A Tale of Two Subreddits: Measuring the Impacts of Quarantines on Political Engagement on Reddit

Qinlan Shen,[1] **Carolyn P. Rosé**[2]
[1] Machine Learning Research Group, Oracle Labs
[2] Language Technologies Institute, Carnegie Mellon University
qinlan.shen@oracle.com, cprose@cs.cmu.edu

Abstract

Concerns over the spread of abusive content in social media have led online platforms to explore large scale community-level forms of moderation. As an alternative to full bans, Reddit has implemented quarantines of subreddits, limiting access to controversial communities without explicit removal. Previous evaluations of the effectiveness of quarantines have focused on their impact on toxicity in affected communities. In this work, however, we focus on examining the impacts of quarantines of political subreddits on issues of political engagement in the broader Reddit community. We present a case study of two ideologically distinct political subreddits, r/The_Donald and r/ChapoTrapHouse, and examine the impact of quarantines on activity, visibility, and political discussion. We find that while quarantines had a homogenizing effect on participation in political subreddits, they had limited impact on the visibility of community-internal issues and political language, due to the entrenchment of subreddit norms.

Introduction

Content warning: *This paper discusses issues regarding the moderation of toxic content in online spaces and shows example content from quarantined subreddits. Reader discretion is advised.*

In response to the spread of abusive content in online spaces, in recent years, social media platforms have experimented with moderation strategies that operate over large communities. Reddit, for example, in 2015 announced the ban of five subcommunities or *subreddits* that participated in coordinated harassment efforts[1]. Targeted deplatforming of communities, however, remains controversial, especially in the space of political discussion, where removal-based moderation has been challenged as a form of ideological censorship (Jiang, Robertson, and Wilson 2020). As an alternative to bans, Reddit has proposed the use of *quarantines*, a softer form of community-level moderation where access to controversial subreddits is made more difficult but communities can continue to operate on the site. Due to the ongoing nature of participation in quarantined communities, the long-term impacts of quarantines and their efficacy as a moderation strategy have been heavily debated among moderation researchers and Redditors alike.

Prior work examining the impact of subreddit quarantines have suggested that they have limited effectiveness for mitigating toxic behavior on Reddit. Chandrasekharan et al. (2020) examined how the quarantines of TheRedPill and The_Donald affected participation and lexical usage, finding that while quarantines reduced user recruitment, they had limited impact on the use of misogynistic and racist terms within those subreddits. Cousineau (2020) and Ribeiro et al. (2020) argue that the soft moderation in quarantines serve as a warning and opportunity for communities to coordinate migration off of Reddit. However, work in this space has been limited to quarantines addressing abusive behavior by the Manosphere (Farrell et al. 2019) and the alt-right, whose audiences share similar values of patriarchy and cyberlibertarianism (Massanari 2020). Shen and Rosé (2019), who analyzed discussions around changes to Reddit's quarantine policy, found that redditors with varying political beliefs differed in how they discussed issues in the announcement. These differences were aligned with moral values favored by liberals and conservatives (Graham, Haidt, and Nosek 2009), suggesting that users with different beliefs highlight different priorities when considering moderation issues.

To address the question of how ideology can interact with responses to quarantining, we present a case study of the quarantines of two prominent political subreddits, r/The_Donald and r/ChapoTrapHouse. As The_Donald and ChapoTrapHouse fall on opposite sides of the left-right political spectrum, we can not only examine behavioral differences in how these subreddits respond to quarantines but also investigate the impact of quarantines on political engagement and polarization on Reddit. To examine the political impacts of quarantine in the broader Reddit space, we focus on three research questions, informed by debates and folk theories on quarantines:

- **RQ1:** What were the impacts of the quarantines on the posting activity of users in the quarantined subreddits? Is there evidence that quarantines have a homogenizing effect on participation within the quarantined subreddits?
- **RQ2:** How did quarantines impact the visibility and monitoring of issues within quarantined subreddits?
- **RQ3:** How did the quarantines impact the language of po-

[1] https://www.reddit.com/r/announcements/comments/39bpam/removing_harassing_subreddits/

litical discussion in and out of the quarantined subreddits?

RQ1 is similar to the causal inference analyses in Chandrasekharan et al. (2017) and (2020), which examined the impact of subreddit-level moderation events on posting activity and new user recruitment in the affected communities. However, we additionally analyze whether the impacts of the quarantines on posting activity changed for different types of users, such as power users or non-ideologically aligned users, to address the issue of whether quarantines have a homogenizing effect on participation in subreddits. In RQ2, we investigate the visibility and discussion of quarantined subreddits in three ideologically distinct subreddits focused on monitoring community issues on Reddit. For RQ3, in addition to measuring toxicity in subreddits over time, we evaluate whether quarantines had an impact on value associations highlighted in political discussion through the lens of Moral Foundations Theory (Graham et al. 2013). Through this analysis, we examine how users carry or maintain linguistic practices across different communities over time.

Background

Evaluating Content Moderation Impacts

Content moderation is considered vital for maintaining the health of online communities, enforcing what types of behaviors are or are not welcome in a community and establishing how anti-social behaviors should be treated (Gillespie 2018). Moderating content on large social media sites, however, is a fundamentally challenging problem, as platforms must balance open interaction with the regulation of acceptable and unacceptable behaviors at scale. Moderation research has explored the tensions and tradeoffs in content moderation strategies by considering the effects of interventions at different levels of impact, from individual users (Kiesler et al. 2012; Jhaver et al. 2019; Srinivasan et al. 2019) to subcommunities (Cousineau 2020) and entire platforms (Sybert 2021). The wide range of strategies, communities, and levels of impact, however, leaves open questions about best practices for moderation, both for specific communities and more general issues of content regulation.

Community-Level Content Moderation on Reddit

Participation on Reddit, a popular content aggregation and discussion platform, is centered around interest-based subcommunities called subreddits. Reddit emphasizes free speech and cyberlibertarianism as central tenets of the platform (Robertson 2015). As a result, subreddits are user-created and run with limited regulation by the Reddit administration. Nevertheless, Reddit has implemented community-level interventions, such as bans and quarantines, on subreddits with high amounts of objectionable content. Chandrasekharan et al. (2017) examined one such intervention in 2015, the removal of r/fatpeoplehate and r/CoonTown, finding evidence that these bans were effective in limiting the spread of hate on Reddit.

The impact of quarantines on offensive content, however, remains under heavy debate. Quarantining introduces design friction to community access by adding a warning to quarantined subreddits and preventing the subreddit from being searched. Quarantines, however, do not explicitly block users from participating in a controversial community. As such, the effectiveness of quarantines as a moderation strategy remains under debate. Some redditors argue that quarantines effectively limit the visibility and impact of objectionable content on Reddit without affecting user engagement[2]:

> *If you quarantine hateful content ... targets won't see it and feel bad; decent folks won't see it and will stay for other subreddits. And Reddit gets to keep the traffic.*

Others, however, argue that by not limiting participation, a high-profile quarantine may be counterproductive and drive up attention and traffic to a controversial community:

> *I lurk in T_D, and I don't know if this is confirmation bias but I've noticed more activity after the quarantine.*

> *We're doing great. The quarantine actually helped us through the Streisand Effect.*

Beyond their impact on user participation, quarantines may have effects on other aspects of content regulation on Reddit. Because quarantines allow subreddits to remain on the platform, controversial groups could be contained and monitored from known, centralized hubs:

> *It's the same reasons why white nationalist sites stay up. They're being monitored and the authorities can act when there's a real threat.*

However, some forms of monitoring, such as reporting by lurkers, could be hindered by visibility restrictions on quarantined subreddits. Redditors also speculate whether quarantines could concentrate objectionable behaviors in subreddits by insulating users in quarantined communities from outside views through an echo chamber effect:

> *Quarantining does a few things: it puts marginalized communities in danger due to the overwhelming encouragement of violence, it keeps people in a hole because their answers come from others in worse situations, and it perpetuates obsolete ideas.*

The relationship between quarantines and radicalization is especially important to understand for political subreddits, where engagement is centered around important social issues and discussions are influential in shaping political outcomes. Guided by these discussions over quarantines as a moderation strategy, in this paper, we explore three research questions focusing the effects of quarantines on participation, visibility, and polarized political discussion. We examine two ideologically distinct subreddits, r/The_Donald and r/ChapoTrapHouse, to investigate how these issues may affect political engagement on Reddit.

The_Donald

r/The_Donald was a subreddit centered around support of former president Donald Trump. Created in June 25, 2015, shortly after the announcement of Trump's candidacy for president, The_Donald has been widely studied as an influential hub for the far-right (Flores-Saviaga, Keegan, and Savage 2018; Massachs et al. 2020), with around 750,000

[2] Quotes from redditors are lightly paraphrased for anonymity

The_Donald	ChapoTrapHouse
The_Durham	EscobarOpiumDen
DonaldJTrumpFanClub	ChapoTrapHouse4
The_MuellerMeltdown	LessTankieChapo
DrainTheSwamp	BlackWolfFeed
TheRightBoycott	ChapoFYM

Table 1: Top 5 control (highest percentage of quarantined users) subreddits for The_Donald and ChapoTrapHouse

The_Donald	ChapoTrapHouse
WatchRedditDie	nfl
kotakuinaction2	CFB
gtaonline	PoliticalCompassMemes
StrangerThings	classic_wow
SubredditDrama	pan_media

Table 2: Top 5 invaded (highest increase in posting behavior from active users post-quarantine) subreddits for The_Donald and ChapoTrapHouse

subscribers at the time of its quarantine (Stewart 2019). Before its quarantine, r/The_Donald was a well-known source of controversial, hateful, and violent content, and Reddit had already implemented measures to prevent posts from the The_Donald from reaching the front page through collective voting (Robertson 2019). The subreddit was eventually quarantined on June 26, 2019, with repeated calls for violence against Oregon police and public officials during a climate change vote cited as the catalyst for the quarantine.

ChapoTrapHouse

r/ChapoTrapHouse is a subreddit centered around the popular left-wing comedy podcast Chapo Trap House, influential in the populist "dirtbag left" movement. The subreddit was reported to have around 130,000 subscribers around the time of its quarantine on August 6, 2019 (Martinez 2019), shortly after the quarantine of The_Donald. While there was speculation that the quarantine was due to *brigading* or targeted invasions of other subreddits and anti-cop sentiment, the reasons the subreddit was quarantined remain under debate[3]. As a prominent left-leaning quarantined community with controversy surrounding its quarantine, ChapoTrapHouse provides an interesting contrast to previously studied quarantined communities from the alt-right and Manosphere.

Data

Data for analysis was collected using full monthly dumps of Reddit activity ranging from May to September 2019 from the Reddit Pushshift API (Baumgartner et al. 2020). For our case studies, we focused on activity 50 days before and after the original quarantine date. We extract both submissions and comments from the quarantined subreddits under our observation period. In addition to the quarantined subreddits, we extract data from related control, invaded, and neighboring subreddits for analysis. In this section, we describe our procedure for finding these subreddits.

Control Subreddits

In this paper, we want to use causal inference techniques in order to establish whether quarantines had a direct impact on the outcomes in our research questions. To control for other factors that may influence outcomes, we want to find *control subreddits*, which are similar to the quarantined subreddit but were not quarantined, to serve as a quasi-experimental comparison. Following Chandrasekharan et al. (2017), we used co-posting behavior from users who actively posted[4] in the quarantined subreddit pre-quarantine to establish subreddit similarity. For control subreddits, we used the 100 subreddits with at least 50 users with the highest percentage of users who were also active users in the quarantined subreddits. Control subreddits were checked to ensure that none were quarantined or banned during the observation period. The top 5 control subreddits for The_Donald and ChapoTrapHouse are listed in Table 1.

Due to the highly interconnected nature of subreddits focused on a particular topic, such as political discussion, unlike in traditional A/B testing where control groups are not influenced by the intervention, we cannot fully guarantee that our selected control subreddits are not affected by the quarantines. For example, in response to the quarantines, users from the original quarantined subreddit may choose to move to a similar, but not quarantined controlled subreddit as an unmoderated alternative. While this is an inherent limitation with working with quasi-experimental techniques in an interconnected community, we argue that using related political subreddits as pseudo-controls, similar to Chandrasekharan et al. (2017), allows us to account for other underlying trends that may influence our dependent variables, such as political events or shifts in public opinion, in comparison to other unrelated subreddits. We primarily use these "control" subreddits as a basis of comparison between similar quarantined and non-quarantined subreddit to isolate the effect of directly experiencing the quarantine itself.

Invaded and Neighboring Subreddits

In addition to finding control subreddits for the causal analyses, we want to investigate the impact of the quarantines on other subreddits that may have had a change in participation. As in Chandrasekharan et al. (2017), we consider subreddits that had a 100% increase in posts by active users from the quarantined subreddits as *invaded subreddits*. Invaded subreddits ordered by increase in total posting behavior are listed in Table 2.

For both The_Donald and ChapoTrapHouse, we note that many of the top invaded subreddits include communities focused on interests outside of politics, such as gaming, sports,

[3] https://www.reddit.com/r/SubredditDrama/comments/cmw7o4/rchapotraphouse_has_been_quarantined_discuss_this/

[4] We define active users as users who have posted at least 10 comments in a subreddit. All user-level analyses in the paper are run on active users to limit the impact of drive-by participation.

The_Donald	ChapoTrapHouse
unpopularopinion	chapotraphouse2
Conservative	BreadTube
PoliticalHumor	LateStageCapitalism
conspiracy	COMPLETEANARCHY
AskThe_Donald	ENLIGHTENEDCENTRISM

Table 3: Top 5 neighboring (highest co-posting percentage) subreddits for The_Donald and ChapoTrapHouse.

	The_Donald	ChapoTrapHouse
# posts	2,584,025	1,124,617
# users	24,194	12,527
- power users	1,708	1,235
- non-power users	15,114	9,005
- aligned users	22,843	12,213
- non-aligned users	1,351	314

Table 4: Number of posts and users collected for The_Donald and ChapoTrapHouse

and television. As such, users may engage in drastically different behaviors in invaded subreddits compared to their original behavior in the quarantined subreddits. Because we are interested in the impact of quarantines on political discussion, we also want to analyze explicitly political subreddits with high sustained participation by users from quarantined subreddits. These *neighboring subreddits* were defined as subreddits where a high percentage of active users in the quarantined subreddit also participate. We first consider all subreddits that at least 1% of active users from each quarantined subreddit posted in, which is the 99th percentile for amount of user overlap between the quarantined subreddit and a candidate neighboring subreddit. We then manually filtered these candidate neighboring subreddits for subreddits focused on political and social issues. Due to the high popularity of the neighboring subreddits compared to invaded and control subreddits, in terms of subscribers and activity levels, we take only the top 25 neighboring subreddits in terms of percen for analysis. Examples of neighboring subreddits for The_Donald and ChapoTrapHouse are listed in Table 3.

Approximate Labeling of User Ideology

Users who participate in a political subreddits may not necessarily be aligned with the beliefs and norms of that community (Datta and Adar 2019; Guimaraes et al. 2019). To better understand the impact of quarantines on political engagement and highlight potential differences in reaction to the quarantine in left-leaning and right-leaning spaces on Reddit, we want to identify the political beliefs of users who participate in The_Donald and ChapoTrapHouse.

To estimate user-level beliefs, we follow previous work leveraging participation across subreddits as a proxy for user interests or ideology (Olson and Neal 2015). We label all subreddits in the monthly dumps as left, right, or neutral based on user co-posting behavior with known ideological subreddits. For each subreddit, we calculate the z-score of the log odds ratio of a user being active in both that subreddit and ChapoTrapHouse (left) vs. The_Donald (right). A subreddit is considered “left” or “right” if the z-score passes a one-tailed Z test at $p = 0.05$ in the corresponding direction. Otherwise, it is assigned the “neutral” label.

Users are then labeled as “left”, “right”, or “neutral” based on their distribution of participation in left and right subreddits. Users who post more often on left subreddits than right will be considered left and vice versa, with ties being neutral users. While all posts by a user could be used to construct the distribution for this assignment, a user’s participation within a subreddit may not be aligned with the underlying beliefs or norms of the community. A user may engage in antagonistic behavior in a subreddit and sustained antagonistic behavior may lead to a user to be labeled with their opposing ideology. To account for potentially antagonistic behavior, we only consider the posts of a user in a subreddit that have a karma score of at least 3 for this assignment.

RQ1: Posting Activity

For our first research question, we are interested in examining whether the quarantines of The_Donald and ChapoTrapHouse had an impact on activity within the quarantined subreddits. For our activity measures, we look at the total volume of posts (submissions and comments) and new users (users who have never participated in the subreddit, with a 10 day buffer before our observation period to account for pre-existing users) that the quarantined subreddit received over time. In addition to examining the impact of quarantines on overall activity, we consider the breakdown of these activity measures for different types of users. In particular, we investigate whether the quarantines of The_Donald or ChapoTrapHouse may have had an isolating or homogenizing effect based on potentially disparate effects for power users or ideologically aligned/unaligned users. We consider the following user categories in our activity analysis:

- **Power users:** Users in the 90th percentile for number of posts in the quarantined subreddit before the quarantine.
- **Non-power users:** Users who participated in the quarantined subreddit pre-quarantine but are below the 90th percentile for posting activity.
- **Aligned users:** Users who participate in the quarantined subreddit and have the same ideological alignment (i.e. “right” for The_Donald, “left” for ChapoTrapHouse.)
- **Non-aligned users:** Users who participate in the quarantined subreddit and have a different ideological alignment than the subreddit (including “neutral” users).

By definition, power users and non-power users are already members of the community before the quarantine. Thus, these categories are not used for the new users analysis. Statistics for these users are located in Table 4

Interrupted Time Series Analysis

To assess whether there is evidence of quarantines having an impact on the level of activity within quarantined subred-

	Posting Activity				New Users			
	β_{td}	p	β_{cth}	p	β_{td}	p	β_{cth}	p
overall	0.082	0.778	-0.703	<0.001***	-0.207	0.043*	-0.233	0.046*
power users	-0.378	0.041*	-0.567	<0.001***	-	-	-	-
non-power users	0.063	0.815	-0.826	<0.001***	-	-	-	-
aligned users	0.168	0.569	-0.680	0.002**	-0.090	0.423	-0.111	0.346
non-aligned users	-0.719	<0.001***	-1.061	<0.001***	-0.657	<0.001***	-0.870	<0.001***

Table 5: Interrupted time series coefficients for posting activity and new users in The_Donald and ChapoTrapHouse across user types. β_s is the level change coefficient for the dependent variable for subreddit s. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

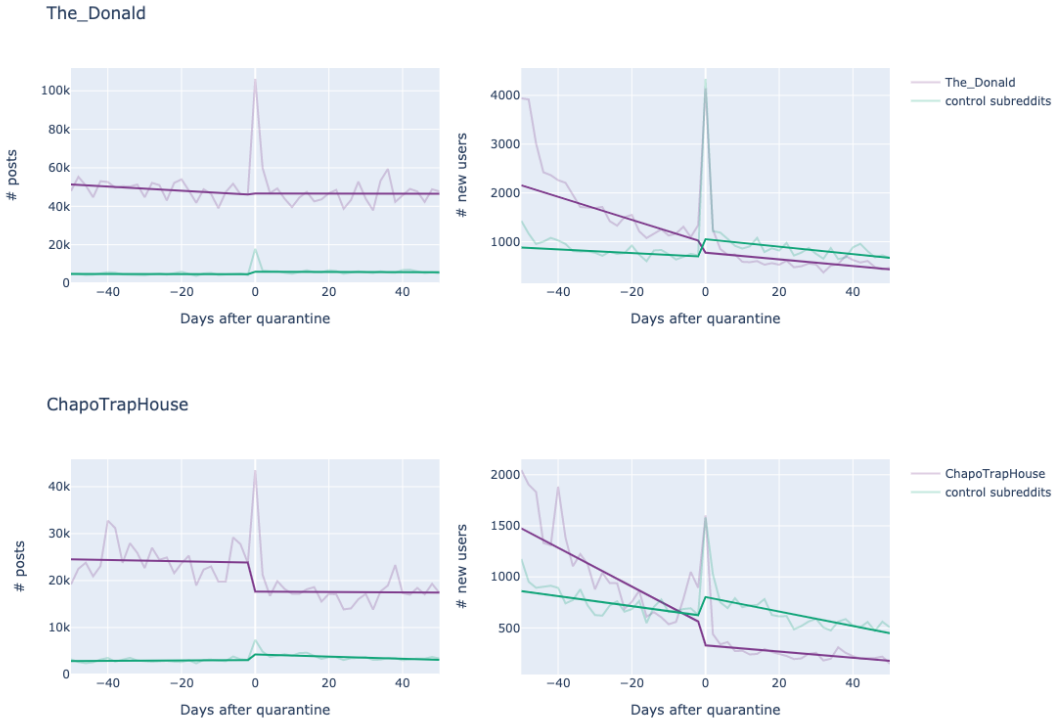

Figure 1: Total number of posts and new users over time in The_Donald and ChapoTrapHouse compared with aggregated control subreddits with fitted ITS regression.

dits, we use Interrupted Time Series (ITS) analyses (Bernal, Cummins, and Gasparrini 2017). In ITS analysis, a time series is used to establish an underlying trend for a dependent variable Y. Given an intervention at a known timepoint, the underlying trend is used as a counterfactual against a model that accounts for changes after the intervention to determine whether there is sufficient evidence that the intervention impacted Y. To account for changes in both the level and slope of the dependent variable after our intervention, a quarantine, we fit the following regression model

$$Y_t = \beta_0 + \beta_1 t + \beta_2 X_t + \beta_3 t X_t \quad (1)$$

where X_t is a binary indicator of whether timepoint t takes place after the intervention. While the slope change coefficient β_3 was included in our regression, for all our models, we found that β_3 either did not show a significant change or indicated a leveling off trend,[5] with the total slope going to zero after the quarantine. This means that while quarantines may have had a sustained impact on some of our dependent variables (i.e. β_2 being significant), the impact on the natural upward or downward trend of these variables is limited. Thus, we only report results on the level coefficient β_2 (hereinafter referred to as β or β_s for subreddit s for ITS analyses throughout the paper).

[5] The leveling off trend was only find in the analyses of new user recruitment. This is likely due to left censoring in how new users are calculated, which we partly address with a 10-day buffer.

Results

Table 5 shows the results for the level coefficient β from the ITS analysis, while Figure 1 illustrates the overall trends for the activity measures. Trend lines are calculated based on the Equation 1, with the central spike on the time bucket of the quarantine date removed from the regression in this and following analyses as an outlier. As in Chandrasekharan et al. (2020), we found that for The_Donald, while there was no significant change in the level of posting activity ($\beta_{td} = 0.082, p = 0.778$), there was a decrease ($\beta_{td} = -0.207, p = 0.043$) in the influx of new users after the quarantine. Using a one-tailed bootstrapping test over β_s for control subreddits s to determine whether the change β_{td} resembles that of changes in the control subreddits, we found evidence that the pattern of the decrease in number of new users was more extreme ($p < 0.001$) than that of the control subreddits. This suggests that the decrease in number of new users can be attributed to the quarantine, rather than general trends seen in other political subreddits. In ChapoTrapHouse, on the other hand, our results showed that there were drops in both overall number of posts ($\beta_{cth} = -0.734, p < 0.001$) and new users ($\beta_{cth} = -0.233, p = 0.046$) that were more dramatic than those in the control subreddits at $p < 0.001$. For both The_Donald ($\beta_{c(td)} = 1.297, p < 0.001$) and ChapoTrapHouse ($\beta_{c(cth)} = 0.741, p < 0.001$), we see a significant aggregate increase in the number of new users in the control subreddits after the quarantine, suggesting that users from the quarantined subreddits began exploring alternatives to their original quarantined subreddit. However, individual control subreddits varied in whether they had an increase or decrease in user recruitment after the quarantine of their corresponding original subreddit.

Breaking the analysis down by user type, however, suggests that many of the observed decreases may be attributed to certain groups. For The_Donald, there was a significant decrease in posting activity by power users ($\beta_{td} = -0.378, p = 0.041$) that was significantly different from the control subreddits ($p < 0.001$) but not for non-power users ($\beta_{td} = 0.063, p = 0.815$). This suggests that the most active users in The_Donald may have been more strongly impacted by the quarantine, while most users' overall level of activity remained stable. There was also a tendency in both The_Donald and ChapoTrapHouse towards decreased participation by users not ideologically aligned with the subreddit. In The_Donald, there was a significant decrease both in the number of posts ($\beta_{td} = -0.719, p < 0.001$) and recruitment ($\beta_{td} = -0.657, p < 0.001$) for non-aligned users but not for aligned users. We see a similar phenomenon in ChapoTrapHouse where there is a substantial drop in new non-aligned users ($\beta_{cth} = -0.870, p < 0.001$) but not aligned users. All decreases in posting activity and new users from non-aligned users were significantly different from the control subreddits ($p < 0.001$), providing evidence that the quarantines affected the activity of non-aligned users. These patterns suggest that quarantines had a homogenizing effect at the community level for both subreddits.

While we observe some evidence of homogenization at the community level, individual users may still interact with non-aligned content at similar rates before and after the quarantine. Prior research into the social dynamics of online political discussion (Kelly, Fisher, and Smith 2005; Morales, Monti, and Starnini 2021) have shown that cross-ideology interactions are more prevalent than expected compared to user demographics in mixed ideology settings. Thus, we also used ITS analysis to determine whether quarantines had an impact on the amount of cross-ideology interaction experienced by an average user in the subreddit. We define two forms of cross-ideology interaction: *direct* interaction, where a user replies or is replied to by a user with a different ideology label, and *indirect* interaction, where a user participates in the same comment thread (starting one level below a submission) as a user with a different label. Table 6 contains the ITS analysis results for cross-ideology interaction. For both types of interaction and both subreddits, we find a significant decrease in the level of cross-ideology interaction experienced by an average user in the subreddit.

Quarantined Sub	**Interaction**	β_{cross}	p
The_Donald	direct	-0.708	<0.001***
	indirect	-0.903	<0.001***
ChapoTrapHouse	direct	-1.104	<0.001***
	indirect	-0.468	0.025*

Table 6: Interrupted time series coefficients for percentage of cross-ideology interactions per user in The_Donald and ChapoTrapHouse. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Overall, our analyses of activity suggest that quarantines decreased the participation of users not ideologically aligned with the subreddit. This supports the hypothesis that quarantines have a homogenizing effect on political subreddits.

RQ2: Visibility and Monitoring

One argument raised in debates about quarantining was that quarantines allow controversial subreddits to be monitored from a known, centralized space. A competing concern, however, is that the design friction introduced by quarantines could lead subreddits to be isolated from outsiders, who may act as a moderating force as both participants and observers. While RQ1 investigated the isolating impact of quarantines on participants, in RQ2, we examine whether the quarantines impacted outsiders documenting issues in quarantined subreddits. We analyze submissions in 3 monitoring subreddits to examine whether quarantines shifted how much outside attention a community receives.

Monitoring Subreddits

For this analysis, we focus on three subreddits whose primary goal is to document issues of controversy, toxicity, and censorship occurring in other subreddits:

- **SubredditDrama:** "a place where people can come and talk about reddit fights and other dramatic happenings", r/SubredditDrama focuses on summarizing controversial events in and across different subreddits.
- **WatchRedditDie:** Described as "a place to track Reddit's abandonment of free speech and decline into cen-

Monitoring Sub	Quarantined Sub	β_{mon}	p
SubredditDrama	The_Donald	0.644	0.122
	ChapoTrapHouse	0.044	0.915
WatchRedditDie	The_Donald	1.000	0.002**
	ChapoTrapHouse	-0.328	0.288
AHS	The_Donald	0.312	0.456
	ChapoTrapHouse	-0.453	0.190

Table 7: Interrupted time series coefficients for number of monitoring submissions mentioning the quarantined subreddit. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

sorship", r/WatchRedditDie collects examples of removed threads and comments across Reddit to argue that Reddit has abandoned its founding free speech principles.

- **AgainstHateSubreddits:** r/AgainstHateSubreddits describes its primary goal as "[drawing] attention to reddit's contributions to the growing problem of radicalization" The subreddit links to examples of toxic content that are held up or amplified by the subreddits they take place in.

All three monitoring subreddits rely on submissions to document individual incidents aligned with its goals. Thus, for RQ2, we focused on submissions as the unit for analysis.

One notable aspect of the monitoring subreddits is that all three have an ideological lean from our subreddit labeling process (**Labeling user ideology**) – SubredditDrama and AgainstHateSubreddits were labeled "left" and WatchRedditDie was labeled "right". We do not argue that these monitoring subreddits are unbiased in operation. In fact, the ideological biases of these subreddits are reflected in their goals, as discussion focused on censorship in right-leaning WatchRedditDie and harms and consistency in the left-leaning subreddits mirrors the findings from Shen and Rosé (2019). We are instead interested in seeing whether the political leanings of these subreddits led to differences in how they monitored issues for our quarantined subreddits.

Results

To measure the impact of the quarantines on how monitoring subreddits documented incidents in quarantined subreddits, we ran ITS analyses on submissions that each subreddit received mentioning a quarantined subreddit (Table 7). Across the three subreddits, we only found a significant change in WatchRedditDie, where there was a significant increase in submissions mentioning The_Donald after the quarantine ($\beta_{mon} = 1.000, p = 0.002$). Overall, this suggests that quarantines did not reduce the attention that quarantined subreddits received from monitoring subreddits. One explanation for this is that both The_Donald and ChapoTrapHouse were high-profile subreddits before their quarantines. Users in the monitoring subreddits were likely already aware of the issues and reputations of those subreddits and thus, were able to maintain close attention to those communities.

In WatchRedditDie, there was a substantial increase in submissions centered around The_Donald after its quarantine. To better understand what drove this increase, we use a Sparse Additive Generative Model (Eisenstein, Ahmed, and Xing 2011), or SAGE, to compare distinctive terms in WatchRedditDie submissions before and after the quarantine (Table 8). The key intuition behind SAGE is that by modeling the difference in word frequencies compared to a background corpus, it can enforce sparsity in topics or class labels over words. In addition to examining the top SAGE terms by themselves, we also examined example submissions that contained the top terms to provide additional context for trends in WatchRedditDie. We see some evidence that there was a shift in focus for submissions about The_Donald, with terms before the quarantine referring to properties of specific posts or threads, such as "comments" and "link", and terms after focusing more on quarantines and site-wide administration issues. This suggests that discussion around The_Donald on WatchRedditDie shifted from specific examples of content removal in the subreddit to issues surrounding its quarantine.

Quarantined Sub	Time	Top Terms
The_Donald	after	quarantined, users, moderator, reddit, site admin, political, ban censorship, power
	before	report, search, removed, comments, click, link discussion, title, threads, drama
ChapoTrapHouse	after	gone🦀, brigade, alt, quarantine, mentality, racist, evasion, good, 🤡🤡🤡, propaganda
	before	chapo, subs, death, user, white, reddit, The_Donald, nazi, site, banned, hate

Table 8: Distinctive terms in WatchRedditDie before/after the quarantine of The_Donald obtained using SAGE.

We, however, do not see a similar shift in number of submissions or language for ChapoTrapHouse. Instead, many of the terms associated with ChapoTrapHouse submissions after the quarantine on WatchRedditDie are associated with celebration (e.g. the dancing crab emoji), mockery (e.g. clown emojis), or behaviors justifying the quarantine. This shift in behavior suggests that unlike for The_Donald, where the quarantine was used to challenge the Reddit administration moderation strategies, the quarantine of ChapoTrapHouse was seen as justified by WatchRedditDie. For the right-leaning user base of WatchRedditDie, the quarantine of The_Donald may be considered more salient or personal than issues from the quarantine of ChapoTrapHouse. This difference in reaction suggests that, to some degree, monitoring in WatchRedditDie is motivated by personal ideology.

Differences between β_{mon} for The_Donald and ChapoTrapHouse for the other monitoring subreddits suggest that there may be similar ideological effects in those subreddits. However, we observed that the major shifts in top SAGE terms in the other two monitoring subreddits are primarily centered on how the quarantined subreddits discussed specific political events, such as Charlottesville, Israel, Andy Ngo, etc., rather than broader moderation issues on Reddit. Nevertheless, we found some submissions high in ChapoTrapHouse post-quarantine terms suggesting that some users in AgainstHateSubreddits question or oppose the quarantine:

> *AgainstHatesubreddits can't stand that CTH has been quarantined. Quite ironic, isn't it?* (SubredditDrama)

> *I was trying to find out what happened to Chapo. Given that the sub isn't bad, I would've liked to find out about it under other circumstances.* (AgainstHateSubreddits)

Overall, these results suggest that quarantines had a limited impact on visibility, as many quarantined subreddits tend to be high-profile communities Reddit is wary of removing. We, however, see some evidence of ideological motivation in what is discussed about quarantined communities.

RQ3: Linguistic Analysis

The primary goal of quarantines is to limit the spread of objectionable content from controversial communities. Therefore, in order to evaluate their effectiveness, we need to examine their impact on the toxic content produced by the quarantined subreddit. In the space of political discussion, however, quarantines may have linguistic impacts beyond amounts of toxic content. By limiting outside intervention, for example, political ideas could be reinforced or radicalized in more isolated communities. Thus, in RQ3, we examine the impact of quarantines on an additional linguistic phenomena, the association of political issues with moral values. We examine the impact of quarantines on these linguistic features for quarantined subreddits and related communities that may be affected by quarantine events.

Measuring Toxicity

To estimate the amount of toxic content produced by a subreddit, we use Perspective API[6], a popular machine learning system from Google for evaluating the impacts of texts. While Perspective API suffers from limitations, such as lacking consideration of more subjective forms of toxicity, its general purpose nature provides some advantages for us. Because the reasons behind ChapoTrapHouse's quarantine remain unclear, Perspective API can provide grounding for toxic behaviors without having to specify known targets of hate. For the quarantined subreddits, as well as our control, invaded, and neighboring subreddit sets, we collect toxicity scores for 1,000 posts per day in our observation period.

To validate whether Perspective API provides a reasonably good estimate of toxicity in political subreddits, we sample 100 posts each from The_Donald and ChapoTrapHouse. One of the authors than manually identified whether the sampled posts contained texts that intentionally disparages an individual or group on the basis of some identity characteristic, such as race, gender, nationality, sexual orientation, occupation, etc. (Schmidt and Wiegand 2017). We then compare the toxicity scores of Perspective API, thresholding at a score of 0.5 for toxic content. We found that Perspective API achieved an F1 score of 68.18 (precision=75.0, recall=62.5) on The_Donald and 70.59 (precision=64.29,recall=78.26) on ChapoTrapHouse. While the precision for Perspective API was lower than the lexicon-based approaches used in Chandrasekharan et al. (2017) and Chandrasekharan et al. (2020), it was able to achieve reasonably good recall on both quarantined subreddits, which allows us to more effectively measure the overall prevalence of toxicity in these subreddits beyond strictly defined keywords.

Moral Foundations

When expressing stances on issues, individuals draw associations between political subjects and moral values to justify their beliefs. Certain values may be highlighted to strengthen a stance or argument for a particular audience. Moral Foundations Theory (Graham et al. 2013) provides a framework for describing basic moral values held across human cultures and has commonly been used in political studies to describe differences in associations drawn by liberals and conservatives (Graham, Haidt, and Nosek 2009). As political subreddits, the primary goals of The_Donald and ChapoTrapHouse are to discuss issues and express support for a particular political objective. Thus, we propose using moral foundations to measure changes in political associations and discussions in response to the quarantines. We use the expanded moral foundations set that includes liberty as a core moral value:

- **Care/Harm:** care for others, sensitivity to suffering
- **Fairness/Cheating:** fairness, justice, and reciprocity
- **Loyalty/Betrayal:** solidarity with one's in-group
- **Authority/Subversion:** respect for hierarchy, tradition
- **Sanctity/Degradation:** avoidance of the impure/taboo
- **Liberty/Oppression:** desire not to be restricted

Table 9 shows examples of these moral foundations being invoked in political discussion on Reddit. Prior work in NLP on moral foundations has primarily focused on texts by formal political entities, such as news sources or politicians (Johnson and Goldwasser 2018). To account for domain differences with the more informal political discussions on Reddit, we annotate our own moral foundations dataset. Two annotators labeled 50 comments for whether each comment invoked each moral foundation. Inter-annotator agreement for the moral foundations categories were calculated using Cohen's κ (Table 10). Overall, annotators were able to obtain moderate agreement over all moral foundation categories, except Sanctity/Degradation. Deliberation between the annotators revealed that the annotators had disagreements over what was considered taboo in a political context (e.g. sex, drug use, Communism in the U.S.). After discussing these boundary cases, the two annotators then separately annotated 2,200 comments.

[6] https://www.perspectiveapi.com/

	Example Post
Care/Harm	It marked the first time Trump ever gave anything to charity instead of stealing from charities.
Fairness/Cheating	The only reason California went to Hillary was because of cheating like illegals voting and Google/Facebook/Twitter interfering in the election.
Loyalty/Betrayal	I see no conceivable reason to support Weld at any point, including the fact that he's not a conservative in any sense.
Authority/Subversion	Ok, I'm really fed up with Harris bullying for extra time, and the moderators giving it to her, every time.
Sanctity/Degradation	The DNC are sickening little parasites, fucking vermin.
Liberty/Oppresion	Are we allowed to make fun of Obama's hurricane or is that also violent hate speech that will get us banned?

Table 9: Examples of Reddit comments labeled as invoking a particular moral foundation.

Moral foundation	κ_H	κ_L	κ_{DB}
Care/Harm	68.03	23.46	56.60
Fairness/Cheat.	67.65	11.72	49.14
Loyalty/Betray.	49.32	27.06	58.87
Authority/Subv.	63.77	2.75	42.73
Sanctity/Degrad.	18.48	5.47	64.68
Liberty/Oppress.	63.41	10.18	47.38

Table 10: Cohen's κ agreement results for moral foundation annotation by humans (κ_H), the expanded moral foundations lexicon (κ_L), and a fine-tuned Distilbert model (κ_{DB}).

Feature	β_{td}	p	β_{cth}	p
Toxicity	-0.552	0.100	-0.062	0.883
Care/Harm	-0.702	0.100	-0.019	0.963
Fairness/Cheat.	-0.455	0.255	-0.563	0.165
Loyalty/Betray.	0.697	0.085	-0.266	0.521
Authority/Subv.	-0.209	0.610	0.146	0.737
Sanctity/Degrad.	-0.467	0.080	0.187	0.649
Liberty/Oppress.	0.219	0.612	0.461	0.293

Table 11: Interrupted time series coefficients for the value of the linguistic feature in The_Donald and ChapoTrapHouse.

We use 2,000 of these comments as a training/seed set and 100 comments each as validation and test sets for evaluating approaches for labeling our full quarantine dataset. We consider two approaches for propagating our annotated labels:

- **Lexicon:** For our lexicon-based approach, we extend the original moral foundations dictionary from Graham, Haidt, and Nosek (2009) to account for more Reddit-specific invocations of moral foundations using pointwise mutual information (Church and Hanks 1990). Using both the original dictionary and our set of annotated posts, we calculate the PMI between every word in our corpus and posts containing a specific moral foundation F. The 100 words with the highest PMI for each moral foundation F that were not included in the original dictionary are then added as additional indicators for foundation F. We consider a post to contain a moral foundation if it has at least one occurrence of a term for that foundation in the extended dictionary.
- **DistilBERT:** We fine-tune a DistilBERT (Sanh et al. 2019) pretrained language model to perform the moral foundations classification task. We first fine-tune the base language model on a sample corpus of r/politics from May to August 2019, using the masked language model objective. We then train the fine-tuned model as a moral foundations classifier on our training/seed set for 10 epochs.

Table 10 compares the κ of our two approaches on our test set. We find that the trained DistilBERT model performs adequately and significantly better than the lexicon-based approach for all of our moral foundation categories. Thus, we use our trained classifier to label the remainder of our data.

Results

For both toxicity and the moral foundations features, we track the prevalence or percentage of posts in a subreddit that contain each linguistic feature over time. Table 11 gives the interrupted time series coefficient results for the linguistic features within a quarantined subreddit. Overall, we found no evidence that the quarantines caused a change in either the toxicity or the moral associations expressed in the subreddit. We see similar results when running ITS analysis over the control, invaded, and neighboring subreddits for both The_Donald and ChapoTrapHouse. To illustrate how linguistic trends compare between the quarantined subreddit and its control, invaded, and neighboring subreddits, Figure 2 shows the average toxicity score for sampled posts in each subreddit category over time. Again, we see that the level of toxicity remains around stable for all categories, before and after the quarantine.

Overall, these results seem to suggest that certain elements of language within subreddits, such as toxicity and moral values, remain stable to the interruptions introduced by quarantines. This stability holds even for invaded and neighboring subreddits, which represent priority shifts and

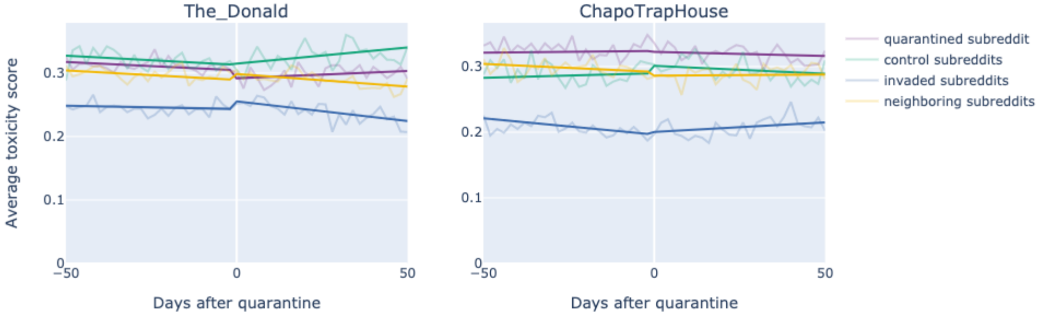

Figure 2: Average toxicity scores over time in The_Donald and ChapoTrapHouse compared with aggregated control, invaded, and neighboring subreddits with fitted ITS regression models.

alternatives to the quarantined subreddits for affected users. One potential explanation for this stability is that when users participate in a subreddit, they adjust their own behaviors to be more similar to that of the general community. As a result, the linguistic norms of a subreddit become entrenched and are very difficult to change. Quarantines, which allow for ongoing participation in controversial subreddits, may therefore not provide a sufficient disruption to substantially change behaviors in the Reddit ecosystem.

Analysis of Linguistic Entrenchment

To test whether a user's language in a subreddit is primarily the result of their accommodation to community norms, we propose an analysis of user-level linguistic trends based on Granger causality (Granger 1969). Granger causality is a statistical method for determining whether a time series X can be used to forecast changes in a target time series Y. X is said to Granger-cause Y if its prior values are significant predictors Y_t beyond previous values of Y themselves. This can be determined by checking the γ coefficients in the following regression model:

$$Y_t = \sum_{i=1}^{n} \alpha_i Y_{t-i} + \sum_{j=1}^{m} \gamma_j X_{t-j} \qquad (2)$$

While Granger causality does not necessarily imply that Y is directly caused by X, it does indicate X has both precedence and significant predictive power over Y.

For our Granger analysis, our goal is to determine whether a subreddit s's prior linguistic tendencies are predictive of the average linguistic feature value of a user's posts in that subreddit at timepoint t. We use a 1-lag Granger model, meaning that we look back one timepoint for the user and subreddit for prediction. We, however, modify the regression setup so that we only predict values for users u who did not participate in subreddit s in the previous timepoint. This is to ensure that the calculated prior linguistic tendencies of users and subreddits are independent of each other.

Table 12 shows the results of the Granger analysis. Due to sampling limitations for Perspective API, we only run this analysis over moral foundation categories. We found that the γ coefficients for subreddit language tendencies were significant for all moral foundations and subreddit types for The_Donald and ChapoTrapHouse. Additionally, the γ coefficients are higher than the α coefficients for authors in all but one of our settings, suggesting that an author's language in a subreddit is more reflective of that subreddit's linguistic tendencies than their own. Overall, this suggests that subreddit linguistic norms are quite stable and users adjust to these norms when participating. As such, quarantines, with their lack of true restrictions on participation, may be limited in their ability to address content issues within communities.

Discussion

From our analyses, we found that quarantines were associated with a general decrease in participation from users who were not ideologically aligned with the subreddits. However, we found no evidence that quarantines impacted the visibility of issues within quarantined subreddits or language in the general Reddit ecosystem. We similarly found few differences in effects between The_Donald and ChapoTrapHouse. Instead, our analyses support the idea that subreddits have stable linguistic norms and users adjust to these norms when participating in different communities.

Implications for Platform Moderation

Prior work examining the impacts of quarantines suggested that they had limited effectiveness for addressing toxicity in online spaces (Chandrasekharan et al. 2020; Ribeiro et al. 2020). Our results support these previous findings, but also provide more insight into why quarantines were ineffective at addressing toxic content. The linguistic norms of subreddits are strongly entrenched, suggesting that interventions that allow ongoing participation but do not directly address content issues may be ineffective as moderation strategies.

Despite the limited effectiveness of quarantines for addressing content issues, we found some evidence that quarantines may have had unintended effects on political polarization. Our results show that quarantines had a homogeniz-

	Invaded Subreddits				Neighboring Subreddits			
Feature	α_{td}	γ_{td}	α_{cth}	γ_{cth}	α_{td}	γ_{td}	α_{cth}	γ_{cth}
Care/Harm	0.313	**0.554**	0.360	**0.520**	0.313	**0.707**	0.360	**0.690**
Fairness/Cheat.	0.305	**0.627**	0.405	**0.437**	0.339	**0.687**	0.381	**0.666**
Loyalty/Betray.	0.404	**0.572**	0.391	**0.511**	0.442	**0.621**	0.363	**0.696**
Authority/Subv.	0.248	**0.672**	0.390	**0.467**	0.291	**0.709**	0.380	**0.657**
Sanctity/Degrad.	0.317	**0.603**	0.320	**0.625**	0.384	**0.640**	0.375	**0.642**
Liberty/Oppress.	0.274	**0.616**	**0.399**	0.387	0.279	**0.746**	0.427	**0.614**

Table 12: Granger causality regression coefficients for linguistic feature values of an author in invaded and neighboring subreddits for The_Donald and ChapoTrapHouse. α gives the coefficient for the previous posts of the author and β gives the coefficient for the previous posts in the target subreddit. All coefficients are significant at $p < 0.001$.

ing effect on political subreddits, limiting exposure to users and content with different beliefs. Additionally, quarantines impacted the discourse surrounding moderation issues on Reddit, with the focus of WatchRedditDie shifting to debate over and antagonism towards the Reddit administration after the quarantine of The_Donald. In exploring effective alternatives to deplatforming political content, further research considering political polarization as a consequence is needed.

Limitations and Future Work

The linguistic insights in this paper are based on labels from two ML systems. While these labels allow us to get a general sense of linguistic trends, our analyses are limited by the capacities of these models. Perspective API, for example, is based on a generalized definition of toxicity, without considering more subjective or community-specific forms of abuse. Similarly, our DistilBERT classifier was trained on a sample of annotated comments which do not fully cover how moral foundations can be invoked on Reddit.

While we examined the quarantines of ideologically distinct political subreddits, we found few differences in effects on The_Donald vs. ChapoTrapHouse. The two subreddits , however, share many properties, such as being high-profile political subreddits. While useful for investigating the interaction between quarantines and ideology, subreddits with very different properties have also been quarantined. Lesser-known subreddits and subreddits centered around other controversial subjects, such as gore or eating disorders, may provide more general insights into quarantine impacts.

In this paper, we take the perspective of examining quarantines as a moderation intervention and evaluate whether or not quarantines lead to outcomes debated by Reddit stakeholders focused on issues of moderation effectiveness. An alternative point of view for examining the impacts of quarantines instead could be to explore the impact of quarantines from the perspective of the communities experiencing the quarantines themselves. Historical tracking of the outcomes of quarantined subreddits suggests that relatively few quarantined subreddits are unquarantined instead of eventually being banned.[7] This is indeed the case with our two subreddits of interest, with The_Donald and ChapoTrapHouse eventually being banned in June 2020 as part of an initiative to crack down on hate speech in light of recent Black Lives Matter protests. One potential consequence from these skewed outcomes, then, is that quarantines can potentially be perceived as a warning or threat to existence for the affected communities. From this perspective and our findings that quarantines have a homogenizing effect on participation, the question remains whether quarantines, viewed as an external threat, pushes affected subreddits to reinforce and advocate for their existence beyond what can be observed with our linguistic features, fundamentally changing how users, new and old, integrate into or influence the community. We leave this threat-centered view of quarantines and whether it changes the participation experiences of users within the affected subreddit for future work.

Conclusion

In this paper, we investigated the quarantines of two ideologically distinct political subreddits, The_Donald and ChapoTrapHouse. Taken as a whole, our analyses suggest that quarantines are ineffective at their intended goal of addressing toxic content and may increase polarization in political spaces. We highlight that future research into moderation should examine the issue of linguistic entrenchment in communities and its implications for intervention design.

References

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. In *ICWSM*.

Bernal, J. L.; Cummins, S.; and Gasparrini, A. 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology* 46(1).

Chandrasekharan, E.; Jhaver, S.; Bruckman, A.; and Gilbert, E. 2020. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *ACM Transactions on Computer-Human Interaction* .

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *CSCW*.

[7] https://www.reddit.com/r/reclassified/

Church, K.; and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1).

Cousineau, L. S. 2020. Displaced Discussion: The Implications of Reddit Quarantine and the Movement of TheRedPill to Self-Hosting. *Selected Papers of Internet Research* .

Datta, S.; and Adar, E. 2019. Extracting Inter-Community Conflicts in Reddit. In *ICWSM*.

Eisenstein, J.; Ahmed, A.; and Xing, E. P. 2011. Sparse additive generative models of text. In *ICML*.

Farrell, T.; Fernandez, M.; Novotny, J.; and Alani, H. 2019. Exploring Misogyny across the Manosphere in Reddit. In *WebSci*.

Flores-Saviaga, C.; Keegan, B.; and Savage, S. 2018. Mobilizing the trump train: Understanding collective action in a political trolling community. In *ICWSM*.

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

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology* 47.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96(5).

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* .

Guimaraes, A.; Balalau, O.; Terolli, E.; and Weikum, G. 2019. Analyzing the traits and anomalies of political discussions on reddit. In *ICWSM*.

Jhaver, S.; Appling, D. S.; Gilbert, E.; and Bruckman, A. 2019. "Did You Suspect the Post Would be Removed?" Understanding User Reactions to Content Removals on Reddit. In *CSCW*.

Jiang, S.; Robertson, R. E.; and Wilson, C. 2020. Reasoning about political bias in content moderation. In *AAAI*.

Johnson, K.; and Goldwasser, D. 2018. Classification of moral foundations in microblog political discourse. In *ACL*.

Kelly, J.; Fisher, D.; and Smith, M. 2005. Debate, division, and diversity: Political discourse networks in USENET newsgroups. In *Online Deliberation Conference*.

Kiesler, S.; Kraut, R.; Resnick, P.; and Kittur, A. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design* .

Martinez, I. 2019. Chapo Trap House subreddit quarantined for allegedly encouraging violence. *The Daily Dot* .

Massachs, J.; Monti, C.; Morales, G. D. F.; and Bonchi, F. 2020. Roots of Trumpism: Homophily and Social Feedbackin Donald Trump Support on Reddit. In *WebSci*.

Massanari, A. 2020. Reddit's Alt-Right: Toxic Masculinity, Free Speech, and/r/The_Donald. *Fake News: Understanding Media and Misinformation in the Digital Age* .

Morales, G. D. F.; Monti, C.; and Starnini, M. 2021. No echo in the chambers of political interactions on Reddit. *Scientific Reports* 11(1).

Olson, R. S.; and Neal, Z. P. 2015. Navigating the massive world of Reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science* .

Ribeiro, M. H.; Jhaver, S.; Zannettou, S.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and West, R. 2020. Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *arXiv preprint* .

Robertson, A. 2015. Was Reddit always about free speech? Yes, and no. *The Verge* .

Robertson, A. 2019. Reddit quarantines Trump subreddit r/The_Donald for violent comments. *The Verge* .

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint* .

Schmidt, A.; and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *SocialNLP*.

Shen, Q.; and Rosé, C. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in reddit's quarantine policy. In *Third Workshop on Abusive Language Online*.

Srinivasan, K. B.; Danescu-Niculescu-Mizil, C.; Lee, L.; and Tan, C. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. In *CSCW*.

Stewart, E. 2019. Reddit restricts its biggest pro-Trump board over violent threats. *Vox* .

Sybert, J. 2021. The demise of #NSFW: Contested platform governance and Tumblr's 2018 adult content ban. *New Media & Society* .