number of times they were retweeted by any other account. The majority of the discussion were driven by non-elite users. Members of the Congress were more active and influential than news outlets and political journalists. Think thanks and individual experts had little effect.

label all other users as "Others", which include ordinary citizens, celebrities, influencers, organizations, and anyone else who posted on content moderation during our observation period.

Four observations stand out in Figure 2. First, non-elite users generate most of the political conversation about content moderation. Second, the number of tweets about content moderation increased significantly around major political events such as the Black Lives Matter protests, the US election results, and especially around Trump's ban from major social media platforms. Third, before January 2021, it was rather Members of Congress than news outlets and political journalists driving the online discussion, but their activity decreased after Trump's ban. Fourth, relevant think thanks (e.g., Data and Society Institute, Pew Research Center, and Brookings Institution) and individual experts (e.g., Daphne Keller, Joan Donovan, and David Hoffman) played a marginal role in the discourse, compared to other elite actors.

(a) Retweet Network of Users



(b) Timeline of the Activities of Users Across Communities

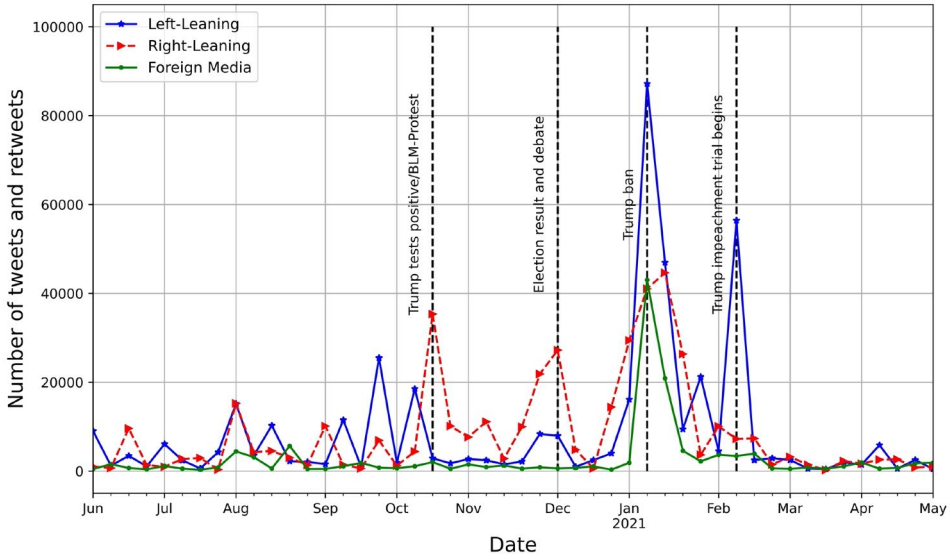


Figure 3: Identifying and characterizing various communities of users. (a) Retweet network of users who posted about content moderation. Each node represents a user, and there is an edge between two nodes if one of them retweeted the other. Node label size represents *PageRank* score. Color reflects different communities identified by the *Louvain* algorithm (Blue: liberal, Red: conservative, Green: Foreign media). Edges are colored by source. (b) Timeline of the activity of different communities.

4.2 Polarization of the Discourse

Figure 3a illustrates the retweet network of Twitter users who posted about content moderation and either retweeted or were retweeted at least 10 times during our data collection period. The network includes 28,971 nodes and 184,967 directed edges. Figure 3b plots the timeline of activity per cluster, where the vertical axis shows the number of tweets originated from users in each community plus the number of times they were retweeted by others.

Community	% Users	% Tweets	Top Hashtags	Label
Blue	48.4	45.2	DiaperDon, BigLie, Section230, JoeBiden	Left-leaning
Red	32.1	41.2	Section230, BigTech, antifa, Twitter, 2A	Right-leaning
Green	19.5	13.6	EXCLUSIVE, Trump, Deplatforming, Twitter	Foreign Media

Table 1: Characterizing Various Communities Within the Retweet Network

In Figure 3a, we can see that the majority of nodes are concentrated in two big, separated clusters representing left (blue) and right (red) leaning users. More particularly, almost 88% of users in the retweet network are clustered in the blue and red communities containing 55% and 33% of users respectively (Table 1). This suggests a polarized discourse in which users with similar political affinities retweet each other, but do not retweet out-groups. There is also one smaller community in Figure 3a which includes foreign (non-US) media accounts (green). This cluster contains 8% of users in the retweet network (Table 1).

The most frequently occurring hashtags found in the blue community's tweets are #DiaperDon (referring to unsubstantiated claims about Trump using adult diapers), #BigLie, #Section230, #transgender, and #JoeBiden. In addition, the most frequent hashtags found in the profile descriptions of users in the blue community are #BLM, #Resist, and #BlackLivesMatter. We also see that liberal political commentators and activists such as @briantylercohen, @robreiner, and @kylegriffin1 are among the top accounts with the highest *PageRank* (a measure of identifying influential nodes in a network) score in this community. As can be seen in Figure 3b, users within this cluster were highly active around the time of Trump's ban from social media and Trump's impeachment proceedings. This evidence leads us to conclude that the blue community leans to the left of the US political spectrum. In contrast, the most frequent hashtags found in the red cluster's tweets are #Section230, #BigTech, #antifa, #Twitter, and #2A (referring to the Second Amendment), while the most frequently occurring hashtags in their accounts description are #MAGA, #KAG (referring to Keep America Great, Trump's 2020 campaign slogan), and #2A. Famous conservative politicians and activists such as @HawleyMO, @RyanAFournier, and @RudyGiuliani are among the top users with the highest *PageRank* score in this community. Finally, we find that users in this community

were highly active when Trump tested positive for COVID19 and also around the time of the US election results (Figure 3b). Hence, we label this community as right-leaning on the U.S. political spectrum. The other smaller community is clearly associated with foreign media accounts. The most frequent hashtags in the green cluster's tweets were #EXCLUSIVE, #Trump, #Deplatforming, and #Twitter. For a complete list of the key identifying characteristics of the communities see Section S4 in SI.

4.3 Topics

Table 2 illustrates the six topics extracted by LDA along with the top 10 words for each topic, description of the topic, and our suggested labels for them. The first topic includes words such as "Trump", "president", "election", and "Biden", which we label as *Election* since most of the tweets are about claims and rumors related to the election results, election manipulation, and how Congress ought to prevent it. The second topic contains top words such as "ban", "Trump", "Twitter", and "account", which is clearly about social media platforms' ban of Donald Trump. Hence, we label it as *Trump Ban*. Furthermore, "social", "censorship", and "tech" are among the top 10 words of the third topic (*Censorship*), which implies discussion about speech censorship imposed by major tech platforms. The fourth topic includes top words such as "section", "repeal", "protect", and "amendment", which is labelled as *Section 230*. In case of the fifth topic, we see "deplatforming", "people", "speech", and "free" in the top words. This is suggestive of users' opinion about *Deplatforming* in its general meaning and not exclusively about Trump's deplatforming from major social media platforms (our review of 25 random tweets associated with this topic confirmed our label for this topic). Looking at the top words for the sixth topic and reading a sample of 25 random tweets associated with it, it became clear to us that this topic is about social media platforms' actions on content moderation, which is why we labelled it as *Platforms*. We provide three representative sample tweets for each topic in Table S5 in SI.

As can be seen in the last column of Table 2, most of the tweets were about "Trump

No.	Top 10 Words	Description	Label	% of Tweets
1	ban, trump, twitter, donald, account, tweet, facebook, president, permanent, shadow, lie	Platforms' Ban of Trump	Trump Ban	31%
2	trump, ban, president, election, bill, biden, news, democrat, american, republican, vote	Election Manipulation and Results	Election	29%
3	Section, repeal, tech, platform, company, big, protect, law, speech, sue, amendment	Section 230 Reform/Repeal	Section 230	18%
4	deplatforming, people, speech, free, work, make, conservative, hate, liberal, leave	Deplatforming of Special Users	Deplatforming	13%
5	Medium, social, censorship, sign, tech, big, petition, company, platform, corporate, conservative	Big Tech Speech Censorship	Censorship	5%
6	content, facebook, twitter, platform, moderation, post, harmful, hateful, youtube, ceo	Platforms' Actions on Content Moderation	Platforms	4%

Table 2: Extracted Topics from LDA and their Top Words and Suggested Labels.

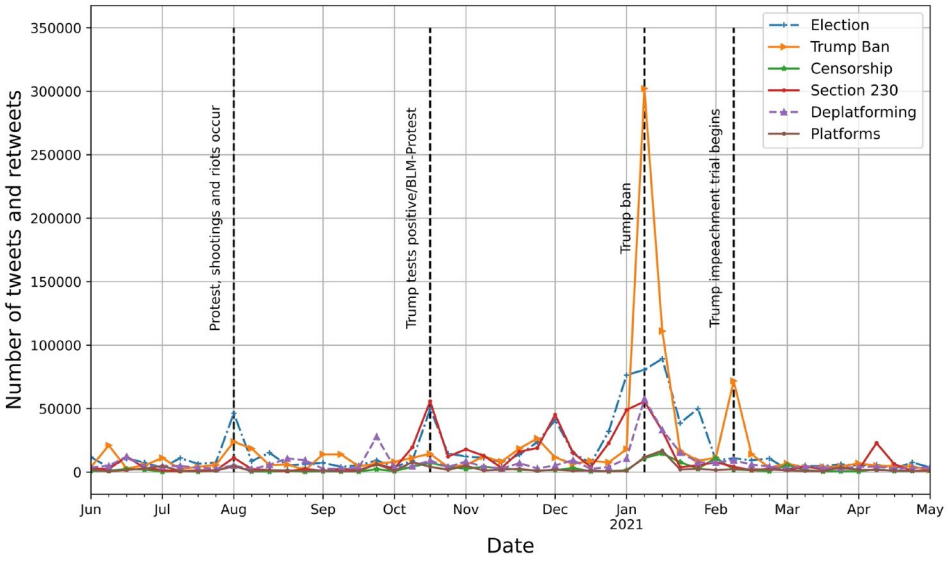


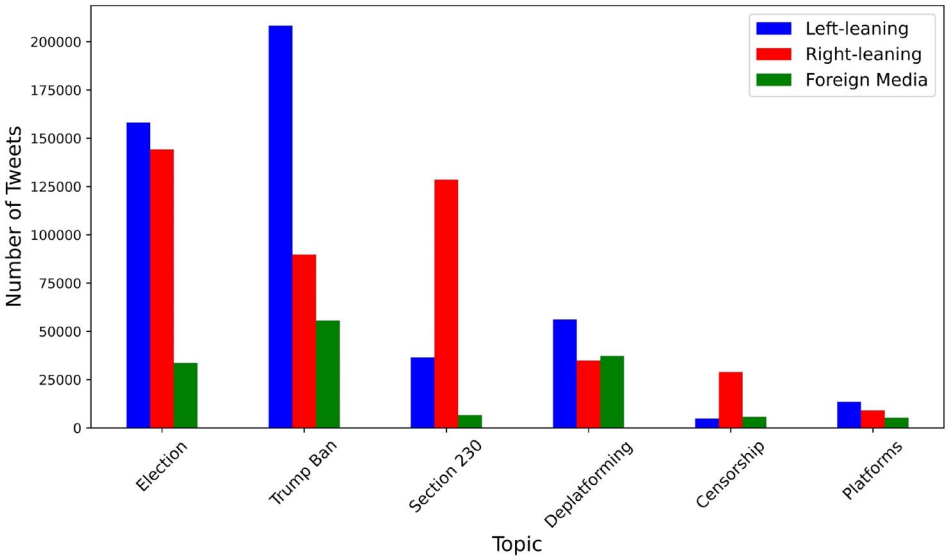
Figure 4: **Topic Trends Over Time.**

Ban" (31% of all relevant tweets and retweets) and "Election" (29%). Tweets about "Section 230" and "Deplatforming" made up 18% and 13% of all tweets respectively, whereas "Censorship" and "Platforms" comprise 9% of all tweets combined. However, this does not tell us much about the dynamics of topics over time. Therefore, we plot the weekly aggregated numbers of tweets and retweets associated with each topic in Figure 4. Unsurprisingly, "Trump Ban" peaked in January 2021, around the time he was effectively banned by most social media platforms. Moreover, we notice that "Section 230" has been a recurring topic that occasionally dominated the discourse.

To better understand the underlying characteristics of the political discourse on content moderation, we consider the distribution of topics across the three communities identified in our network analysis, and in particular left-leaning and right-leaning users. Figure 5a plots the distribution of topics across the three communities of users. We can see that while all three communities were posting about all six topics, the amount of activity varies. Users in the left- and right-leaning groups were almost equally posting about "Election", "Deplatforming", and "Platforms". However, whereas much of the discussion about "Trump Ban" was driven by left-leaning users, right-leaning users produced the majority of "Section 230" and "Censorship" posts. In fact, left-leaning users posted 2.5 times as many tweets as right-leaning users about "Trump Ban", and right-leaning users posted 3.7 and 3.5 times as many tweets as left-leaning users about "Section 230" and "Censorship" respectively. These differences highlight the contrasting ways in which users across the political spectrum frame the issue of content moderation. Foreign media accounts were mostly posting about "Trump ban", "Deplatforming", and "Election" and had less interest in the other three topics.

Furthermore, 5b displays the share of tweets across the six topics posted by the elite actors identified above (Congress members, Think tanks, Experts, and news media), that is, excluding our residual category ("other users"). The first observation is that across all topics, at most 27% of all tweets within a topic were written by members of Congress. For example, we can see that only 9% of the tweets about "Election" were posted by news outlets and political journalists. The second interesting observation is that members of

(a) Distribution of Topics Across Ideological Communities



(b) Distribution of Topics Across Types of Actors

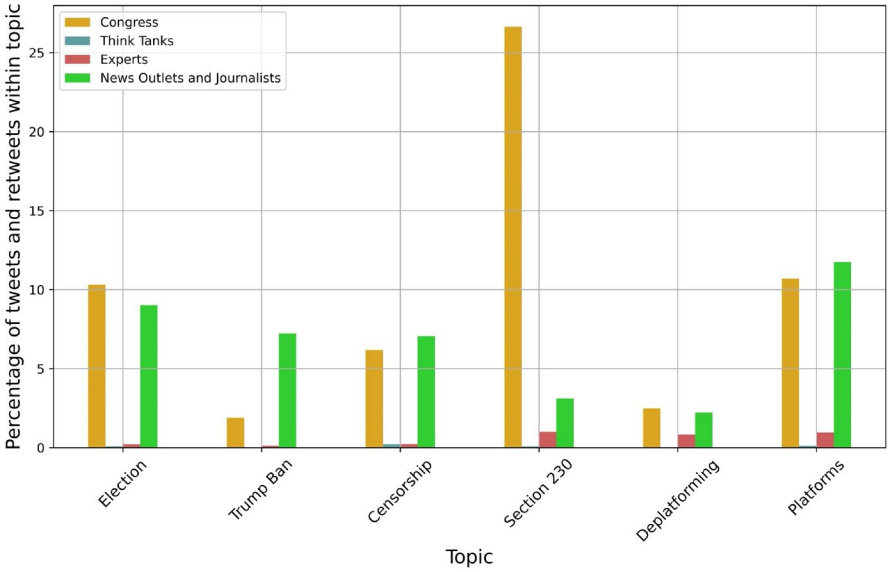


Figure 5: **Distribution of topics over political elites and communities.** (a) Both left- and right-leaning users posted almost equally about 'Election' and 'deplatforming'. However, while much of the discussion about 'Trump Ban' was driven by left-leaning users, the discussion around'Section 230' and 'Censorship' was mostly generated by right-leaning users. Foreign media outlets were mostly posting about "Trump ban", "deplatforming", and "election" and had less interest in the other topics. (b) At most, 27% of all tweets and retweets within a topic are attributed to a single political elite group. Members of Congress, news outlets and journalists produced the majority of the discussion across all topics. Members of the Congress were much more active on "Section 230". Individual experts were more active and successful in driving the conversation compared to think tanks.

Congress, news outlets and journalists had the biggest portion of discussion across all topics. The third observation is that members of Congress were much more active than any other group on the "Section 230" topic. Finally, individual experts were much more active and successful in pushing the discussion compared to think tanks.

4.4 User Engagements

We display density plots of observed user engagements (i.e., retweets and likes) across our three communities in Figure 6. The width of the curves represents the relative frequency of observations at any point on the vertical axis. We see that user engagement with posts created by the three communities do not considerably differ, but the left-leaning users have clearly attracted more user engagements than the other two communities. However, if we remove the likes and only consider the number of retweets, the retweet volume is markedly higher in the left-leaning community compared to the right-leaning and foreign media ones (see Figure S2 in SI).

We identify the top 10 accounts whose Twitter posts gained the highest engagement in each community (see Tables S10 and S11 for complete lists). In the left-leaning community, almost 13% of all retweets and likes (9.4% if only consider retweets) were attracted by a single user (i.e., @briantylercohen), and more than 42% of all retweets and likes were attributed to only 10 users (38% if only consider retweets). We see a similar pattern within the right-leaning community, with @LindsayGrahamSC inviting 14.5% of all retweets and likes (9.2% if only consider retweets) and 46% of all retweets and likes gained by only 10 users (45% if only consider retweets). The picture is more equal within the foreign media cluster, where the top users only account for 6% of all retweets and likes, with the top-10 users inviting 31%.

The density plots in Figure 7 show observed user engagements (likes and retweets) across our six LDA-detected topics. Considering the width of the curves, we see that user engagements of "Election", "Trump Ban" topics are approximately similar, and engagement volume of "Censorship" is almost equivalent to engagements of "Platform" topic. We also see a similar pattern between "Section 230" and "Deplatforming". Among

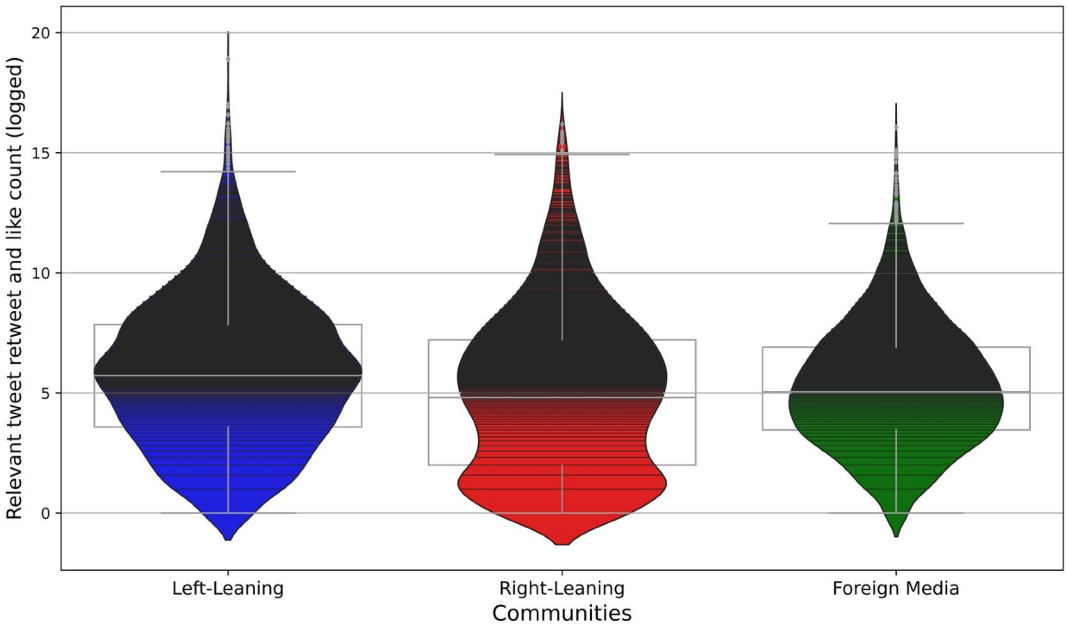


Figure 6: **User engagements across three communities.** Kernel density plots with overlaid boxplots showing median, first, and third quartiles of the distribution. The engagement volume do not considerably differ between the three communities, but the left-leaning users have higher user engagement compared to others.

the six topics, while "Election" and "Trump Ban" clearly have higher user engagements than others, the volume is slightly higher in "Trump Ban". Removing the number of likes and only plotting the retweets does not change the results.

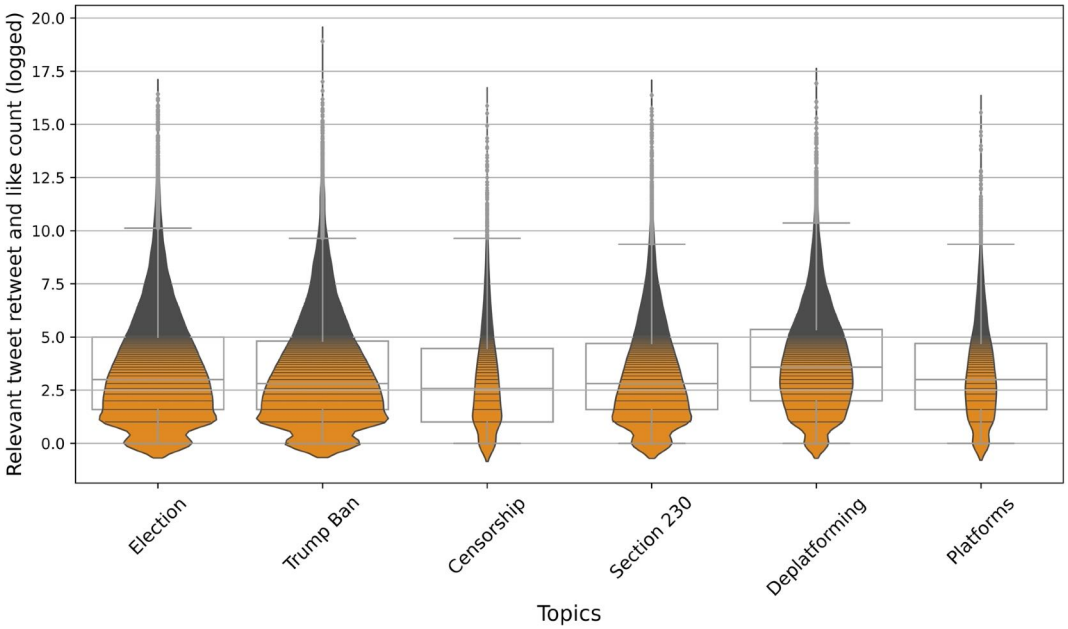


Figure 7: **User engagements across six topics.** Kernel density plots with overlaid boxplots showing median, first, and third quartiles of the distribution. Among the six topics, while "Election" and "Trump Ban" clearly have higher user engagements than others, the volume is slightly higher in "Trump Ban".

5 Conclusion

In this paper, we discussed how content moderation—the "organized practice of screening user-generated content" (Roberts, 2019, 12)—has become a salient political issue in the United States. Our initial analysis of millions of tweets posted between January 2020 and April 2021 shows that public attention to content moderation increased following social media companies' decisions to regulate the activity of prominent political figures. Moreover, the structure of the discourse network, inferred from retweets, reveals a familiar pattern of polarization within US political discourse, with right-leaning and left-leaning users discussing the issue mostly in isolation from each other. The US Twitter discourse on content moderation revolved either around general discussions of deplatforming, platform "censorship", and Section 230, or around discussions of specific events, such as Twitter's decision to block profiles, as well as the almost simultaneous ban of Donald Trump from major social media platforms. Importantly, different kinds of users engaged with these topics to varying degrees. Discussions of content moderation from the perspective of "Section 230" were particularly frequent among right-leaning users as well as members of Congress, whereas the "Trump ban" angle was over-represented among left-leaning users as well as foreign media. Finally, user engagement was higher around discussions of the US election and Trump's ban than any other topics. These result highlight relevant elements of the Twitter discourse around content moderation, which point to different ways in which the issue has been politicized in the United States. Although content moderation may be seen as a largely practical problem, it is in fact a process with deep political implications linked to the role of private and public actors in the regulation of speech. How the "problem" is defined matters for the kinds of "solutions" (policy actions) that are advanced to address it.

Our analysis provides an initial basis for a better understanding of the politics of content moderation, which could be extended in a number of ways. First, while the topic models are helpful to understand the general themes associated with content moderation, future research should aim to generate a more accurate conceptualization and measurement of the frames used to describe content moderation as a political problem. Second,

different kinds of arenas should be considered, including traditional media and statements made by politicians (for example in Congressional debates) as well as by tech companies. Third, the process by which content moderation has emerged as politically salient issues should be theorized more thoroughly, including the role of different kinds of actors and the connections between different arenas.

Content moderation is a widespread practice which has become increasingly contested due to its significant consequences for democratic accountability and civil liberties. By discussing how content moderation is a relevant political issue (as opposed to a purely technical one) and providing descriptive evidence on the social media discourse around it, we hope that this paper will stimulate further social science research on this important topic.

References

Alizadeh, M., J. N. Shapiro, C. Buntain, and J. A. Tucker (2020). Content-based features predict social media influence operations. *Science advances 6*(30), eabb5824.

Article 19, I. (2019, Jul). France: Analysis of draft hate speech bill.

Baumgartner, F. R. and B. D. Jones (2010). *Agendas and instability in American politics*. University of Chicago Press.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research 3*, 993–1022.

Dang, S. and E. Culliford (2021, Jul). Twitter sees jump in govt demands to remove content of reporters, news outlets.

Douek, E. (2021). Governing online speech: From 'posts-as-trumps' to proportionality and probability. *Columbia Law Review 121*(3), 759–833.

Eady, G., R. Bonneau, J. A. Tucker, and J. Nagler (2020). News sharing on social media: Mapping the ideology of news media content, citizens, and politicians.

Gerrard, Y. and H. Thornham (2020). Content moderation: Social media's sexist assemblages. *new media & society 22*(7), 1266–1286.

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

Gillespie, T. (2020). Content moderation, ai, and the question of scale. *Big Data & Society 7*(2), 2053951720943234.

Gorwa, R., R. Binns, and C. Katzenbach (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society 7*(1), 2053951719897945.

Grimmelmann, J. (2015). The virtues of moderation. *Yale JL & Tech. 17*, 42.

Hallinan, B., R. Scharlach, and L. Shifman (2021). Beyond neutrality: Conceptualizing platform values. *Communication Theory*.

Human Rights Watch, I. (2018, Feb). Germany: Flawed social media law.

Kaye, D. (2019). *Speech police: The global struggle to govern the Internet*. Columbia Global Reports.

Keller, D. (2018). Internet Platforms: Observations on Speech, Danger, and Money. (1807), 44.

Keller, D., P. Leerssen, et al. (2020). Facts and where to find them: Empirical research on internet platforms and content moderation. *Social Media and Democracy: The State of the Field, Prospects for Reform 220*.

Klonick, K. (2018). The New Governors: The people, rules and processes governing online speech. *Yale Law Journal 131*(6), 1599–1699.

Klonick, K. (2020). The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression. *Yale Law Journal 129*(8), 2418–2499.

Leurs, K. and M. Zimmer (2017). Platform values: an introduction to the# aoir16 special issue.

Nicas, J. (2021, Sep). Brazil's president bans social networks from removing some posts.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research 12*(Oct), 2825–2830.

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

Rosen, G. and M. Bickert (2021, Jan). Our response to the violence in washington.

Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge University Press.

Twitter, I. (2021, Jan). Permanent suspension of @realdonaldtrump.

Wiener, A. (2021, Jan). Trump's been unplugged. now what?

York, J. C. (2021). *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. Verso Books.

S1 Twitter data

S1.1 Keywords and stopwords for Twitter API query

Keywords

{(hateful OR harmful OR extremist OR conservative OR liberal OR objectionable) AND content}, {(trump OR qanon) AND ban}, #contentfiltering , #contentmoderation, #deplatforming, #onlinecensorship, #section230, #shadowbanning, blanket ban, content filtering , content moderation, content regulation, corporate censorship, deplatforming, disputed claims, left wing content, misleading information, moderation policies, moderation tools, new public square, no platforming, online censorship, right wing content, section 230, shadow banning, shadowbanning, social media ban, social media censorship, unverified claims.

Stopwords

#Covid19, #noflightlist, chloroquine, corona, coronavirus, covid, e-cigarette, flavor, flights, healthcare , immigrant, immigration, muslim, noflight, PGA, travel, vape, vaping, virus.

S1.2 Original tweets and retweets

Retweet	Total Tweets	Unique Users
FALSE	1'415'810	666'063
TRUE	5'070'044	1'608'418

Table S1: *Retweets vs original tweets*

S2 Classification

For coding purposes, we defined content moderation as follows: "Content moderation refers to the practice of screening and monitoring content posted by users on social media

Retweet	Min	Max	Median	Mean
FALSE	1	3'390	1	2.13
TRUE	1	1'146	1	3.15

Table S2: *User Statistics for retweets and original tweets*

	Total Tweets	RA 1	RA 2	Average
Test Questions Correctly Classified	365	75.5 %	80.2 %	77.9 %
Classified as relevant	4'716	35.0 %	40.3 %	37.7 %
Classified as relevant and agreed on	4'282			36.4 %

Table S3: *Training and Test Data Quality*

sites to determine if the content should be published or not, based on specific rules and guidelines." Using this definition, we instructed coders to tweets as relevant or irrelevant. Relevant tweets include those that discuss social media platforms content moderation rules and practices, and tweets that discuss governments regulation of online content moderation. It also includes tweets that discuss mild forms of content moderation, like flagging tweets and tweets when they directly relate to content moderation. Tweets should moreover be coded as irrelevant if they do not refer to content moderation or if they are themselves examples of moderated content. This includes for example tweets by Donald Trump that Twitter has labeled as disputed a tweet claiming that something is false or a tweet containing sensitive content.

To construct the training set, we initially relied on mTurk crowd-workers but then switched to research assistants to achieved the required quality. Our two research assistants classified 4,716 tweets with 365 test questions. Table S3 shows that the agreement with our test questions is fairly good with 77 % and the overall distribution of relevant and irrelevant tweets lies within our expectations. Furthermore the ICC between the two research assistants is very high with 89.1%.

To construct the classifier, we first applied standard text pre-processing steps: we converted all tweet texts to lowercase and remove all URLs, user mentions, and punctuations from the text (we crafted separate features for URLs and user mentions); we removed

	Precision	Recall	F1	Accuracy
Random Forest	0.94	0.92	0.93	0.93
XGBoost	0.95	0.94	0.94	0.95
Logistic Regression	0.90	0.88	0.89	0.90
Linear SVC	0.94	0.93	0.94	0.94

Table S4: *Classification Performance*

retweets; and we randomly selected 80% of tweets and train our classifiers on them and report the performance on the remaining unseen out-of-sample 20% of tweets.

We then calculated five types of human-interpretable features for each post-URL pair: (1) *content*: e.g., word count of a post and URL, polarity and readability of tweets, and *tf-idf* scores for those words that appear in at least 5 tweets; (2) *meta-content*: e.g., top 25 words or bigrams in relevant and irrelevant tweets; (3) *URL domain*: e.g., whether a URL domain is a news, political, satire, or national/local website; (4) *meta URL domain*: e.g., whether a URL domain is in the top 25 political, left, right, national, or local domains shared in relevant and irrelevant tweets; and (5) *communication*: e.g., whether a relevant think tank or congress member has been mentioned in a tweet. In total, we represent each tweet-URL pair as a vector of 3,761 features.

We trained *Random Forests*, *Logistic Regression*, *Linear Support Vector (Linear SVC)* and *XGBoost* classifiers using the *scikit-learn* library for Python Pedregosa et al. (2011). For the *random forests* classifier, we set the number of trees in the forest at 1000 (i.e. *n_estimators = 1000*), and the number of features to consider when looking for the best split as the square root of the number of potential features (i.e. *max_features = sqrt*). We use *scikit-learn*'s default setting for the rest of hyperparameters.

Table 6 reports macro-averaged precision, recall, accuracy, and F1 scores for each class (i.e. relevant or irrelevant to content moderation) using the default classification threshold of 0.5. F1 is the harmonic mean of precision and recall and is a standard metric for binary classification tasks. Precision is the fraction of true positives over the sum of true positives and false positives. Recall is the fraction of true positives over the sum of true positives and false negatives. Our main evaluation metric of interest to choose the

best classification model is the accuracy. We achieved the accuracy of 0.95 and 0.93 for the *XGBoost* and *Random Forest* classifiers on unseen out-of-sample test data respectively. Therefore, we choose *XGBoost* as the best model and use it to label the rest of the tweets and exclude the irrelevant ones.

S3 Topic Modeling Evaluation

Figure S1 plots the coherence scores for different number of topics ranging from 3 to 100. Since the coherence score is maximized at K = 6, we picked it as the optimum number of topics.

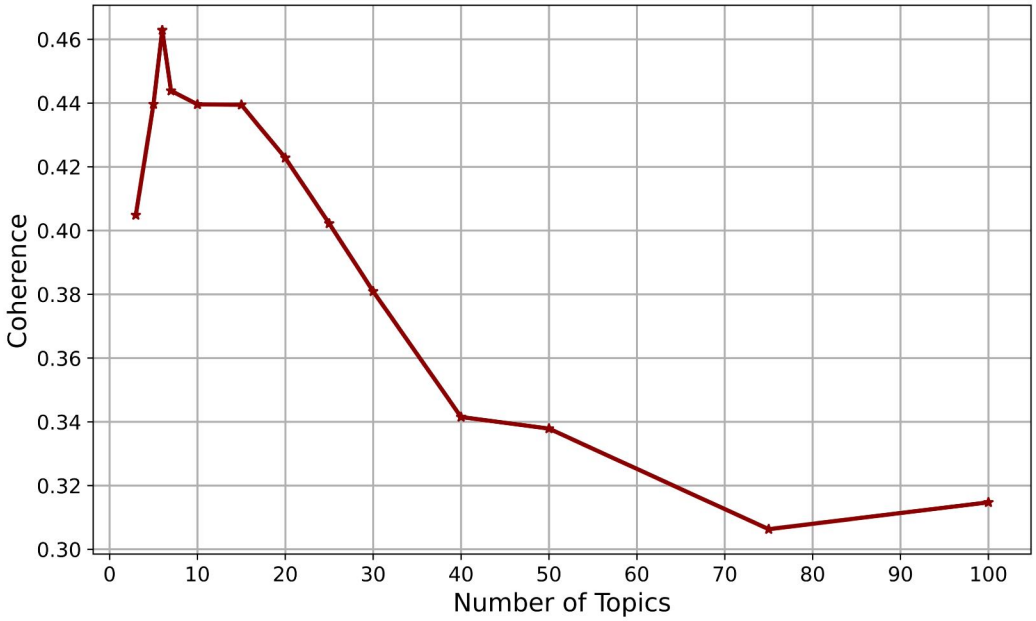


Figure S1: **Topic Coherence (Determining Optimal Number of Topics).** We used the *Gensim* library of Python for this task with its parameters set as chunksize = 10,000, passes = 50, iterations = 100, alpha = asymmetric, and the default values for the rest of parameters.

Topic	Sample Tweets
Election	- @AjitPaiFCC roll back section 230. It is obviously a tool to manipulate our election. - As Trump suggests issuing EOs to extend federal benefits, delay the election, and/or ban vote by mail, we should remind him that doing so would subvert the Constitution. - Joe Pollak live-tweets Trump campaign's 'Path to Victory' conference, still gets tagged by Twitter for spreading 'disputed' claims https://t.co/sISh0ADU0A @TwitchyTeam AAG AAG2020
Trump Ban	- BREAKING: Pressure mounts on social platforms to ban Trump for good. - I'm surprised Trump didn't do this from day one of his Twitter ban... https://t.co/jaROtE8p9y. - Facebook bans Trump 'indefinitely' https://t.co/Y6bZquWF6h via @Verge.
Censorship	- Fact Checking = CENSORSHIP Shadow Banning = CENSORSHIP Big Tech promotes... CENSORSHIP. - Jim Jordan accuses big tech of censoring, 'shadow banning' conservatives https://t.co/TABuUMyZmK via @YouTube. - Trump's Executive Order Aims to End Left-Wing Social Media Censorship of Conservatives https://t.co/k0omJZpaRl via @realmattcouch.
Section 230	- @realDonaldTrump @GStephanopoulos Repeal Section 230! - END SECTION 230 PROTECTION NOW!!! RT to all! https://t.co/VjDqwrcx9Z. - Revoking Section230 would change @Wikipedia as we know it. By completely destroying it and half the internet as well.https://t.co/eVu4tuzd9Z
Deplatforming	- Sometimes violence deplatforming is exactly the answer. Punch racists in the face and make the world a better place. https://t.co/nyQcCDECSN. - @mechanimyn @ggreenwald The people calling for the deplatforming are the fascists. https://t.co/roKXHbmrk6. - @RachelRileyRR @Facebook This whole deplatforming pple u dnt agree with is getting out of hand now.
Platforms	- The big, messy business of content moderation on Facebook, Twitter, YouTube https://t.co/8unkDja61i. - @tomphillipsin Facebook claims it blocked ResignModi "by mistake"?? Political content moderation is an unsustainable model. Social media will implode taking on content moderation for the whole world. - Facebook Finally Launches Its New Oversight Board for Content Moderation: https://t.co/fBvZLTDtup slashdot NBC News reports that "Social media users who believe their posts have been unfairly removed from Facebook or Instagram can now file an appeal to Facebook's Independent O...

Table S5: Sample Tweets for the Extracted Topics.

S4 Key Characteristics of the Communities

This section provides key features of the detected communities described in Section 4.2.

	Top Tweeter	Top Retweeted	Top Influential	Top Shared URLs	Top Hashtags	Top Mentioned	Top Words
1	tsartbot	briantylercohen	briantylercohen	hill.cm	#diaperdon	@briantylercohen	trump
2	GovSherazKhan	SachaBaronCohen	tedlieu	trib.al	#markzuckerberg	@sachabaroncohen	ban
3	Eilonwy24	sociauxling	Acosta	cnn.it	#biglie	@blackmarxist	twitter
4	MaryJan62216031	tedlieu	juliettekayyem	washingtonpost.com	#section230	@jack	president
5	bethlevin	HKrassenstein	kylegriffin1	cnn.com	#breaking	@realdonaldtrump	permanent
6	ewindham3	SenWarren	HKrassenstein	rawstory.com	#executiveorder	@twitter	run
7	givemepie360	Acosta	SachaBaronCohen	newsweek.com	#annomynous	@acosta	break
8	BeHappyandCivil	CNN	clairecmc	politico.com	#annonymous	@tedlieu	section
9	ScoobyLady27	juliettekayyem	CaslerNoel	variety.com	#weirdo	@hkrassenstein	people
10	Elizabe29599604	scottjshapiro	SenWarren	dlvr.it	#weirdotrump	@senwarren	violent
11	PamUnplugged	clairecmc	scottjshapiro	axios.com	#crybabytrump	@cnn	announce
12	SandyKAlano	NPR	AnaCabrera	politi.co	#dontaskdonttell	@juliettekayyem	deplatforming
13	DemocratsUp	CaslerNoel	robertjdenault	kateklonick.com	#consequencesfortrump	@susanwojcicki	fake
14	proudCanadavet	kylegriffin1	JoeBiden	scotsman.com	#donstake	@sundarpichai	medium
15	Tanis42	thehill	CNN	abc.net.au	#tiktokban	@scottjshapiro	facebook
16	ToniRagusa	JoeBiden	TheRickyDavila	thehill.com	#sikh	@clairecmc	social
17	TechmemeChatter	AnaCabrera	MSNBC	nbcnews.com	#trump	@npr	time
18	kat223	MSNBC	JuddLegum	thedailybeast.com	#parlerapp	@caslernoel	donald
19	bassher	Meidas_Michelle	thehill	abc7.la	#bears	@kylegriffin1	understand
20	roripierpont1	David_Leavitt	NormEisen	usatoday.com	#smartnews	@realdonaldtr	celebrity

Table S6: Characteristics of the Left-Leaning Community.

	Top Tweeter	Top Retweeted	Top Influential	Top Shared URLs	Top Hashtags	Top Mentioned	Top Words	Top Bigrams
1	real_Stephanie	LindseyGrahamSC	HawleyMO	trib.al	#section230	@realdonaldtrump	section	
2	ldga123456	HawleyMO	LindseyGrahamSC	hann.it	#bigtech	@lindseygrahamsc	ban	
3	PatriotMarie	RyanAFournier	RyanAFournier	childrenshealthdefense.org	#censorship	@hawleymo	trump	
4	MtRushmore2016	RudyGiuliani	RudyGiuliani	ow.ly	#thedefender	@ryanafournier	twitter	
5	Micropan44	GOPLeader	GOPLeader	foxnews.com	#instagram's	@rudygiuliani	president	
6	Samtaylorrose	RealJamesWoods	RealJamesWoods	buff.ly	#sect...	@gopleader	social	
7	StayTheCourse7	GayStr8Shooter	jsolomonReports	justthenews.com	#section230.	@realjameswoods	medium	
8	perchance99	laurenboebert	laurenboebert	justice.gov	#antifa	@gaystr8shooter	repeal	
9	21stcenturycrim	jsolomonReports	GayStr8Shooter	wsj.com	#section23...	@laurenboebert	tech	
10	mike71914654	Jim_Jordan	Jim_Jordan	dlvr.it	#2a...	@jsolomonreports	big	
11	Matthew52802818	MaryVought	RepTedBudd	reut.rs	#section230?	@twitter	censorship	
12	BeanK511	nypost	seanhannity	babylonbee.com	#walkaway	@jim_jordan	bill	
13	M_William_1985	RepTedBudd	MaryVought	truepundit.com	#nsl]	@maryvought	vote	
14	MyPlace4U	seanhannity	nypost	whitehouse.gov	#antielab	@nypost	biden	
15	rollypoly31	TheLeoTerrell	TheLeoTerrell	proofpoint.com	#antifadomesticterrorists	@reptedbudd	fracking	
16	Nmharley2	barronjohn1946	CharlesPHerring	thegreggjarrett.com	#new	@jack	time	
17	OldSalz	CharlesPHerring	SeanLangille	wnd.com	#liberalstudies	@seanhannity	speech	
18	lori_clydesdale	MrAndyNgo	RepGregSteube	peoplesgazette.com	#hongkong	@fdrlst	break	
19	edmgail1944	RepGregSteube	MrAndyNgo	appledaily.com	#foxnews	@theleoterrell	conservative	
20	Draggen75	CawthornforNC	CawthornforNC	hannity.com	#sotu	@barronjohn1946	account	

Table S7: Characteristics of the Right-Leaning Community.

	Top Tweeter	Top Retweeted	Top Influential	Top Shared URLs	Top Hashtags	Top Mentioned	Top Words	Top Bigrams
1	shadesmaclean	mtracey	mtracey	independent.co.uk	#exclusive:	@mtracey	ban	
2	virgiliocorrado	navalny	navalny	bbc.in	#toolkits,	@navalny	trump	
3	cgsheldon	authoramish	PeterSweden7	theguardian.com	#censorship	@authoramish	twitter	
4	vmrwanda	dhruv_rathee	officialmcafee	zerohedge.com	#farmerspro...	@dhruv_rathee	donald	
5	world_news_eng	PeterSweden7	zerohedge	tdrt.io	#twitter	@petersweden7	medium	
6	Varun8Vijay	spectatorindex	OpIndia_com	patreon.com	#trump	@amitmalviya	censorship	
7	meghnabasu2	OpIndia_com	DailyMail	trib.al	#kooapp...	@spectatorindex	social	
8	newscenterPHL1	amitmalviya	Independent	indy100.com	#dominiccummngs	@opindia_com	content	
9	PHLNewsInsider	TheBrando2	kittypurrzog	opindia.com	#dominiccummings	@thebrando2	speech	
10	TechmemeChatter	BefittingFacts	spectatorindex	tcrn.ch	#breaking	@befittingfacts	free	
11	SouravDindaBJP	NorbertElekes	techdirt	republicworld.com	#narendramodi	@norbertelekes	deplatforming	
12	BitesData	officialmcafee	NorbertElekes	theverge.com	#deplatforming	@twitter	president	
13	varun18vijay	piersmorgan	authoramish	sptnkne.ws	#tommyrobinson	@officialmcafee	platform	
14	VirtualPartyBla	Independent	amitmalviya	punchng.com	#left:	@piersmorgan	account	
15	dhruvbhim	KapilSibal	Snowden	engt.co	#europeanparliament	@independent	amp	
16	allymrowe	saltydkdan	MeghUpdates	ndtv.com	#kashmir	@kapilsibal	facebook	
17	jazmasigan_2	Snowden	SkyNews	dlvr.it	#facebook	@saltydkdan	extend	
18	PhilDeCarolis	zerohedge	coolfunnytshirt	rol.st	#section230	@youtube	act	
19	rmrbq	DailyMail	MichaelChongMP	ctvnews.ca	#update	@joerogan	thread	
20	InnovativeHindu	newslaundry	ellymelly	indiatimes.com	#gravitas	@snowden	unacceptable	

Table S8: Characteristics of the Foreign Media Community.

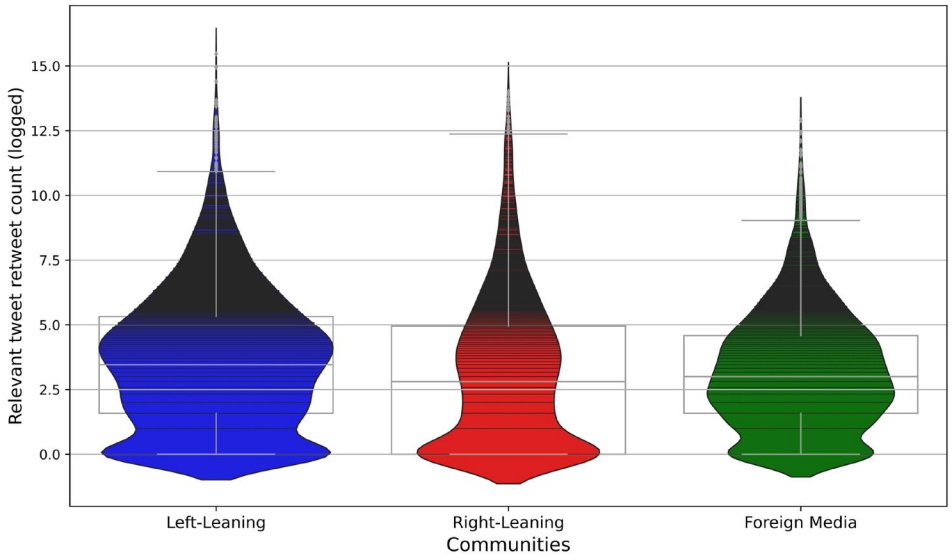


Figure S2: **Retweets distribution across three communities.** Kernel density plots with overlaid boxplots showing median, first, and third quartiles of the distribution.

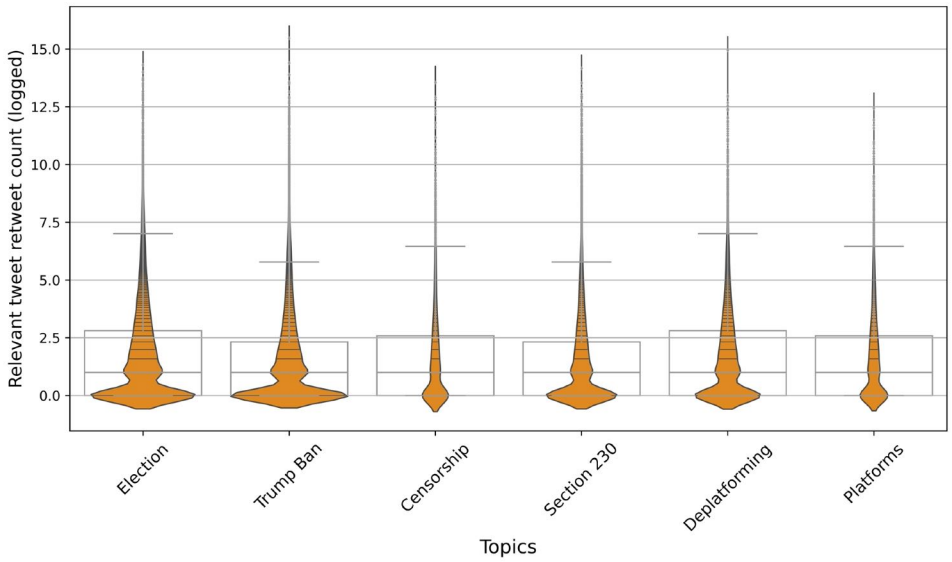


Figure S3: **Retweets distribution across six topics.** Kernel density plots with overlaid boxplots showing median, first, and third quartiles of the distribution.

Name	Twitter Account
Ajit Pai	@AjitPai
Jessica Rosenworcel	@JRosenworcel
Kelly Quinn	@_kquinn_
Sarah Lamdan	@greenarchives1
Johannes M. Bauer	@jm_bauer
Dr. Ruth, Ph.D.	@LCCWC_Ruth
Jasmine McNealy	@JasmineMcNealy
Jonathan Obar	@CDNJobar
Woodrow Hartzog	@hartzog
Peter Swire	@peterswire
Rich DeMillo	@richde
Shaheen Kanthawala	@ItsShaheenK
Azer Bestavros	@Bestavros
alessandro acquisti	@ssnstudy
Laura Brandimarte	@thefirstred
Philip Napoli	@pmnapoli
Robyn Caplan	@robyncaplan
Jacob Metcalf	@undersequoias
Elizabeth Anne Watkins	@watkins_welcome
Tarleton Gillespie	@TarletonG
danah boyd	@zephoria
Chris Bail	@chris_bail
Daniel Kreiss	@kreissdaniel
Shannon McGregor	@shannimcg
Robyn Caplan	@robyncaplan
Casey Fiesler	@cfiesler
Brianna Dym	@BriannaDym
Michael Zimmer	@michaelzimmer
Aaron Jiang	@aaroniidx
Hannah Bloch-Wehba	@HBWHBWHBW
Nathaniel Raymond	@nattyray11
Sofia Ranchordas	@SRanchordas
Elisabeth Sylvan	@lisard
Joan Donovan	@BostonJoan
Deen Freelon	@dfreelon
Alexandra Siegel	@aasiegel
Kelly Quinn	@kelly__quinn
JaneCombs	@jane_combs
Philip Napoli	@pmnapoli
Ken Rogerson	@rogerson
Seth C. Lewis	@SethCLewis
Sarah Stonbely	@SarahStonbely
Tom Glaisyer	@tglaisyer
colaresi	@colaresi
Jen Easterly	@CISAJen
Daphne Keller	@daphnehk
marius dragomir	@mariusdrag
David Hoffman	@hofftechpolicy
Jolynn Dellinger	@MindingPrivacy
Justin Sherman	@jshermcyber
Francesca Tripodi	@ftripodi

Table S9: List of the individual experts on the topic of content moderation and their Twitter handles.

Left-Leaning Screen Name	%	Right-Leaning Screen Name	%	Foreign Media Screen Name	%
briantylercohen	12.8	LindseyGrahamSC	14.5	Jack_Septic_Eye	6.1
rickygervais	7.3	HawleyMO	8.4	mtracey	5.1
SachaBaronCohen	5.9	RyanAFournier	4.9	saltydkdan	3.1
sociauxling	2.7	RudyGiuliani	4.7	navalny	3.1
CNN	2.6	GOPLeader	2.7	jazz_inmypants	2.7
BetteMidler	2.6	laurenboebert	2.6	dhruv_rathee	2.5
scottjshapiro	2.3	RealJamesWoods	2.3	authoramish	2.5
HKrassenstein	2.2	jsolomonReports	2.2	piersmorgan	2.3
kylegriffin1	1.9	barronjohn1946	1.9	NorbertElekes	1.9
SenWarren	1.9	GayStr8Shooter	1.8	ContraPoints	1.7

Table S10: Top 10 users in each community who gained highest retweets and likes.

Left-Leaning Screen Name	%	Right-Leaning Screen Name	%	Foreign Media Screen Name	%
briantylercohen	9.36	LindseyGrahamSC	9.29	mtracey	5.27
SachaBaronCohen	7.18	HawleyMO	9.09	navalny	4.20
sociauxling	4.77	RyanAFournier	5.54	authoramish	3.03
tedlieu	2.58	RudyGiuliani	4.69	dhruv_rathee	2.48
CNN	2.54	GOPLeader	3.48	OpIndia_com	2.36
SenWarren	2.53	RealJamesWoods	2.94	PeterSweden7	1.91
HKrassenstein	2.51	jsolomonReports	2.70	spectatorindex	1.84
Acosta	2.30	laurenboebert	2.60	Independent	1.78
juliettekayyem	2.18	GayStr8Shooter	2.53	BefittingFacts	1.70
scottjshapiro	2.07	Jim_Jordan	2.16	amitmalviya	1.69

Table S11: Top 10 users in each community who gained highest retweets.