The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy

Qinlan Shen
Carnegie Mellon University
qinlans@cs.cmu.edu

Carolyn P. Rosé
Carnegie Mellon University
cprose@cs.cmu.edu

Abstract

Recent concerns over abusive behavior on their platforms have pressured social media companies to strengthen their content moderation policies. However, user opinions on these policies have been relatively understudied. In this paper, we present an analysis of user responses to a September 27, 2018 announcement about the quarantine policy on Reddit as a case study of to what extent the discourse on content moderation is polarized by users' ideological viewpoint. We introduce a novel partitioning approach for characterizing user polarization based on their distribution of participation across interest subreddits. We then use automated techniques for capturing framing to examine how users with different viewpoints discuss moderation issues, finding that right-leaning users invoked censorship while left-leaning users highlighted inconsistencies on how content policies are applied. Overall, we argue for a more nuanced approach to moderation by highlighting the intersection of behavior and ideology in considering how abusive language is defined and regulated.

1 Introduction

In response to the rising surge of abusive behavior online, large social media platforms, such as Facebook, Twitter, and Youtube have been pressured to strengthen their stances against offensive content and increase their transparency in how content policies are enforced. Facebook, for example, first released its community standards publicly in April 2018 and has made efforts to ban white nationalist and separatist content (Stack, 2018), while Twitter announced a new policy against "dehumanizing speech" in September 2018 (Matsakis, 2018).

Nevertheless, the problem of how to define what behaviors are abusive and how these behaviors should be handled remains a challenge. One major issue in terms of defining a content policy for a major platform is that defining what abusive behavior is requires consideration of both behavior *and* ideology – political ideology is inextricably tied with abusive language on major platforms where sensitive discussion can occur. For example, Reddit (Statt, 2018) and Twitter (Newton, 2019) have faced recent backlash for allowing racist content to remain on their platforms over concerns of bias against right-leaning viewpoints. Prior research (Shen et al., 2018; Jiang et al., 2019) has also demonstrated that ideology can be used as a tool to challenge moderation decisions.

In this paper, we argue that ideology is inextricably tied to how abusive language is defined and regulated in real-world applications in social media. To demonstrate the role of political ideology in the problem of defining abusive language, we present the first NLP study of polarized user responses towards policy. We examine how users frame their arguments in supporting or opposing stronger moderation policies to draw insight into ideologically-related user concerns over their impact. As a case study, we focus on users' responses towards changes to the quarantine policy on Reddit.[1] Reddit provides an interesting site of study into content moderation issues due to a culture of debate over whether free speech is a principal tenet of the platform (Robertson, 2015). Here, we focus on a specific policy change to provide an in-depth analysis of the polarized stances users take.

The rest of the paper is organized as follows. First, we give an overview of related work and describe the recent Reddit quarantine policy update. Next, we present a general topic analysis of discussion surrounding the quarantine policy. We then describe how we operationalized polarization by characterizing users based on their participation across subreddits, then examine how different

[1] https://www.reddit.com/r/announcements/comments/9jf8nh/

Proceedings of the Third Workshop on Abusive Language Online, pages 58–69
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

users frame issues within topics. Finally, we discuss the implications and limitations of our work.

2 Related Work

One of the primary roles of moderation in online spacesis the regulation of anti-social behaviors (Kiesler et al., 2012), such as spamming, cyberbullying, and hate speech. The design and best practices for moderating abusive content on large social media platforms, however, is a fundamentally challenging issue (Gillespie, 2018), due to the tension between providing a space for open and meaningful interaction and determining what behaviors are acceptable and how unacceptable behaviors should be handled. While social media companies, as private organizations, can legally curate content on their platforms (Robertson, 2015), cracking down on content can lead to tension with users, who may view it as setting a precedent for banning behaviors or even political ideologies in the future. Previous research Shen et al. (2018); Jiang et al. (2019), has demonstrated that tensions and backlash can arise in communities if participants perceive moderation decisions as biased against minority viewpoints, even if decisions seem "fair" after accounting for behavior.

Previous research on the effect of moderation policies has focused primarily on the effect of moderation on directly affected users. For example, Chandrasekharan et al. (2017) investigated the impact of the 2015 Reddit hateful content ban on users who participated on the banned subreddits, while Chang and Danescu-Niculescu-Mizil (2019) examined the participation trajectories of users blocked by community moderators on Wikipedia. User opinions on moderation policies, however, remains relatively understudied from a large-scale quantitative perspective, though previous work has drawn insights from structured interviews and surveys with users. Jhaver et al. (2018) interviewed both users who used blocklists on Twitter and users who have been blocked on their insights about harassment and blocking. Myers West (2018) surveyed participants on OnlineCensorship.org about their experiences with content moderation to gather insights into folk theories about how moderation policies work.

Most closely related to our work, which focuses on ideologically motivated user viewpoints, Jhaver et al. (2017) used a mixed-methods approach to investigate how users on r/KotakuInAction, a subreddit associated with the Gamergate movement, view free expression, harassment, and censorship within their own community. Rather than focusing on users who share certain views within a particular subreddit, however, we focus on users who responded to a Reddit-wide moderation policy change. This allows us to examine how users who have participated across a wide range of subreddits present their opinions, with the goal of understanding what elements of the debate between moderation and censorship are polarized.

3 Reddit Quarantine Policy Announcement

On September 27, 2018, Reddit announced changes to their quarantine policy in response to growing concerns over the visibility of offensive content on their platform. The quarantine feature allows site administrators to hide "communities that, while not prohibited, average redditors may nevertheless find highly offensive or upsetting"[2] from being searched, recommended, or monetized. While the quarantine function was initially announced in August 2015 as part of a broader initiative to address offensive content, the September announcement specifically focused on expanding use of the quarantine function. The two major aspects of the announcement were 1) a quarantine wave of 20+ communities of interest or *subreddits* and 2) the introduction of an appeals process for moderators of quarantined subreddits.

The announcement was posted in the r/announcements subreddit, which allows users to respond to major Reddit-internal policy changes. To investigate the discourse surrounding the announcement, we collected comments that were posted in response to the r/announcements over the course of one month using the Pushshift API.[3] After filtering out 6 comments that were deleted by users or removed by moderators, as we no longer had access to the original comment texts, we then identified 13 well-known meta-bots[4] among the remaining users. Both comments by and responses to these meta-bots were removed,

[2] https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/quarantined-subreddits

[3] https://pushshift.io/api-parameters/

[4] CommonMisspellingBot, WikiTextBot, Link-Help-Bot, YTubeInfoBot, HelperBot_, LimbRetrieval-Bot, BigLebowskiBot, FatFingerHelperBot, RemindMeBot, imguralbumbot, opinionated-bot, societybot, svenska_subbar

Topic	Top Words
T0: Accessibility of Quarantined Content (13.6%)	quarantine, reddit, subs, subreddit, content, community, view, find, offensive, list, users, mobile, quarantining, site, access
T1: Heated Outbursts (11.7%)	shit, fuck, lol, racist, ca [CringeAnarchy], literally, stop, td [the_Donald], stupid, show, love, dude, alt, call, thread, leftist
T2: Content in r/The_Donald (11.2%)	t_d, ban, post, subreddit, the_donald, propaganda, admins, rules, russian, subs, users, violence, racism, page, link
T3: Conservative vs. Liberal Politics [U.S.] (10.1%)	trump, politics, left, time, wing, posts, evidence, comments, day, stuff, donald, top, ago, hard, conservative
T4: Censorship of Political Views/Debate (9.8%)	people, bad, censorship, agree, make, wrong, political, point, opinions, disagree, thought, fact, ideas, understand, discussion, feel
T5: Moderation/Free Speech on Social Media (9.2%)	reddit, speech, free, hate, hitler, site, heil, internet, platform, thing, censorship, website, private, open, freedom
T6: Far-Right/Far-Left Ideologies (9.0%)	white, nazi, anti, people, genocide, holocaust, support, great, fascist, jews, communism, capitalism, country, claim, socialism
T7: Personal Experience (7.1%)	people, things, talking, thing, time, men, matter, person, real, years, talk, life, made, lot, world
T8: Laws/Government-level Policies (6.2%)	people, society, violence, person, power, words, point, world, rights, groups, political, majority, control, argue, definition, part
T9: Miscellaneous (12.0%)	good, make, ca, yeah, read, back, man, money, question, side, wo, big, end, full, care

Table 1: Identified topics, proportion in our dataset, and top 15 associated words. Topic names were assigned after examining both the top words and the top comments associated with each topic.

as they are usually formulaic and unrelated to the content of our analyses (e.g. "Good bot", complaints about bot responses), leaving us with a final announcement dataset containing 9,836 posts from 3,640 users.

4 Topical Analysis

Topic choice has been commonly used in NLP (Tsur et al., 2015; Field et al., 2018; Demszky et al., 2019) as a proxy for *agenda-setting*, the strategic highlighting of what aspects of a subject are worth discussing (McCombs, 2002). Here, we first describe our preliminary topic analysis for discovering the range of topics discussed.

4.1 Models

We used Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to construct our topics. While Structural Topic Models (STM) (Roberts et al., 2013) are popular for social science analyses for enabling document metadata to act as topic covariates, STM consistently performed worse than LDA on our data, both in topical coherence measures and human interpretability.[5]

For the LDA models, we considered each comment to be a document. Comments were tokenized using SpaCy (Honnibal and Montani, 2017) and stopwords and punctuation-only tokens were removed. We trained models with 5, 10, 15, 20, 25, 30, 40, and 50 topics. We selected the model with 10 topics for further analysis for having the highest CV coherence, which has been shown to more closely correlate with human ratings of interpretability (Röder et al., 2015) than semantic coherence (Mimno et al., 2011). When analyzing and interpreting the topics discovered, we examined both the highest weighted words and example comments associated with each topic.

4.2 Results

Table 1 presents the topics discovered by the model. The most prevalent topic (**T0**) in the discussion thread focuses on accessibility to quarantined subreddits. This is unsurprising, as this

[5] A potential challenge for STM for our data is the lack of global consistency in our metadata. Comments in Reddit threads are organized in broad semi-topical hierarchical trees and threads can contain thousands of comments (Weninger, 2014). As a result, user participation on a single thread can be scattered and upvoted comments in one subthread may substantially overlap in content with downvoted comments in another. Thus, the simpler LDA model, with fewer global priors on the structure and content of the data, may have better generalization.

topic directly addresses the short-term impact of the quarantine wave, such as the ability to search for and list quarantined subreddits, access to quarantined content on the mobile app, and whether quarantined content will generate ad revenue. The proportion of **T0** across comments, however, is relatively low (13.6%), compared to discussion centered around the broader implications of quarantining. For example, **T3:** Conservative vs. Liberal Politics and **T6:** Far-Right/Far-Left Ideologies center around broader ideologies associated with controversial content, while **T4:** Censorship of Political Views/Debate, **T5:** Moderation/Free Speech on Social Media Platforms, and **T8:** Laws/Government-Level Policies discuss the legal implications of online content moderation.

One notable topic in our model was **T2:** Content in r/The_Donald. Despite not being one of the subreddits quarantined during the quarantine wave, much of the discussion surrounding the announcement centered around The_Donald, due to its prominent reputation for controversial behavior. We can see evidence of discussion about controversial behavior on The_Donald, as many of the highly weighted words in the discussion of The_Donald are words describing negative behaviors that have been associated with the subreddit in past research, such as propaganda/fake news (Kang, 2016), promotion of violence and racism (Squirrell, 2017), and visibility manipulation and mobilization through bots (Carman et al., 2018; Flores-Saviaga et al., 2018). The_Donald is often considered an “elephant in the room” with regards to content moderation on Reddit, as the subreddit remains one of the most visible and active subreddits on the site despite its controversial reputation.

A somewhat surprising omission from the topics discovered was discussion around the new appeals process for quarantined subreddits. While the bulk of the text in the original post of the thread centered around the introduction of the appeals process, only 0.13% of the posts explicitly used the words “appeal” and “appeals” in reference to the appeals policy. The addition of an appeals process is relatively uncontroversial for increasing the transparency of quarantines and primarily affects moderators of quarantined subreddits. This suggests that what *is* driving discussion within the thread are the more controversial issues that may have a personal, ideological impact on users. As a result, we expect that users with differing viewpoints may highlight different aspects within the general topics discussed here.

5 Characterizing User Participation on Reddit

In order to better understand how different users highlight or *frame* particular aspects within each topic (Entman, 2007; Nguyen et al., 2013; Card et al., 2016), we first want to characterize the types of users who participated in the r/announcements discussion. Because subreddits on Reddit represent interest-based subcommunities, previous work has used participation across subreddits as a signal of user interests or viewpoint (Olson and Neal, 2015; Chandrasekharan et al., 2017). We follow in the lines of this work by characterizing users using their participation in subreddits prior to the announcement. In this section, we describe a graph-partitioning approach for characterizing common interests across subreddits.

5.1 Constructing the Interest Graph

For each user who participated in the r/announcements quarantine thread, we collect all submissions and comments posted by the user in the month preceding the quarantine policy update (August 27 - September 26). We then counted how many times each user posted in each subreddit. In order to ensure that users both showed sustained interest in a subreddit and to limit the number of users who participate in subreddits to challenge the widely held view of a subreddit, we consider a user to be interested in a subreddit if they have posted at least 3 times[6] in the preceding month with a positive score.

To capture similarities between the subreddits users participate in, we then cluster them by performing graph partitioning over a subreddit interest graph (Olson and Neal, 2015). We construct a subreddit interest graph by drawing an undirected edge e_{ij} between two subreddit nodes i and j if the same user participates in both subreddits. A_{ij}, the weight of e_{ij}, is set equal to the number of users in common between i and j. We reduce the number of edges in the graph by setting a global edge threshold $A_{ij} >= 5$.[7]

[6] The threshold was determined based on the distribution of user-subreddit participation pairs across users who participated in the r/announcements thread.

[7] While we can threshold the edges of a graph using a significance-based backbone extraction algorithm, our subreddit graph is based only on the users from the

Category	Central Subreddits	Accuracy	Cohen's κ
C0: Tech/Sports	technology, Games, pcmasterrace, nba, PS4,	56.25	68.31
C1: Internet Compilation	WTF, WhitePeopleTwitter, trashy, BlackPeopleTwitter, mildlyinfuriating	84.38	75.13
C2: Right-Leaning	CringeAnarchy, unpopularopinion, the_Donald, Libertarian, TumblrInAction	78.13	66.14
C3: Memes	greentext, starterpacks, dankmemes, PrequelMemes, MemeEconomy	50.00	27.64
C4: Left-Leaning	TopMindsOfReddit, SubredditDrama, ChapoTrapHouse, The_Mueller, FuckTheAltRight	81.25	52.71

Table 2: Identified subreddit categories, central subreddits, averaged annotator performance and agreement on intrusion task.

5.2 Louvain Community Detection

We use the Louvain community detection algorithm (Blondel et al., 2008) to define a partition over the constructed subreddit interest graph. The objective of the Louvain algorithm is to maximize the *modularity* of a partition, which measures the density of links within vs. between communities. The Louvain modularity Q is defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \qquad (1)$$

where $k_i = \sum_j A_{ij}$ is the sum of the weights of edges attached to node i, $\delta(c_i, c_j) = 1$ if nodes i and j belong to the same community, 0 otherwise, and $m = \frac{1}{2} \sum_{i,j} A_{ij}$. Because ΔQ from moving node i from one community to another is easy to compute, the algorithm finds the best partition through a simple two-stage process:

1. Assign each node to its own community
2. Repeat until convergence
   (a) Iterate through nodes i, moving i into the community that gives the highest increase in modularity, until convergence.
   (b) Construct new graph where nodes are communities and edge weights between communities are equal to sum of edge weights between lower-level nodes.

We use a resolution factor (Lambiotte et al., 2008) of 1.0 and select the highest modularity partition of the dendrogram for our subreddit categories. The resulting 5 categories are shown in Table 2.

5.3 Evaluation

To ensure that the 5 discovered subreddit categories gave us high-quality and coherent notions of user interests, we run a human evaluation of the discovered categories using a subreddit intrusion task, analogous to word intrusion tasks used for evaluating topic model interpretability (Chang et al., 2009). The subreddit intrusion task was presented to two native English speaker annotators who used Reddit on a daily basis to ensure familiarity with the types of user interests on Reddit. Given a set of four subreddits belonging to one of the categories, and an "intruder" subreddit from another category, annotators were asked to identify the intruder. Annotators were provided with the description and 5 highly-ranked thread titles for each subreddit for additional context in determining the intruder. For each category, all the other categories were selected as an intruder instance 4 times, giving us 16 sets per category. After completing the intrusion task, the annotators discussed their decision-making process during the intrusion task and assigned labels to the five discovered subreddit categories.

Results for the intrusion task for each category are included in Table 2. For all the subreddit categories except **C3:** Memes, the annotators achieved moderate-to-high agreement and performed significantly better than a random baseline. The category of **C3:** Memes is more abstract compared to the other categories and contains many subreddits that are not easily identifiable by name and description alone. Nevertheless, the annotators were able to reach an agreement on the interests covered by **C3** in discussion after the intrusion task.

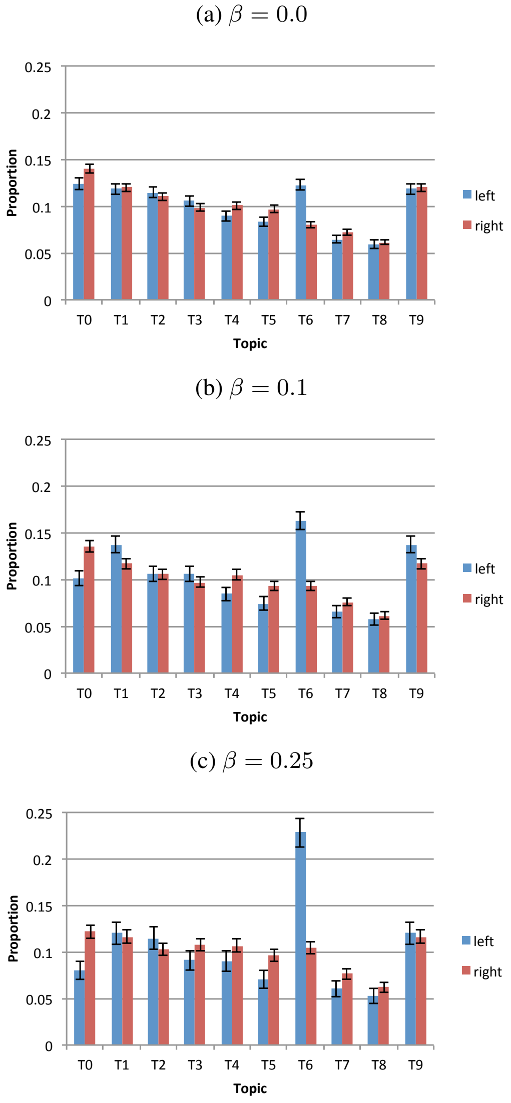
From these discovered subreddit categories, for each user, we calculate their distribution of par-

r/announcements thread. As a result, a significance-based method of thresholding edges can give uneven results based on how many users were sampled from each subreddit.

ticipation across the five categories and an additional category for unidentified subreddits. One limitation of considering user viewpoints based on these categories, however, is that only **C2:** Right-Leaning and **C4:** Left-Leaning are directly related to political viewpoint. Rather, these five categories more closely represent shared sets of interests or personas users can engage in. While this limits what we can say in terms of polarization across the traditional definitions of left-leaning vs. right-leaning political ideologies, we argue that considering user participation in these interest categories is more representative of how users on Reddit engage in politics across the site.

6 Analyzing Polarized Viewpoints Towards the Quarantine Policy

In the previous sections, we first identified the general topics discussed within the r/announcements thread about the quarantine policy. We then characterized users who participated in the r/announcements thread based on their distribution of participation across different subreddits in the month preceding the announcement. In this section, we examine the relationship between a user's ideological views and how they strategically highlight particular aspects of each topic. Rather than using a static left vs. right framework for operationalizing user viewpoint, we examine how users highlight different aspects as they move along the left-right spectrum. We then analyze the relationship between users' polarization and their framing within the topics identified in Section 3 in an unsupervised manner.


(a) $\beta = 0.0$

(b) $\beta = 0.1$

(c) $\beta = 0.25$

Figure 1: Topic prevalence across left and right-leaning users at different levels of polarization, with 95% confidence intervals.

6.1 User Polarization

While we can label users strictly as left vs. right based on whether they spend more of their time on left-leaning and right-leaning subreddits in their participation distribution, we can get a more nuanced view of the differences between left-leaning and right-leaning users by additionally considering how polarized users are along the left-right spectrum. Rather than using a simple majority-based assignment, we introduce a polarization margin hyperparameter β that controls for how skewed a user must be towards one side to be considered a left-leaning or right-leaning user. For a given β, we can assign the class of each user u_i based on their participation distribution p:

$$C_\beta(u_i) = \begin{cases} \text{left}, & \text{if } p_l(u_i) - p_r(u_i) > \beta \\ \text{right}, & \text{if } p_r(u_i) - p_l(u_i) > \beta \\ \text{neutral}, & \text{otherwise} \end{cases} \tag{2}$$

$\beta = 0$ is equal to the majority case. For our remaining analyses on agenda-setting and framing, we compare results for $\beta = \{0, 0.1, 0.25\}$.

6.2 Polarized Agenda-Setting

Figure 1 shows the prevalence of each topic across left-leaning and right-leaning users at differing values of β. We found that right-leaning users were significantly more likely to invoke **T0:** Accessibility of Quarantined Content, **T4:** Censorship of Political Views/Debate, and **T5:** Moderation/Free Speech on Social Media for all values of β. The high prevalence **T0** is unsurprising, as the majority of the newly quarantined subreddits (listed in the Supplementary Material) were associated with conservative views and users. Thus, accessibility to the newly quarantined subreddits would be a concern for many right-leaning users. The increased prevalence of topics **T4** and **T5**, which are focused on the relationship between content moderation online spaces and censorship, suggests that right-leaning users may be challenging the ability or approach of Reddit administrators to expand the quarantine policy as a form of censorship. Finally, the higher prevalence of **T7:** Personal Experience topic, which is focused on users' personal participation on the quarantined or other controversial subreddits, suggests that to some extent, right-leaning users are leaning into their participation on controversial subreddits in their responses towards the announcement.

Across all values of β, left-leaning users use **T6:** Far-Right/Far-Left Ideologies significantly more than right-leaning users. This difference increases as the polarization margin β increases. This suggests that left-leaning users were likely to invoke the controversial behaviors associated with the extremism, particularly the far-right. Interestingly, while extremist ideology is more likely to be invoked by left-leaning users, there was no significant difference in prevalence between left-leaning and right-leaning users for discussion of US politics (**T3:** Conservative vs. Liberal Politics).

Overall, we note that while the relative prevalence of topics for left-leaning and right-leaning users generally remained the same at different values of β, the major differences between left-leaning and right-leaning users became larger as we increase the polarity margin.

6.3 Within-Topic Framing

We expect users who have different positions to highlight different aspects of each topic. To separate out the salient words within each topic t for left-leaning and right-leaning users, for each word w, we use the z-score of the log-odds ratio with a Dirichlet prior (Monroe et al., 2008) as a salience score, $\delta_w^{r(t)-l(t)}$:

$$\delta_w^{c(t)} = \log \frac{y_w^{c(t)} + \alpha_w^t}{n^{c(t)} + \alpha_0^t - (y_w^{c(t)} + \alpha_w^t)} \tag{3}$$

$$\delta_w^{r(t)-l(t)} = \delta_w^{r(t)} - \delta_w^{l(t)} \tag{4}$$

$$\sigma(\delta_w^{r(t)-l(t)}) = \frac{1}{y_w^{r(t)} + \alpha_w^t} + \frac{1}{y_w^{l(t)} + \alpha_w^t} \tag{5}$$

$$z(\delta_w^{r(t)-l(t)}) = \frac{\delta_w^{r(t)-l(t)}}{\sqrt{\sigma(\delta_w^{r(t)-l(t)})}} \tag{6}$$

where $n^{c(t)}$ is the number of words in corpus c, $y_w^{c(t)}$ is the count of word w in corpus $c(t)$, $l(t)$ and $r(t)$ are the left-leaning and right-leaning corpora for topic t, and α_0^t and α_w^t are corpus and word priors from a background corpus. We set the Dirichlet prior by using the posts from "neutral" users as a background corpus, with the size and count of words in the background corpus as the corpus and word priors respectively.. We extend the salience score to bigrams and trigrams and sampled posts containing the top 50 salient terms for each topic and faction to analyze framing strategies at different levels of polarization.

First, we found that, across topics, right-leaning users framed the issues surrounding content moderation in terms of censorship and suppression, while left-leaning users tended to frame issues in terms of consistency. For example, in **T4:** Censorship of Political Views/Debate, right-leaning users consistently used terms such as "silencing", "echo chamber", and "censorship" in reference to impact of the announcement, directly accusing the quarantine policy of being used to silence political viewpoints. This supports our hypothesis from Section 5.2 that right-leaning users invoked **T4** to criticize the quarantine policy as a form of censorship. On the other hand, when left-leaning users invoked **T4**, they used terms such as "picking and choosing", "bad faith" in reference to uneven and insufficient application of the policy. Left-leaning users also often compared the quarantine feature to "bans" in **T4**, arguing that many subreddits quarantined under the announcement shared similarities with subreddits that were banned in the past.

We see similar patterns in **T5**: Moderation/Free Speech on Social Media, though many of the salient terms used are specific to internet platforms. Right-leaning users emphasize the ideal of a free and open internet, using terms such as "open platforms" and invoking the name of "Aaron Swartz", the late Reddit co-founder known for his anti-censorship views. Left-leaning users, on the other hand, consistently highlighted that private organizations like Reddit ("private company", "privately owned") had the right to remove or hide content in violation of their policies.

One of the more salient framing strategies related to consistency by left-leaning users is the comparison of quarantines with Reddit's handling of pornographic content, primarily in **T0:** Accessibility of Quarantined Content and **T8:** Laws/Government-level Policies. While opinions about how to handle porn on Reddit are mixed, porn is commonly used as an analogue for many of the consistency issues involved with quarantining subreddits with abusive language. For example, some users argue that the intent and functionality of quarantining should be similar to the not-safe-for-work (NSFW) filtering system already in place for pornographic subreddits, which does not explicitly block a subreddit from being searched or shown in r/all. Others compare the liability of hosting pornography vs. other forms of offensive content, such as violence or hate speech.

We also found that across factions, users tried to highlight controversial, even violent, behavior by users on the opposite side. In Section 5.2, while we suggested that left-leaning users invoked **T6:** Far-Right/Far-Left Ideologies to highlight controversial behaviors in far-right subreddits, **T6** is also associated with talk surrounding the quarantine of r/FULLCOMMUNISM, described as a "self-aware socialist satire sub". Thus, invocation of **T6** may also be reflective of their personal investment in participating in a quarantined subreddit. We see, however, that discussions about "socialism" and "communism" are highly salient for right-leaning users, who commonly accused subreddits associated with these ideologies of supporting dictatorships and inciting violence. Similarly, for left-leaning users,"nazi", "ethnic", "fascist", and "genocide" are highly salient in **T6**, which were used to argue that many right-leaning subreddits, quarantined or not, expressed racist views, supported fascism, and denied genocides.

The framing strategy of highlighting controversial behavior from the opposing viewpoint was also apparent in **T2:** Content in r/The_Donald. While the most salient terms for right-leaning users focused on the how The_Donald governs itself ("admins", "moderators", "users", "rules"), left-leaning users explicitly emphasized that the_Donald has content encouraging violence ("kill", "doxxing", "encouraged", "attacking", "spread"). One of the most common associations between The_Donald and incitement of violence cited by left-leaning users was the case of u/Seattle4Truth, a The_Donald user, who murdered his own father (Neiwert, 2017).

Like with our analysis of topic choice, the specific strategies on each side remained generally consistent at the different levels of polarity.

7 Discussion

From our analysis, we find that right-leaning users tend to frame the issues surrounding content moderation in terms of censorship of political viewpoints, while left-leaning users highlight the issues surrounding consistency in how moderation is applied, especially in regards to unmoderated offensive content. On the surface, these findings seem to reflect stereotypes about how freedom of expression is viewed by liberals and conservatives offline in the debate over campus free speech (Friedman, 2019). However, we argue that the emphasis on censorship vs. consistency is not entirely reflective of stereotypical, surface-level differences between conservative and liberal viewpoints on the tension between moderation and free speech. Both left-leaning and right-leaning users, for example, used statements decrying both hate speech and censorship and highlighted concerns with how the Reddit quarantine policy was implemented. Instead, we argue that these strategies are employed as a defense of a user's legitimate participation on Reddit. While previous work has examined the use of free speech discourse as a defense against ego or expressive threat (White et al., 2017), further exploration is needed into why the specific strategies of censorship vs. consistency are applied in the context of online discussion.

As an example for needing more nuance in understanding how opinions on policy are used strategically in argumentation, one common framing strategy we see across both sides is the association of opposing viewpoints with the incitement or encouragement of violence. The question of whether something incites or encourages violence is important, as the encouragement and incitement of violence is explicitly prohibited by

Reddit's content policy.[8] While "encouraging and inciting violence" provides a more concrete frame of judgment than broader definitions of offensive language, there still is ambiguity in terms of how administrators should respond to content that violates Reddit policy, especially on the level of broader communities. At the level of subreddits, it is unclear to what extent a community has to demonstrate violent behavior before the administrators take action to quarantine or ban a subreddit. Many users[9] argue that this ambiguity allows for the Reddit administration to protect popular but controversial subreddits like The_Donald.

7.1 Limitations and Future Work

Our work in this paper is focused on polarized responses to a specific content moderation policy change on Reddit. While we perform an in-depth analysis of the issues raised by the quarantine policy change, our findings may be specific to the context surrounding this particular event, such as the majority of subreddits quarantined in conjunction with the announcement being right-leaning. A longitudinal analysis, where we examine responses to announcements affecting content moderation on Reddit over time may give us a more general view of how users on Reddit talk about free speech and how the discourse of free speech on Reddit has evolved in response to major events. As of June 2019, there have not been other major notifications regarding moderation policy changes in the r/announcements subreddit since the quarantine policy changes. Nevertheless, finding textual signals of user opinions for other moderation-related events, like the progression and eventual banning of quarantined subreddits (e.g. CringeAnarchy, watchpeopledie), remains an interesting area of study.

While we introduced the polarization margin as a method for capturing differences beyond a static left vs. right ideological assignment over users, we found very few differences between users in the same class at different levels of polarization. One limitation of our approach, however, is that we still rely on a hard left-right distinction at the different values of polarization margin β. Relaxing the assumption that users must be assigned to a class for our topic choice and salience analyses and instead using the raw distribution of participation across all subreddit categories may give us better insight into the range of users' framing strategies across a wider, more nuanced range viewpoints.

7.2 Ethical Considerations

The investigation of the discourse surrounding the Reddit quarantine policy requires us to handle sensitive information related to users' political leanings. To limit the impact of this study on users' privacy and participation on Reddit (Fiesler and Proferes, 2018), usernames were only used to collect user activity outside of the r/announcements thread. After data collection, all usernames were anonymized by replacement with a random numeric id. Additionally, this study focuses on the relationship between discussion about moderation and polarization in aggregate. Though individual researchers viewed example posts, these posts were not matched with individual users by either username or id. Finally, while the full anonymized data from the r/announcements thread is publicly available[10], we only release the user distribution across subreddit categories to prevent the user tracking across subreddits.

8 Conclusion

In this paper, we used techniques for examining agenda-setting and framing to investigate how users discuss their opinions on an update to Reddit's quarantine policy. We presented a novel approach for operationalizing user polarization for our framing analyses, finding that as a whole, right-leaning users tended to invoke censorship while left-leaning users tended to invoke consistency in how policies are applied. While this seems to reflect stereotypes about how freedom of expression is viewed by conservatives and liberals, we argue for a more nuanced view of formalizing differences in how users frame their opinions about policy. Overall, this work builds towards understanding the relationship between ideology and policy with regards to offensive language.

Acknowledgments

This research was supported in part by NSF Grant DGE1745016 and the K&L Gates Presidential Scholarship Fund. We thank Michael Miller Yoder, Daniel Clothiaux, and the anonymous reviewers for their helpful comments and feedback.

[8] https://www.redditinc.com/policies/content-policy

[9] See r/AgainstHateSubreddits, which tracks behaviors across subreddits that violate Reddit's content policy.

[10] https://github.com/qinlans/alw3_data

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR).*

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment.*

Dallas Card, Justin Gross, Amber Boydstun, and Noah A Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Mark Carman, Mark Koerber, Jiuyong Li, Kim-Kwang Raymond Choo, and Helen Ashman. 2018. Manipulating Visibility of Political and Apolitical Threads on Reddit via Score Boosting. In *Proceedings of the IEEE International Conference On Trust, Security And Privacy In Computing And Communications.*

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW).*

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems (NIPS).*

Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. *arXiv preprint arXiv:1902.08628.*

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL).*

Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication.*

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society.*

Claudia I. Flores-Saviaga, Brian C. Keegan, and Saiph Savage. 2018. Mobilizing the Trump train: Understanding collective action in a political trolling community. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).*

Jonathan Friedman. 2019. Chasm in the Classroom: Campus Free Speech in a Divided America. Technical report, PEN America.

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

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear.*

Shagun Jhaver, Larry Chan, and Amy Bruckman. 2017. The view from the other side: The border between controversial speech and harassment on kotaku in action. *arXiv preprint arXiv:1712.05851.*

Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI).*

Shan Jiang, Ronald E Robertson, and Christo Wilson. 2019. Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation.

Cecilia Kang. 2016. Fake news onslaught targets pizzeria as nest of child-trafficking. *The New York Times. https://www.nytimes.com/2016/11/21/technology/fact-check-this-pizzeria-is-not-a-child-trafficking-site.html. Accessed.*

Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design.*

Renaud Lambiotte, J-C Delvenne, and Mauricio Barahona. 2008. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770.*

Louise Matsakis. 2018. Twitter releases new policy on "dehumanizing speech". *Wired. https://www.wired.com/story/twitter-dehumanizing-speech-policy/. Accessed.*

Maxwell McCombs. 2002. The agenda-setting role of the mass media in the shaping of public opinion. In *Proceedings of the 2002 Conference of Mass Media Economics.*

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*.

Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*.

David Neiwert. 2017. Alt-righter 'Seattle4Truth' charged with killing father over conspiracy theories. *Southern Poverty Law Center. https://www.splcenter.org/hatewatch/2017/10/23/alt-righter-seattle4truth-charged-killing-father-over-conspiracy-theories. Accessed.*

Casey Newton. 2019. Why Twitter has been slow to ban white nationalists. *The Verge. https://www.theverge.com/interface/2019/4/26/18516997/why-doesnt-twitter-ban-nazis-white-nationalism. Accessed.*

Viet-An Nguyen, Jordan L Ying, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1106–1114.

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

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. 2013. The structural topic model and applied social science. In *Advances in Neural Information Processing Systems (NIPS) Workshop on Topic Models: Computation, Application, and Evaluation*.

Adi Robertson. 2015. Was Reddit always about free speech? Yes, and no. *The Verge. https://www.theverge.com/2015/7/15/8964995/reddit-free-speech-history. Accessed.*

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*.

Qinlan Shen, Michael Miller Yoder, Yohan Jo, and Carolyn P Rosé. 2018. Perceptions of Censorship and Moderation Bias in Political Debate Forums. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Tim Squirrell. 2017. Linguistic data analysis of 3 billion Reddit comments shows the alt-right is getting stronger. *Quartz. https://qz. com/1056319/what-is-the-alt-righta-linguistic-data-analysis-of-3-billion-reddit-comments-shows-a-disparate-group-thatis-quickly-uniting/. Accessed.*

Liam Stack. 2018. Facebook Announces New Policy to Ban White Nationalist Content". *The New York Times. https://www.nytimes.com/2019/03/27/business/facebook-white-nationalist-supremacist.html. Accessed.*

Nick Statt. 2018. Reddit CEO says racism is permitted on the platform, and users are up in arms. *The Verge. https://www.theverge.com/2018/4/11/17226416/reddit-ceo-steve-huffman-racism-racist-slurs-are-okay. Accessed.*

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the Annual Meeting of the Assocation for Computational Linguistics (ACL)*.

Tim Weninger. 2014. An exploration of submissions and discussions in social news: Mining collective intelligence of Reddit. *Social Network Analysis and Mining*.

II White, H Mark, and Christian S Crandall. 2017. Freedom of racist speech: Ego and expressive threats. *Journal of Personality and Social Psychology*.

A Quarantined Subreddits

Here, we list the subreddits included in the quarantine wave associated with the announcement, with their status as of May 3rd, 2019. All these following subreddits were quarantined on September 27-28th, though some have been banned or privatized by their moderators in the meantime:

- **Quarantined:** theredpill, Ice_Poseidon, FULLCOMMUNISM, Braincels, 911truth, WhiteBeauty, fragilejewishredditor, WhiteNationalism, GentilesUnited, ZOG, AmericanJewishPower, CringeChaos, NorthwestFront, BritishJewishPower, mayo_town, Ice_Poseidon2
- **Banned:** watchpeopledie, CringeAnarchy, hearpeopledie, SubOfPeace, White_Pride, GoyimDefenseForce
- **Privatized:** BlackPillCentral, AgainstGayMarriage