Evaluating and Recontextualizing the Social Impacts of Moderating Online Discussions

Qinlan Shen

CMU-LTI-21-007

September 16, 2021

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
www.lti.cs.cmu.edu/

Thesis Committee:

Carolyn P. Rosé, Carnegie Mellon University (Chair)
Yulia Tsvetkov, Carnegie Mellon University
Geoff Kaufman, Carnegie Mellon University
David Jurgens, University of Michigan
Cliff Lampe, University of Michigan

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Content warning: *This thesis discusses issues regarding the moderation offensive language in online spaces and thus contains examples of abusive language and reflection about the social impacts of moderation on sensitive online political discussions. Reader discretion is advised.*

Keywords: abusive language, moderation, online communities, social media, community norms, nlp for social good

Abstract

In response to recent surges in abusive content in online spaces, large social media companies, such as Facebook, Twitter, and Reddit, have been pressured both legally and socially to strengthen their stances against offensive content on their platforms. While the standard practice for addressing abusive content is for human moderators to review whether content is appropriate, the vast scale of online content and psychological toll of abusive material on moderators has led to growing interest in natural language processing in developing technologies to aid in the moderation of offensive language. However, while there has been steady progress on the development of models centered on classifying offensive texts, there is limited consensus over what abusive language is and how NLP models can address practical issues within online moderation. In the complex sociotechnical systems where content moderation takes place, the answers to the questions of "what is abusive language?" and "how should language technologies be used to address abusive language?" can have a major impact on the participation experiences of users in online platforms. Research in online moderation from other disciplines, such as human-computer interaction, platform design, and law, often addresses these social consequences by taking a more interaction-focused view of the problem of moderating abusive language. However, when evaluating moderation issues at scale, these studies of interaction often end up relying on simplified approaches for considering sociolinguistic issues in online communities.

In this thesis, my goal is to bridge the gap between the language-focused view of content moderation from NLP and the interaction-based view from platform design in two directions. In the first direction, I develop and apply more sophisticated language technologies techniques for evaluating the sociolinguistic impacts of moderation strategies at scale. Under this **evaluation** paradigm, I demonstrate the use of NLP techniques in measuring the social impacts of moderation strategies through three case studies over different online communities at different levels of impact. In the first case study, I examine volunteer-based moderation in Big Issues Debate, a political debate community on Ravelry, by investigating how users can perceive reactive moderation as a form of censorship. I introduce a framework for evaluating moderation bias while controlling for user behaviors displayed during the judgment. Using a probabilistic graphical model that accounts for user preferences and transitions between utterances, I then identify intent-based speech acts associated with a high-risk of moderation to examine whether users with minority viewpoints were targeted for moderation decisions. In the second case study, I apply techniques in framing analysis to examine user responses to a platform-wide policy change announcement regarding the expansion of the quarantine feature on Reddit. Through this analysis, I highlighted the ideological nature in how users discuss the tension between content moderation and free speech and the prioritization of different user experience goals with regards to moderation practices on the left and the right. As

a followup to this second case study looking at quarantine policy, in the third case study, I examine the impacts of the quarantines of two major political subreddits, r/the_Donald and r/ChapoTrapHouse. In addition to measuring changes in activity and the use of toxic language within the quarantined and related subreddits, I explore the impact of quarantines on different signals of polarization, engagement, and value association unique to political discussion communities. Based on the findings from these three evaluation case studies, I discuss some of the major social implications of the different moderation strategies used and provide recommendations for additional considerations when designing and evaluating moderation interventions.

The second direction I propose to bridge the gap between language and interaction in abusive language studies is using insights from social theories and the study of online communities to recontextualize how normative and abusive language is defined in language technologies. Under this **contextualization** paradigm, I introduce an examination of how to operationalize descriptive linguistic norm differences across political subcommunities on Reddit. I first present an annotation experiment investigating how experiential factors influence human perception of ideological differences between content from different subreddits. Based on findings from the annotation experiment that the use of specific political associations may distinguish different interpretation patterns by annotators, I then introduce a framework for characterizing common types of assertions and associations made with political entities. Using this framework, I analyze differences in fine-grained assertion usage tendencies between various political subreddits on the left and right. Finally, I evaluate a subreddit embedding model on its ability to capture these differences in tendencies in how different political subreddits use assertion types. Based on these analyses, I reflect on common assumptions within NLP regarding the relationship between labels and normative linguistic behavior and provide future recommendations for taking a more interactional view of language issues in online communities.

Acknowledgments

This thesis is the result of a 6-year long journey, through times of inspiration and excitement and times of frustration and tears. There were moments I was not sure if I would make it here, but I am proud of what I was able to achieve in these six long years. I did not embark on this adventure alone, so I would like to take the time to appreciate people who were able to help me along the way.

First and foremost, I would like to thank my advisor, Carolyn Rosé, for guiding me along the way throughout this journey. She was an excellent mentor and my greatest advocate, and I can never thank her enough for everything she was able to teach me. I will miss our times poring over some odd observation or other from Reddit, trying to make sense of the weird, wonderful world of conversation, which she has given me a newfound appreciation for.

I am deeply grateful to my committee members Yulia Tsvetkov, Geoff Kaufman, David Jurgens, and Cliff Lampe. Thank you for all the invaluable feedback you have given me over the course of this thesis. I always looked forward to hearing the different perspectives and exciting new ideas from each and every one of you whenever we met, and I hope we have the chance to work together again some time in the future.

The work in this thesis could not have been done without the help of various labmates and collaborators over the years. I want to thank all of them for their support, with special thanks to Michael Miller Yoder, Yohan Jo, Sopan Khosla, Jill Fain Lehman, Daniel Clothiaux, and Chris Bogart for directly helping with the planning, annotation, implementation, or writing of parts of this thesis. I also want to give thanks to the members of the Social Computing Reading Group and all the leaders and co-leaders who have helped with running the group along the way (Diyi Yang, Joseph Seering, Tianshi Li, and Pranav Khadpe). It was an honor to lead the group for 2 and a half years, and the discussions we have had over the years have inspired many of the ideas in this thesis.

Beyond just the work in this thesis, I have had a variety of great research mentors over my years at the LTI. Thank you to Dr. Christianne Fellbaum for inspiring a lifelong love of computational linguistics. I would never have been able to accomplish all of this without your faith and support. Thank you to the Descartes team at Google for being so welcoming during my multiple internships there and giving me the opportunity to work on exciting cutting-edge language technologies projects in industry. Special thanks to my host Yinfei Yang for being so patient in putting up with all my questions and helping me make connections throughout the team. I am grateful to other faculty members and scientists at the LTI, namely Chris Dyer, Patrick Littell, Graham Neubig, Lori Levin, and Ashique KhudaBukhsh, who I have had the honor of working together with on other major NLP projects during my Ph.D.

I could not have completed this thesis without the help of my friends, both old and new, who kept me sane throughout this entire process. Thank you for all the meals, game nights, watch parties, skating sessions, and emotional support you have offered throughout the years. I cannot possibly list everyone's names here, but I would like to give special appreciation to Daniel Clothiaux (again), Michael Miller Yoder (again), Hyeju Jang, Peppar Cyr, Harrison Lee, Doug Ashley, Jingyou Xu, Elliot Schumacher, and Barb Gahagen.

And last, but certainly not least, thank you to my family. Whether you are on the other side of the world or willing to open your doors to offer a home away from home here in the U.S., I am truly grateful for everything you have taught me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Content warning: *This thesis discusses issues regarding the moderation offensive language in online spaces and thus contains examples of abusive language and reflection about the social impacts of moderation on sensitive online political discussions. Reader discretion is advised.*

Since its earliest origins, the Web has been celebrated for its enormous potential for facilitating communication. As early as 1968, J.C.R. Licklider and Robert Taylor predicted in their article "The Computer as a Communication Device" that "In a few years, men will be able to communicate more effectively through a machine than face to face" [139]. With modern foresight, it is not difficult for us to see how. Accessibility to the Internet across the world has been steadily increasing, connecting people from all over the globe In 2018, for example, the ITU reported that over half the world's population is connected to the Internet [93]. The growth of social media platforms, such as Facebook and Twitter, has provided spaces for global communication to take place. Thanks to these platforms, people across the world can communicate directly with each other, meet individuals with a wider range of backgrounds and experiences, and organize into communities with shared interests and values. The growing influence of social media in our lives has had a particularly notable impact in the space of online political discussions. By increasing access to key political figures and other users with different political beliefs and goals, social media platforms have democratized participation in the shaping of political discourse. As a result, online political discussions have had an increasingly influential role in political outcomes over a wide range of social issues.

However, even as modern social media companies provide platforms that enable diverse participation and constructive engagement, they must also contend with increased access to abusive content on their platforms. The openness and ease of participation on social media, coupled with an "online dishinbition effect", where disassociation lowers social inhibitions [202], has led to behaviors such as trolling, cyberbullying, and hate speech to proliferate online. These issues are particularly prevalent in discussions centered around sensitive political or social issues important to protected or marginalized groups. In response to this growing wave of abusive content, social media platforms 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 recent efforts to ban

white nationalist and separatist content [199], while Twitter announced a new policy against "dehumanizing speech" in September 2018 [152] and an initiative to investigate the impacts of deplatforming [127]. Nevertheless, the problem of how to address abusive behaviors in online spaces remains an interdisciplinary challenge.

1.1 The role of language technologies in content moderation

In the field of natural language processing, there is growing interest in developing automated technologies to replace or augment humans in abusive language moderation. While human moderation of abusive content is considered standard practice today, there is a clear incentive for using automated technologies for moderation in order to alleviate the burden on human moderators. Viewing abusive material for the sustained periods of time needed to address the scale of content production online, for example, has been demonstrated to take a heavy psychological toll on human moderators [178]. As a result, there is a strong push in NLP to develop automated moderation technologies that can operate at scale to address issues with abusive language online.

Despite the incentives in deploying language technologies for moderation, there remain unresolved issues with current approaches in NLP for addressing the problem of abusive language. Much of the progress in recent abusive language research in NLP has focused primarily on developing and refining modeling techniques to identify abusive language [125]. However, there is limited reflection within NLP as a field as to whether the techniques being developed can actually address issues of abusive content in complex real-world sociotechnical settings. Abusive language detection systems, for example, often operate on limited or ambiguously defined notions of abuse [112, 212]. What kinds of language should be considered abusive are not easily defined, as abuse can take many forms (e.g. inter-personal hate, identity-based attacks, community norm violations) and interact with stylistically complex linguistic phenomena, such as sarcasm or irony [209]. Abusive language is also not a stable target – whether or not content is abusive is heavily dependent on the social contexts in which it is situated. An exchange that takes place between close friends that might be harmless could be considered abusive when it takes place between strangers with different views. Similarly, a comment considered widely acceptable in one community may be considered deeply inappropriate in another [33]. Finally, the question remains as to how abusive language detection systems should interact with the complex sociotechnical systems found in online communities. Prior research in human-computer interaction and platform design has called into question whether automated, classification-based systems can address or amplify existing issues with transparency, policy management, and power in the moderation of online communities [14, 71, 74].

In this thesis, I argue that with our current understanding of how to define abusive language and how NLP systems can contribute to content moderation, the focus in language technologies on developing abusive language detection systems to *perform* content moderation is premature. The simple act of classifying user-generated content as inappropriate or abusive can impact users' ability to freely express themselves in online spaces and their reputation within community circles. Thus, focusing on abusive language detection without understanding the nature of abusive language or the interventions they can lead to the development and deployment of systems that can unintentionally reshape power structures or reinforce problematic biases. Instead, I argue

that NLP can help address abusive online content by building tools that can help us better understand these issues. By developing methods that formalize more nuanced linguistic phenomena, NLP can serve a role in helping to evaluate the social impacts of moderation strategies at scale, while being more conscious of critical issues in both language and interaction when considering abusive language moderation.

1.2 Bridging the gap between language and interaction

In order to situate our view of the role of NLP in the interdisciplinary space of content moderation research, I introduce the concept of levels of impact. These levels of impact primarily correspond to different points where moderation interventions can take place but also provide a scaffold for describing how different disciplines focus on different issues when thinking about moderation strategies and interventions.

Language technologies as a field primarily views the issue of addressing abusive language at the level of *content*, an individual artifact where some form of abusive behavior can occur. From the perspective of intervention strategies, moderating at the content level appears straightforward; if a post or comment contains abusive language, a moderator can directly respond to the offending piece of content, usually through some form of deleting, hiding, or editing. Under a primarily content-based view, it is easy to imagine how an automated abusive language detection system could fit into a moderation strategy – given some texts, we could build a model to determine whether a specific text artifact should be moderated for abusive or offensive content. As a result, the primary concern of considering interventions from this level of impact is understanding the linguistic features of abusive language and building models that capture distinctions between abusive and benign content. I will refer to this view and its concerns as the language-oriented view of abusive language moderation.

However, in online platforms, the content produced by users is situated in complex sociotechnical systems of interaction. Content is not the only level where moderation interventions can have an impact, as platforms provide many different modes of interaction in which potential abuse can occur. As a result, in contrast to abusive language research in NLP, much of the research centered on content moderation in fields such as HCI, platform design, and law are focused on effects within higher-order, interactional levels of impact. Some examples of other levels of impact commonly examined under this more interaction-oriented view of moderation are listed below:

- **Platform:** This level considers interventions and impacts from the perspective of an entire self-contained platform, such as Facebook, Twitter, Reddit, Tumblr, etc. as a whole. A key example of an intervention studied at this level is the controversial December 2018 ban of adult content on Tumblr [53, 138, 171].
- **Community:** Social media sites provide a space for users with similar interests or values to gather or form subcommunities under the broader umbrella of the platform. In many cases, these subcommunities are an explicit affordance provided by the platform, such as groups on Facebook [167] and subreddits on Reddit. However, on platforms where subcommunities are not explicitly defined, subcommunities may also form around other methods of interaction provided by the site. Examples of the second, more implicit form

of subcommunities include mutual follow networks on Twitter [79, 157] and fandom communities on Tumblr centered on tags and follows [85]. For the purposes of this thesis, I will primarily use the term *community* to refer to explicit communities, as top-down interventions on implicit subcommunities tend to be rare.[1] An example of an intervention examined at this level is the 2015 Reddit ban of the subreddits r/FatPeopleHate and r/Coontown [32].

- **Coalition:** Within subcommunities, users may form particular attachments with other users based on shared interests and identities. Coalitions are considered to be subgroups of individuals within a community that share interests that are differentiated in some way from those of other groups or users participating within the same community.
- **User:** Individual users participating on the platform. User-level interventions are among the most common forms of moderation, as users are commonly suspended and banned in platforms for their abusive acts. Kiesler et al. [121], in their chapter on behavior regulation for example, focused on discussing design principles for incentivizing users to comply with community norms.

While interventions tend to be applied at to a specific level of impact, the overarching effects of a particular moderation strategy are not limited to the immediate level in which is applied. Due to the interconnectedness of online communities and the interactional nature of discussion, there are often secondary consequences and impacts of a given intervention at other levels of impact. For example, the practice of deleting posts containing abusive language is an example of a content-level intervention, with the immediate effect of removing a text artifact. At first glance, this process seems entirely contained within the language-focused view of NLP for how to manage abusive content. However, if texts expressing certain viewpoints or discontent are consistently moderated, the coalition of users who hold those specific views may no longer feel welcome on a platform or community and choose to leave. Thus, in order for language technologies to successfully contribute to efforts towards moderating abusive language, we need to consider the broader implications of content moderation strategies beyond the level of the text. On the other hand, language technologies can also contribute towards the more interaction-focused views of moderation by developing more sophisticated models of the normative and abusive language issues that may contribute to content being moderated. Having a more accurate representation of linguistic phenomena within online platforms, as well as being more aware of limitations of these language technologies, will allow us to have a more accurate and descriptive view of the social impacts of moderation strategies at scale.

In this thesis, my goal is to bridge the gap between the language-focused view of abusive language moderation from NLP and the interaction-focused view abusive language moderation from platform design in both directions. I contribute to the more interactional view of moderation by introducing and applying more sophisticated conceptions of language to aid in the **evaluation** of social impacts of moderation strategies using quantitative methods at scale. On the other hand, I contribute to the more language-focused view of offensive language moderation by leveraging insights from social theories and the study of online communities to better **contextualize** what

[1]Bottom-up interventions, such as chain-blocking or tag muting may occur in implicit subcommunities. Due to their user-driven nature, however, these strategies are beyond the scope of the analyses defined in the thesis, which are more concerned with a platforms and policies view of moderation [190]

linguistic variation looks like in online spaces. I focus the analyses in this thesis on language issues related to moderating online political discussion, as the moderation of political discourse has some unique interactional tensions, while also providing domain constraints on the types of normative and abusive language we expect to see. In the next section, I will describe the organization of this thesis and how different chapters fall under this framework.

1.3 Structure of this thesis

In Chapter 2, I begin by describing what is meant by "abusive language" for the purposes of this thesis. I then highlight work in NLP and HCI related to abusive language moderation to provide context for the recent advances and limitations of both disciplines in addressing abusive language online.

Chapters 3-6 in this thesis describe projects falling under one of two goals for bridging the gap between content and interaction in moderation research:

- **Evaluation:** The primary goal of these chapters is to develop and apply NLP techniques in order to derive more detailed insights into the social impacts of a particular moderation intervention or strategy at scale. I use these techniques in three case studies of moderation interventions or events operating at different levels of impact. **Chapter 3** uses a model that considers individual speaker preferences and common transition patterns between sentences and users in order to discover intent-based speech acts associated with norm violations in Big Issues Debate (BID), a political debate community on Ravelry. These speech acts were used to control for user behaviors in a regression-based framework for measuring moderation bias to examine whether there was evidence of unfair judgments against users holding certain viewpoints and history with the moderation team on BID. **Chapter 4** draws upon techniques from framing analyses to examine how users with different ideological viewpoints emphasize different aspects of policy when discussing their experiences and opinions. Applied to user responses to an announcement on the expansion of the quarantine policy on Reddit, I analyze how users on the left or right highlight different priorities with regards to issues in content moderation from a user experience standpoint. **Chapter 5** examines the impacts of subreddit quarantines, a soft, community-level intervention, on two ideologically distinct political subreddits, r/The_Donald and r/ChapoTrapHouse, their users, and the broader space of political subcommunities on Reddit. In addition to evaluating the impact of quarantines on activity levels and the use of toxic language, I analyze whether the quarantines had an effect on issues of polarization and engagement unique to participation in political discussion communities.
- **Contextualization:** The primary goal of contextualization is to re-examine assumptions about how language technologies defines labels and boundaries for sociolinguistic phenomena. Using insights from social theories and studies of online communities, I take a more interactional view of language issues related to normative and violating behaviors in online political communities. **Chapter 6** presents an examination of how to define and measure differences in linguistic norms between political communities. I first present an annotation study examining how experiential factors influence how people perceive ideo-

Project Description	Chapter	Goal	Data
Perceptions of moderator bias	3	Evaluation (user, coalition)	Ravelry Big Issues Debate (BID)
Ideological framing of policy	4	Evaluation (platform)	Reddit quarantine r/announcements thread
Soft moderation, polarization, and community impacts	5	Evaluation (community)	r/The_Donald, r/ChapoTraphouse and related subreddits
Evaluating differences in political community norms	6	Contextualization	Political subreddits

Table 1.1: Chapter summaries

logical differences between content from different political subreddits. Based on insights from the annotation analysis, I introduce a framework for categorizing common forms of assertions made in relation to political entities. This framework is then used to examine fine-grained linguistic differences between political subreddits and evaluate whether an embedding-based model designed to capture subreddit-level linguistic tendencies can detect more subtle community distinctions.

Table 1.1 compares and summarizes the analyses in Chapters 3-6. Finally, in **Chapter 7**, I summarize the contributions of this thesis and discuss the implications of my findings for both language technologies and platform design. The question of how language technologies can help to address abusive behaviors in online platforms remains open. However, insights from my work can contribute to the design and evaluation new intervention strategies, as well as the development of models of language more attuned to the complex interactional dynamics within online communities. Thus, I conclude by discussing future directions in both language technologies and platform design for helping mitigate abusive language in online spaces. Synthesizing common themes from the results of the evaluation case studies, I present suggestions for considering alternatives to common reactive and deplatforming centered moderation interventions. I then reflect on common assumptions made in computational social science research in NLP and provide recommendations for taking a more interactional view of the relationship between language, norms, and identity in online spaces.

Chapter 2

Background

In this chapter, I first give an overview of what is meant by abusive language in this thesis and what are the main issues we are concerned with regarding the regulation of abusive language in online platforms. I then describe research in both NLP and HCI related to abusive language moderation in order to provide context for the recent advances and limitations of work in those fields.

2.1 Defining abusive language

The question of "what is abusive language" is surprisingly hard to answer. Within NLP itself, several efforts have been made towards defining and organizing different types of harms stemming from social interaction. There, however, remains little consensus on the terms and definitions of concepts related to abuse. Schmidt and Wiegand [187], for example, in their early survey of abusive language detection in NLP, define "hate speech" as communication that disparages a individual or group based on their identity, then use "hate speech" as a broader umbrella term for user-created insulting content. Davidson et al. [45], on the other hand, distinguish the use of the term "hate speech" in relation to other milder forms of offensive language that are not targeted towards specific groups. Waseem et al. [212] refers to the broad umbrella of harms as "abusive language", then presents a central 4-way typology for categorizing differences between abusive language subtasks, such as cyberbullying and hate speech, based on explicitness and targeting. Some prior work reflecting on definitions in the space of abusive language focuses on categorizing aspects of harm specific to certain subtypes of abuse. Van Hee et al. [208], for example, presented an annotation scheme for on cyberbullying, which incorporates dimensions related to severity, role, and intent commonly found in abuse targeted towards individuals. Anzovino et al. [4] used their analysis of misogynistic tweets to develop a typology of forms of misogynistic abuse, such as objectification and sexual harassment. Breitfeller et al. [19], in their exploration of microaggressions, categorized different forms of microaggressions based on themes and intents expressed across various demographic groups. Overall, however, these efforts at defining and categorizing forms of abuse have highlighted the wide variety of definitions and tasks falling under the umbrella of abusive language studies within NLP.

In this thesis, I use a conceptualization of abusive language centered around issues covered

by prescriptive norms and content guidelines across different online communities. Most similar to my concept of abuse is the unified taxonomy from Banko et al. [8], who synthesized community guidelines across different online platforms and recommendations from civil society organizations in their definition of abuse. Under this conceptualization of abuse, I define *abuse* as behaviors that violate the norms or standards of a community or platform in a way that compromises the trust and safety of participants in that space. Taking a community norm and policy-oriented approach to defining abuse provides some advantages when reflecting on how language technologies can be used to address issues of abuse in real-world online settings. By centering the definition of what is abusive according to community standards, this conceptualization of abuse more closely mirrors the judgments that moderators operate on than general purpose definitions of abuse that may not be socially relevant for particular online communities. This definition of abuse also allows us to incorporate harms that are commonly considered in violation of community standards but are rarely explored in NLP research, such as doxxing, misinformation, and self-harm. Finally, this policy-centric definition of abuse explicitly allows for differences in community standards for how to define abuse. Using this definition of abuse, however, requires us to focus our analyses on actual cases of norm violation and moderation in the communities of interest. As a result, in this thesis, the bulk of the language-centered methodology focuses on the operationalization of linguistic norms, descriptive and injunctive, for the specific communities I am interested in studying.

2.2 Abusive language research in NLP

While most of the work in NLP is not centered around our conceptualization of abuse, recent work in NLP has highlighted a variety of interesting areas of research in regards to concerns with building technologies for detecting some definition of abuse. In this section, I cover some of the major areas of study in recent abusive language research in NLP to provide context for the current limitations of language technologies for detecting abusive language.

2.2.1 Dataset construction

Under the predominant supervised classification paradigm in abusive language detection, machine learning models trained to detect abuse rely on large datasets annotated with abusive language labels. While traditionally, these datasets are usually constructed from sampled content, then labeled through crowdsourced annotation efforts, recent work has highlighted some potential concerns with this approach for constructing abusive language datasets.

One challenge in the construction of datasets for abusive language is finding instances of abusive language in the wild. Benign content tends to greatly outnumber abusive content when randomly sampling posts [187], and abusive content is often removed from platforms when it is moderated [31]. The most common approach for addressing the skew between abusive and benign content in online communities, then, is instead of randomly sampling content from a community, we can query for content from contexts more likely to be associated with abuse. Two of the most commonly used datasets in abusive language detection that use this approach are Waseem and Hovy [211] and Davidson et al. [45]. Both datasets were constructed from content

sampled from Twitter based on querying seed words commonly associated with abuse. However, later work has highlighted some of the flaws and disadvantages of using these focused querying approaches. Arango et al. [5], for example, highlighted that querying using keywords can bias datasets towards a small number of users who consistently use hateful keywords, while Wiegand et al. [215] noted a similar effect in biasing abusive content towards certain topics. Vidgen and Derczynski [209] also observed that the selection process for keywords used in focused queries tended to bias sampling of abusive content towards content from far-right communities. This could, in turn, potentially bias trained abusive language models against markers of right-wing discourse. Nevertheless, due to the natural imbalance between abusive and benign content online, focused querying is arguably the necessary standard for obtaining abusive language data.

After obtaining content, the annotation of the collected abusive language data presents additional challenges for dataset construction. Ross et al. [183], for example, demonstrated that while presenting annotators with a definition of hate speech improved annotator alignment with the defined standard, overall reliability in hate speech annotation remained low. Waseem [210] noted differences in how expert annotators and amateur annotators labeled hate speech, with amateurs being more likely to interpret comments as hate speech compared to experts. These findings suggest that the task of annotating hate speech may be inherently ambiguous and unreliable, due to individual differences in how hate speech is perceived, even with instructional guidance. In addition to the problem of unreliable annotation agreement, Castelle [28] calls into question the reliance on third-party annotation for abusive language datasets itself. Comparing annotation agreement and model performance on abusive language datasets where labels were obtained from either third-party annotators or actual moderation outcomes, Castelle found that labels generated by third-party annotators were artificially easy for deep learning classifiers to learn. This is likely the result of annotators relying on surface-level cues to make abusiveness judgments, due to being decontextualized from the actual community in which the text took place in. As a result of the findings from these different annotation studies, there is growing interest in NLP research in figuring out ways to incorporate additional contextual information when annotating abusive language. de Gibert et al. [48], for example, introduced an annotation scheme that allows annotators to extend and mark the visible context around a text when judgments may be unclear from the text itself.

2.2.2 Modeling abusive language

The predominant paradigm in abusive language modeling approaches is based on classification – given a text, determine whether that text is abusive under some categorization of abusive and benign language [45, 133, 137, 211]. However, recent work in the modeling of abusive language has proposed other formulations of the problem beyond a single classification decision. A common approach for incorporating more nuance into the judgment of whether or not a text is abusive is to include the prediction of other information related to abusiveness. For example, recent shared tasks regarding offensive language detection now include subtasks, such as identifying the target or severity of abuse, in addition to the main judgment of offense [9, 222]. Sap et al. [186] similarly introduced social bias frames as an explanatory framework for capturing and describing the power implications behind why more implicit expressions may be considered abusive. In another approach for investigating the potential sociolinguistic mechanisms behind

abuse, Carton et al. [27] use an adversarial neural model to generate extractive explanations for why texts were identified as personal attacks.

While making judgments based on the text itself is still the predominant approach in abusive language detection, similar to the observations in dataset construction, one growing area of exploration in modeling is the role of contextual features outside of the text being judged. Gao and Huang [66] presented early work showing that the inclusion of context features, such as username and title of the article a comment responds to, improved the performance of a simple logistic regression abusive language classifier beyond using just the comment itself. More recent work on the role of context in abusive language detection has explored complete reformulations of the abusive language judgment problem that incorporate additional social and/or conversational context. For example, *escalation detection*, or predicting whether a conversation will become abusive in the future based on markers in early conversation, has been introduced as a more proactive version of the traditional abusive language classification task. Zhang et al. [223] employed techniques from causal inference to examine how preceding context can be used to predict whether conversational threads will eventually go awry on Wikipedia talk pages. In a followup paper, Zhang et al. [224] used a similar approach to predict future anti-social behaviors in Facebook conversations. Hessel and Lee [83] incorporated features from the tree structure of early comments to predict the future controversiality of posts in Reddit, finding that early network-based features were less brittle to domain shifts than text features. The true impact of modeling context for abusive language detection, however, is still unknown, due to the limited public availability of context-aware abusive language datasets [168].

2.2.3 Bias and fairness

Aligned with growing concerns over bias in machine learning models, there is a growing body of research within NLP highlighting biases in abusive language detection datasets and models. Issues of bias and power are particularly important to address in abusive language detection models, as marginalized groups often find themselves the target of online abuse. As a result, biases in abusive language detection systems will have direct impact on the ability of marginalized groups to both safely and freely participate in online spaces Due to the intersection of identity and abuse in hate speech, however, many of the existing abusive language detection datasets and systems often contain biases against demographic groups commonly targeted by online harassment. Park et al. [166], for example, found evidence of gender bias against female identity words in an abusive language dataset, due to these words being correlated with content that was labeled as "sexist", with few negative samples. Sap et al. [185] and Davidson et al. [46] found similar biases against markers of African American English in several Twitter abusive language datasets, which are then replicated and propagated by models trained on these datasets. Kim et al. [122] presented an analysis of intersectional bias in hate speech datasets, finding that tweets by African American men were much more likely to be considered hateful than those of other intersectional groups. Work on how to mitigate these biases without compromising performance, however, remains in its early stages [166, 219].

2.3 Content moderation research in HCI

In this section, I describe recent work in HCI investigating issues of content moderation. While there are existing frameworks for considering different perspectives in content moderation within HCI [190], I divided work in HCI into themes broadly based on methodology in order to discuss insights and methodological advantages of thinking about moderation from multiple perspectives: (1) qualitative and mixed-methods approaches for examining interactional and experiential issues in moderation, (2) large-scale quantitative approaches for evaluating moderation impacts, and (3) the design of tools and strategies for aiding moderators.

2.3.1 Qualitative/mixed-methods studies of moderation

Much of the work examining governance issues in moderation from HCI is centered around qualitative and mixed-methods studies that utilize interviews, surveys, and content analysis to derive insights about the functional and interactional experiences of participating in moderated spaces. These studies can highlight the experiences within existing systems of moderation from specific participatory perspectives. Some studies, for example, have drawn attention to key limitations with existing moderation practices from the perspective of moderators of online communities themselves. Seering et al. [192] interviewed 56 moderators across three different platforms, identifying three key processes through which moderation can affect community development – moderator recruitment and development, handling of misbehavior, and rule-setting. Jhaver et al. [100] similarly conducted semi-structured interviews with 16 Reddit moderators and participated in moderation activities themselves to gain more insight into how Reddit Automod, a configurable automatic moderation program, is integrated into the moderation workflow. The study provided many key findings regarding the experience of coordinating with automatic moderation tools, such as the relative simplicity of Automod being valued by moderators due to easy configurability and the desire for more support/resources for developing, sharing, and evaluating automatic systems. Dosono and Semaan [51] conducted semi-structured interviews with 21 moderators from Asian American and Pacific Islander (AAPI) subreddits on how they handle the emotional labor of moderating conversations about identity. Guided by these interviews, the authors discussed issues of expectation, safety, and burnout when engaging in moderation of personally emotional content, highlighting the need for social and emotional support in specific moderation contexts. While not focused on moderator experiences directly, Fiesler et al. [60] investigate the process of rule-setting by moderators by characterizing types of rules on Reddit and tracing the origin of rules across the broader subreddit ecosystem.

Other studies under this paradigm have focused on moderation from a user experiences perspective. Blackwell et al. [14], for example, examined the experiences of users on HeartMob, a private platform providing support to victims of harassment. Through semi-structured interviews with 18 HeartMob users, the authors discussed the role of classification-based approaches in moderation in both the maintenance of social norms but also the reification of power structures and oppression. From this study, the authors argue for a need for more democratic, user-driven practices for mitigating harassment, as well as systems for supporting the experiences of marginalized and vulnerable users. Myers West [158], on the other hand, ran a survey of users who had experienced a content takedown to investigate folk theories and impacts of moderation

in social media. Survey responses highlighted frustrations with transparency and responsiveness in moderation, which led to users forming folk theories blaming human intervention and biases for having their content removed.

2.3.2 Quantitative studies of moderation impacts

While the qualitative analyses discussed in the previous section highlighted key issues with moderation in online spaces, questions often remain whether the insights from these analyses hold over the wide variety of communities and contexts moderation can take place in. Broad quantitative analyses of moderation, however, are comparatively less common and generally centered around examining the impact of moderation strategies at scale. Chandrasekharan et al. [32] investigated the impact of a 2015 Reddit ban of the subreddits r/FatPeopleHate and r/Coontown on the users who participated within those communities, providing early evidence on the efficacy of community bans for limiting the spread of hate speech across the broader Reddit platform. Jhaver et al. [101] analyzed the usage of flairs and comments for providing content removal explanations on Reddit, finding some broad evidence that providing explanations may help users adhere to community norms in the future. Chancellor et al. [30] examined how pro-ED communities on Instagram skirted moderation restrictions by measuring the adoption and evolution of lexical variations of moderated tags over time. Garland et al. [68] presented a longitudinal study of hate speech and counter-speech on German Twitter, finding that both organized hate speech and counter-speech efforts were able to steer public discourse in particular directions. The main advantage of these large-scale analyses of moderation is their ability to broadly measure the impacts of different intervention strategies across entire communities and platforms. As such, these studies often used to explore more general trends with moderation that can be shared across multiple different social media communities and may provide insight into the overall health of communities and platforms from a more centralized perspective.

While quantitative studies of moderation interventions may capture the broader impacts of interventions across a platform, these studies often rely on simplified operationalizations of language as metrics for evaluating sociolinguistic effects. Chandrasekharan et al. [32], for example, used a count and lexicon-based approach as a measure for the use of abusive language related to the original banned subreddits. Although this approach is simple to implement and scales easily across large amounts of data, it limits analyses of toxic language within communities to obvious surface-form realizations of abuse, missing out on more implicit, pragmatically complex forms of hate. Similarly, while Jhaver et al. [101] used topic analysis to discover common themes in content removal explanations, their analyses of the impacts of removal explanations did not end up exploring how these different themes contributed to how users learned community norms. One approach for getting at interactional phenomena at scale without relying on more complex NLP tools and techniques is using simpler quantitative techniques to find broad patterns in large-scale data, then performing qualitative analysis over the discovered patterns. Juneja et al. [111], for example, used a topic modeling based approach to find meta-communities with similar norms on Reddit, before performing qualitative analysis within each meta-community to associate norm violations with known rules. Nevertheless, recent developments in NLP research could potentially aid in finding more robust, contextually aware linguistic patterns across large-scale social media data. These techniques, however, need to be able to scale and adapt to large social me-

dia data and be easy to integrate with the statistical methods used in these types of quantitative analyses.

2.3.3 Designing for moderation

Some work in HCI chooses to focus on developing tools to directly aid in moderation, instead of simply analyzing the moderation experience. Mahar et al. [144], for example, explored the idea of *friendsourced moderation* by presenting Squadbox, a tool to help the recipients of email harassment recruit friends to shield them from future attacks. Based on insights from interviews with victims of harassment, the authors designed Squadbox to allow for a variety of moderation activities to be taken on by friends of an email recipient, such as approval, sorting, and summarization, with some machine learning functionality provided through Perspective API. While focused more on user-centric curation of emails due to the variability of platform APIs and concerns over user trust and privacy, an extension of Squadbox with additional automatic filtering and summarization capabilities may be potentially useful for moderator coordination within online communities.

In another example of a tool using machine learning designed to aid in moderation, Chandrasekharan et al. [34] developed Crossmod, a machine learning based classification system that provides recommendations by ensembling moderation decisions from FastText classifiers [110] trained over 100 different subreddits. Unlike many of the abusive language detection systems developed within NLP, Crossmod was built based on moderator testimonies about gaps with current automated moderation systems on Reddit and evaluated through an iterative two-phase deployment process. Crossmod also does not rely on the assumption of one general purpose definition of abuse over communities but rather utilizes cross-community learning to derive insights from existing moderator decisions across Reddit. While the authors found that Crossmod recommendations aligned with moderator intuitions for what should be removed from a subreddit, most of the comments marked for removal by Crossmod remained on the site, however, suggesting a gap between the scope of violating content on Reddit and the current systems in place for moderation.

Work on designing for moderation incorporates aspects of both the qualitative and quantitative studies discussed in this section. In order to know what types of systems may provide actual utility to real-world moderation, design work relies on interviews with actual moderators to learn about the specific limitations and desired capabilities within existing governance structures. On the other hand, the integration of machine learning tools within the designed systems often relies on models that learn from large amounts of data and thus, requires awareness of the simplifying assumptions of made by quantitative approaches. As a result, research in design can be advanced by insights from either qualitative or quantitative studies of moderation, while also potentially providing tools and systems with direct real-world impact. Despite this, there are often still limitations with design research in terms of adoption and integration of technologies into existing sociotechnical systems.

Chapter 3

Perceptions of bias and censorship

3.1 Introduction

Online discussion forums create a platform for communities with similar interests to share thoughts and debate issues. However, the technological facilitation of conversation on these forums does not ensure that high-quality deliberation takes place. Discussion forums are vulnerable to problems such as trolling, flaming, and other types of nonconstructive content [170]. Furthermore, when the topic of conversation is controversial, such as discussions centered on religion, politics, and other social issues, discussions can become toxic or inflammatory. Perceived anonymity in online spaces often exacerbates this problem by weakening self-censorship, as individuals are less likely to regulate their own behavior under the belief that it is difficult to trace back what they say to meaningful consequences [29, 47].

To address issues surrounding toxic or offensive language, online political discussion forums often rely on moderators to enforce rules and boundaries for how users behave and what they can say. However, the line between legitimate forms of regulation, which are used to discourage behavior defined as inappropriate, and *illegitimate censorship*, where particular individuals, opinions, or forms of communication are unfairly suppressed, is often difficult to define [217]. Censorship is usually defined subjectively, and in cases where there is room for interpretation, the unconscious biases of regulators may affect their judgments. On the other hand, a user's own bias may lead them to perceive unfair treatment where there is none. This perception can be compounded by a power differential between regular users and users with moderation privileges.

In this work, we contribute new insight into the differences between perceived and actual bias in an online community's attempt to facilitate productive exchange on controversial issues. Fair moderation without illegitimate censorship is fundamental for creating safe, engaging online spaces for deliberation on controversial topics [26]. Research in this area not only can improve the quality of discussion in online political forums but also can allow insight into the process of developing norms of behavior and effective moderation in online communities. Regardless of whether censorship actually takes place, the perception of illegitimate censorship itself can create an atmosphere where users feel unfairly treated and trust in the forum is undermined [217]. Thus, it is important to understand the sources of perceived censorship and recognize when and how perceived censorship is actually manifested.

Guided by these issues, we explore the following research questions:

(1) Do moderators unfairly target users with specific viewpoints? If so, to what degree?

(2) What are possible sources of bias that could lead moderators to censor unfairly?

(3) What are possible causes for users' perceptions of moderator bias?

To address these questions, we examined the perception of moderation bias against users with unpopular viewpoints in the Big Issues Debate forum on Ravelry. Using a probabilistic graphical model to identify speech acts, we identified high-risk behaviors associated with rule-breaking, then examined the effect of viewpoint on the likelihood of moderation, controlling for high-risk behavior. This allows us to investigate whether users with minority viewpoints are being unfairly moderated, given the behaviors they exhibit. We find that moderators are significantly more likely to moderate posts from users that hold unpopular viewpoints, though the effect size of this bias is small. We find a similar effect for users who have been recently moderated, to a smaller degree. While this supports the perception of minority-view users that the moderation is unfair, we argue that the perception of bias within the group is an issue by itself, as the perception of illegitimate censorship can lead to tension between the moderators and users within a community.

The rest of the chapter is organized as follows. (1) We review prior work on the relationship between moderation and censorship in political discussion. (2) We describe the Big Issues Debate forum and its main characteristics. (3) We present our method for measuring moderator bias that takes into account user behavior. (4) We examine to what extent users' perceptions of moderator bias against minority viewpoints are supported by our findings. (5) We discuss the implications of our findings.

3.2 Moderation issues in political discussion

Moderators play an important role in many online forums by helping to maintain order and facilitate discussion within their community [126, 141]. While conventional wisdom suggests that moderators positively influence the quality of discussion in forums [90], the role of a moderator is often diverse [145], unclear [217], or emergent [91] across different communities. Thus, it is important to consider how moderators operate within the context of the community that they are trying to maintain. In online political forums, moderators are considered critical in ensuring quality discussions by creating and enforcing regulations for proper behavior [52], as useful debates require that participants maintain order, respect, and civility towards each other [26, 216].

However, when these political discussions are facilitated by interested groups, moderation can quickly be labeled as censorship. These claims are common on online political forums administered by national governments, a focus of research on the potential for new forms of deliberative democracy [119, 218]. Wright [217] reviews the process for moderation in two of the UK government's online political discussion forums. They find that moderation must be done carefully to avoid the "shadow of control", the perception that some entity of power can control what is said [52]. In the ideal situation, rules for censorship must be detailed, openly available, and enforced by an independent party [217], while still explicitly facilitating the goals of the forum.

In non-governmental political discussion forums, the concept of a "shadow of control" is less

obvious, as these forums are commonly run by volunteer moderators, rather than an explicit centralized entity with particular goals. Nevertheless, unconscious cognitive biases may arise from the structural organization of political discussion forums and from cognitive tendencies. Bazerman et al. [11], in their investigation into why accountants make biased decisions, noted that ambiguity in interpreting information gave accountants the room to make self-serving decisions. In the context of political discussions, ambiguity in the rules for how to engage appropriately in a debate may allow moderators to make unfair decisions against particularly troublesome users or viewpoints they disagree with. Another (more surprising) condition that often promotes unconscious cognitive biases is the belief in one's personal impartiality [113]. While moderators are expected to act impartially, as they are often removed from debate, they may unconsciously make more biased decisions because they are primed to believe that they are genuinely impartial, instead of recognizing these biases.

In the following section, we describe our platform of study, the Big Issues Debate group on Ravelry, and discuss the organizational elements that make it an ideal platform of studying moderation biases.

3.3 Ravelry and Big Issues Debate

Ravelry is a free social networking site for people interested in the fiber arts, such as knitting, crocheting, weaving, and spinning. With over 7.5 million users in December 2017,[1] Ravelry is one of the largest active online communities that has been relatively understudied. While the broader Ravelry community is primarily focused on the fiber arts, social participation on Ravelry centers around tens of thousands of user-created and -moderated subcommunities, called *groups*. Groups act as discussion boards centered on a certain theme. Any user on Ravelry can create a group covering any variety of topics, special interests, or identities, which may or may not be related to the fiber arts. For example, *Men Who Knit* provides a space for men, an underrepresented group in the fiber arts, while *Remrants* allows users to post rants about nearly any aspect of their lives.

3.3.1 Big Issues Debate (BID)

Our study focuses on the Big Issues Debate group on Ravelry. Big Issues Debate, commonly referred to as BID, is described as a space

> *... for everyone who likes to talk about big issues: religion, politics, gender, or anything that is bound to start a debate.*

Receiving over 3,500 posts a month, BID is the largest group dedicated to political and social issues and one of the most active groups overall on Ravelry (in January 2018).[2]

Debates on BID begin with a user creating a thread and posting their view on an issue. Other users post responses to the original user's post or to other posts in the thread. An example BID post is given in Figure 3.1. Every post in the thread, including the original post, has a set of

[1] `https://www.ravelry.com/statistics/users`
[2] `https://www.ravelry.com/groups/search#sort=active`

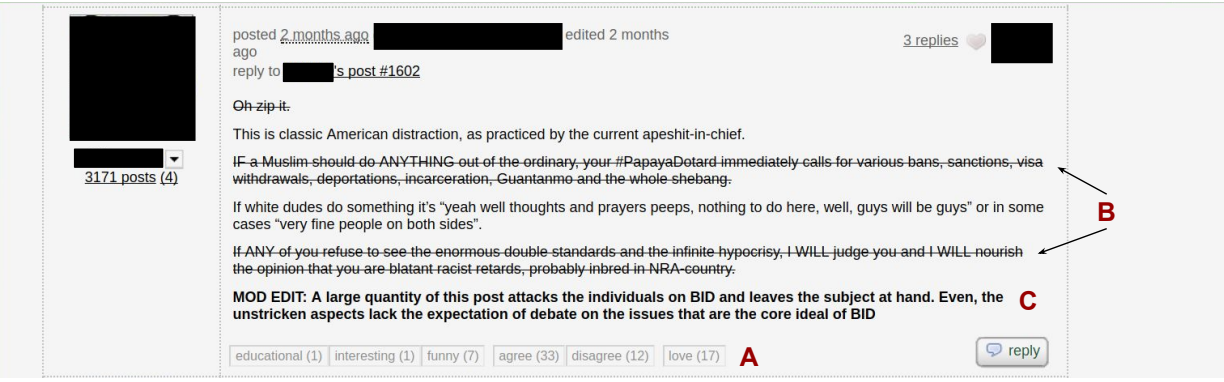


Figure 3.1: Example of a BID post that was also moderated. (A) shows the *tags* associated with the post. The text of the post that was crossed out (B) was not crossed out by the original poster but by the moderators after judging the text as a violation of the rules of BID. (C) gives the moderators' reasoning for how the post violates the rules of BID. Note that although the post was moderated, more users in the group *agree* with the post than *disagree*.

six associated *tags* (Figure 3.1, A) that users can interact with: *educational*, *interesting*, *funny*, *agree*, *disagree*, and *love*. Clicking on one of the tags allows a user to anonymously increase the value of a particular tag once per post, though these values do not affect the order in which posts are displayed.

There are three officially recognized and regulated formats of debate on BID: *Order* (default debate format), *Rigor* (stronger standards for sourcing/citations), and *BID* (discussion about policies and practices on BID). Thread creators can choose which format they want their debate to be in by tagging it in the thread title (e.g. "ORDER - Media Responsibility in Politics", "RIGOR: Bigotry and the 2016 US presidential race"). If not tagged, the thread is assumed to be in the Order format. In all of the recognized formats on BID, users are expected to follow these rules:

1. Abide by Ravelry's Community Guidelines and Terms of Service.
2. No personal attacks.
3. Behave civilly.
4. Debate the topic, not the person.
5. Do not bring in other groups, users not participating in the debate or baggage from one thread to another thread.
6. Don't derail the thread.

Within a discussion thread, users can flag another user as being in violation of one of the 6 main rules. Whether or not a post is flagged is only public to the moderation team, the user who made the flag, and the user who received the flag. Moderators then judge whether flagged posts are in violation of the BID rules. If the post is judged to be in violation of the rules, it is hereinafter referred to as *moderated*. In almost all cases, moderated posts are kept visible, but the offending part of the post is crossed out with a strikethrough (Figure 3.1, B). Moderators are also expected to give reasons for why a post was moderated (Figure 3.1, C), though they do not post their username. Users who repeatedly make offensive posts may have posting privileges

suspended for a period of 24 hours or banned from the group for a longer period of time based on severity of the offense. Moderators may also delete posts, but this is only practiced in the Ask the Mods thread (where only specific types of posts are allowed) or in cases of "extreme spam".[3] BIDs transparent moderation system highlights its surprisingly participatory system of governance. Moderators in BID, for example, are drawn from its general user base, and at one point in its history, moderation was performed by a rotating panel of volunteers, who played the role of a temporary moderator for a week at a time. Users often also commonly engage in meta-discussion with moderators about their judgments through specific threads, such as Ask the Mods, and ordinary users can vote to make changes to the existing moderation structures. To maintain the boundary between ordinary users and moderators, however, one key limitation placed on moderators is that moderators cannot participate in debate threads they moderate. The goal of this limitation is to prevent moderators from explicitly making decisions against users they are debating. Despite this rule, moderators are not required to deanonymize themselves when making a moderation decision.

3.3.2 Issues with moderation

BID provides an interesting setting for studying perceptions of censorship in political discussions not only because it is an active debate group with formal moderation but also because of its controversial reputation. BID's formal moderation is crucial in creating a space where users with different viewpoints can discuss political and social issues, compared to other Ravelry political discussion groups with less formal moderation, which tend to be more homogeneous. However, BID is infamous in the broader Ravelry community for tension between users and its moderation team, providing an ideal setting for studying frustrations about moderation from perceived bias. Meta-discussion threads also provide insight into user opinions and perceptions about the organization of the group. As an example of frustration with the perceived censorship on BID, one conservative-leaning user comments:[4]

> *Never have I seen such bold faced disregard for other's opinions. Am I surprised? Not with this group of mods ... A sorrier bunch of biased, preachy people with unlimited authority seldom seen ... we don't have a freaking chance of having any of our problems addressed when we're outnumbered 50 to 1 (at the least) - seriously????*

expressing their perception that moderators are biased against conservative users, who are in the minority on BID. A liberal-leaning user, on the other hand, commented

> *The thing we can say for sure is that a lot of conservatives have come forward saying they're not being treated fairly. I don't think that's true, but I wouldn't, would I?*

questioning the perception that conservative users in BID are actually unfairly treated.

Some users argue another view on how moderation in BID is biased, where moderators may be biased against certain individuals based on their past behavior:

> *I think there are users who draw moderation when others wouldn't, and I don't think it has anything to do with political orientation. It's embarrassingly obvious at times.*

[3] `https://www.ravelry.com/groups/big-issues-debate/pages/Information-on-Moderation-for-Members`

[4] All quotes are lightly paraphrased for user privacy

It's not weird for people who have been modded to double down, rationalize their behavior, cast blame on others, or toss a word salad to "explain" why they shouldn't have been modded. The mods' reaction to their being modded is just par for the course.

Users who have been moderated in the past or users who have complained about moderation in the past, for example, may be given less leeway for offenses than someone who has never been moderated, as it may be in the moderators' interests to quickly shut down dissent from high-risk individuals.

The widespread idea that the moderators are biased against certain viewpoints or individuals raises the question of what forms these perceived biases take. We find that users on BID primarily consider "censorship" to be a problem of false negatives in moderation. Most users that have been moderated accept that their behavior is inappropriate under the rules of BID. However, users also argue that if their behavior is considered inappropriate, then many similar posts that have escaped moderation should be moderated as well:

However none of those were struck through / given a "mod edit". This was only done to XXXX. Yep. It isn't biased at all

If my posts were deleted why not XXXs?.

I see certain liberals constantly get away with rule breaking. I don't quite understand why. But they do.

I was also modded for not furthering the discussion. I wonder how many other posts don't further the discussion?

Thus, the primary issue of perceived bias appears to be derived not from direct suppression of a user or viewpoint but from issues in how consistently the rules are applied.

3.3.3 Contrasting views of bias

Based on our examination of the organizational structure of BID, we hypothesize that there is opportunity for moderator bias in deciding whether to moderate a post. The guidelines of BID are ambiguous, using vague statements such as "Behave civilly" and "Debate the topic", which leaves room for interpretation. This ambiguity may allow moderators to make self-serving judgments in favor of users who they agree with. Thus, one hypothesis is that moderators could be biased against certain viewpoints or users. On the other hand, this same ambiguity in the rules could allow users to make the self-serving interpretation that moderators are unfair against them or their viewpoints. This supports the hypothesis that there is little to no actual moderator bias, only a user's strong perception of bias. The goal of our analysis is to test these hypotheses through a series of statistical modeling experiments.

3.4 Method

To assess whether the moderation team is making biased decisions, we present an approach for evaluating moderator decisions alongside users' actual behavior in posts considered for moderation. In order to determine whether or not user viewpoint plays a role in these moderation

decisions, we need some operationalization of a user viewpoint on BID. We also need to identify the behaviors that may put a user at risk of being moderated, as certain users may contribute offensive content more often. If users of a certain group express inappropriate behaviors more often, them receiving more moderation than another group could be considered fair, under the notion that all inappropriate behaviors should be moderated. After operationalizing these relevant variables of viewpoint and behavior, we include them in a binary logistic regression model with odds ratios (OR) to predict whether a given post is moderated. This model allows interpretation of the factors that may increase the likelihood that a post would be moderated; odds ratios allows us to estimate the effect of a variable on the probability that the post is moderated.

3.4.1 Dataset

Post data was scraped from the Big Issues Debate group on Ravelry from the beginning of the group in October 16, 2007 until June 6, 2017, including posts from threads that were publicly archived by the moderators and ignoring posts that were deleted. For each post, we collect its thread number, title, post number, author, date of creation, and the value of its tags on June 6, 2017. We also determined whether the post was moderated. We consider a post to be a candidate moderated post if it contains or was replied to by a post containing the phrases "mod post", "mod edit", or "this post was moderated for", which all signal that a moderator has edited the post for inappropriate behavior. Moderators are expected to cross out the portions of text that were judged to have violated the BID rules, so in almost all cases we can recover the original text of the post that was moderated. We remove the very few candidate posts that do not have any portions that have been crossed out from our dataset, as we cannot ensure that these posts still contain the original behavior that they were moderated for. Our final dataset from BID consists of 350,376 posts by 3,320 users over 4,213 threads.

3.4.2 Model specification

Our model is designed to measure the effect of user viewpoint on the likelihood of being moderated. To control for the alternative theory that users' histories with moderation affect moderation decisions, we include an additional lag variable *mod_prev*, in addition to our main effect variables indicating viewpoint and high-risk behavior.

We also define pairwise interaction terms among our three main effect variables (*high_risk*, *mod_prev*, and *minority*) as an input to the regression to tease apart the relationships between the main effect variables in conjunction with each other. The final set of variables that we use as input to the regression are:

Dependent variable

- *moderated*: A binary variable indicating whether the given post was moderated or not.

variable	1	2	3	VIF
1. mod_prev	1.000			1.00
2. high_risk	0.033	1.000		1.00
3. minority	0.141	0.061	1.000	1.00
Mean VIF				1.02

Table 3.1: Correlation and multi-collinearity checks for main effect variables.

Independent variables

- *mod_prev*: The number of times the user has been moderated in the previous 30 days. We normalize this variable to have a mean of 0 and standard deviation of 1 across all posts in our dataset for rescaling purposes.
- *minority*: A binary variable indicating whether the user who made the post is a minority-view user in BID (see "Assigning Viewpoint" section).
- *high_risk*: A continuous variable indicating whether a post has an unusually large amount of high-risk behaviors (see "Characterizing Behavior in BID Posts" section).
- *high_risk* $\times$ *mod_prev*
- *high_risk* $\times$ *minority*
- *mod_prev* $\times$ *minority*

Correlation and multi-collinearity checks for the main effect variables are found in Table 3.1.

3.4.3 Assigning viewpoint

Assigning viewpoints to posts

In order to determine whether users who hold unpopular views are moderated more, we need to label users with whether or not they tend to hold the same view as the majority of the group. To determine whether a user holds majority or minority views, we use the agree and disagree tags on the posts they have made. The agree and disagree tags on a user's post provide an indication of how closely the post aligns with the views of the general user-base on BID.

The general perception on BID is that right-leaning, conservative users and viewpoints are in the minority while left-leaning, liberal users and viewpoints make up the majority. This pattern generally appears to align with the broader composition of users on Ravelry; while there exist groups dedicated to conservative viewpoints, such as Conservative Knitters and The Bunker, some of the most popular groups on Ravelry, such as Lazy, Stupid, Godless and Ravelry Rubberneckers had historical tensions with these groups over the moderation of content within conservative communities, leading to the eventual removal of The Bunker and revocation of moderation privileges for its users. To verify that the agree and disagree tags align with this liberal-conservative conception of majority-minority on BID, we sampled 20 posts with higher agree than disagree tag values and 20 posts with higher disagree than agree tag values. Posts were sampled across threads to determine the general trend of views on BID on a variety of issues. We

then presented the posts, along with the title of the relevant thread and the preceding post in the reply structure as context, to two native English speakers with moderate political knowledge and asked them to separately determine whether the opinion expressed in a post leaned more towards a liberal viewpoint or a conservative viewpoint. We define *liberal* viewpoints as those that favor social progressivism and government action for equal opportunity and *conservative* viewpoints as those that favor limited government, personal responsibility, and traditional values.

We then treat the agree/disagree tags on the sampled posts as another annotator who rates a post as liberal if the post has a higher agree than disagree tag value and conservative otherwise. Comparing this "agree/disagree" annotator with our human judges, we obtain a Fleiss' kappa of 0.916. This indicates high agreement among the human annotators' judgment of liberal and conservative and the agree/disagree tags associated with the post. Thus, we can aggregate the values of the agree and disagree tags of a particular user across BID to get an overview of their political viewpoint within particular posts.

Assigning viewpoints to users

To label the viewpoint of a particular user, we first find every thread they have participated in on BID. For each thread, we sum the agree tag values for each post the user made in that thread. We repeat the same process for the disagree tag values in the same thread. As threads on BID are intended to be centered on a particular issue of debate (e.g. gun control, immigration, tax reform), the summed agree and disagree tag values should indicate how much the other users on BID agree or disagree with the user on that particular issue. If the total disagree tag value is greater than the total agree tag value for a user on a particular thread, we label that user as having the minority viewpoint on the issue discussed in the thread. This thread-level notion of viewpoint is analogous to the *issue-oriented viewpoint* described in Kelly et al. [118].

However, simply holding a minority view on one thread does not indicate that a user holds the minority viewpoint across BID – users may have particular issues where their viewpoints do not align with the ideological group closest to their general beliefs (e.g. a primarily liberal user who is pro-life). Thus, in order to get a general viewpoint for each user, we compare the number of threads where they hold the majority viewpoint with the number of threads where they hold the minority viewpoint. If the number of threads where they hold the minority viewpoint is greater, we label that user as a *minority-view user*. This notion of viewpoint is analogous to the *ideological viewpoints* described in Kelly et al. [118], which are coherent systems of positions across issues. We focus on ideological viewpoints in our analyses because users participate across threads and recognizably carry their ideological positions with them. This is apparent in BID meta-discussion threads where users will refer to each other with ideological labels (e.g. "conservative", "liberal"). Thus, we predict that moderator impressions of users are based on their activity beyond the level of single-issue threads.

3.4.4 Characterizing behavior in BID posts

In the section "Issues with Moderation", we presented evidence that the primary sources of the perception of bias in BID are false negative judgments, where posts that contained behavior that seemingly violated the rules set on BID were ultimately not moderated. Thus, in our analyses,

we want to control for the case where users make high-risk, potentially offensive acts in their posts.

In order to identify the types of behavior that are associated with getting moderated, we choose to focus on speech acts within posts. While previous work has characterized offensive behavior using lists of curated terms associated with hate speech or profanity and other surface-level, keyword-based features [32, 86], we found that this method is unsuited for identifying the types of behavior associated with moderation. First, surface-form word and phrase-level features will not fully capture more subtle, implicit ways of attacking or offending other users, such as sarcasm or passive aggressive statements. Second, the use of offensive terms is acceptable behavior on BID in certain contexts. Profanity is generally accepted (e.g. "We do not mod for profanity, no matter what people have tried to flag.", "I have no issues whatsoever with profanity and sprinkle my posts with it just for amusement."), while hateful terms are often quoted or referenced in debates about language use (e.g. "I nearly blew a gasket when my mother referred to Obama as 'that n*gger in the White House", "Do you think homosexual people bully others when they speak up about people using 'gay' and 'f*ggot' as insults?").

We instead focus on the intent behind each utterance. The literature on speech acts argues that utterances in discussions function to achieve some conversational goal, called a speech act [6, 189]. Communicative intents present in discussions and the intents considered to be harmful depend on the norms in the community being examined. Therefore, we use an unsupervised model to capture the speech acts present in BID. Specifically, we adapt the Content Word Filtering and Speaker Preferences Model (CSM) [105], which has been demonstrated to separate the intentions of utterances from their content. CSM identifies dialogue acts in conversation by assuming that the conversation takes place against a backdrop of underlying topics that change more slowly in the conversation than dialogue acts. With the assumption that these two processes have different transition speeds, CSM learns a set of fast-transitioning *foreground topics* that capture dialogue act-related words and slower-transitioning *background topics* that capture more content-related words. This property of the model is desirable because we are interested in speech acts uncorrelated with topics being discussed.

Each thread in BID is considered a conversation in CSM, and each post in the thread as an utterance in the conversation. CSM assumes that the given data has a set of sentence-level speech acts, each of which is defined as a probability distribution over words, like traditional topic models. Thus, we segment posts into sentences using *sent_tokenize* from NLTK [13]. For moderated posts, we remove sentences with the associated moderated justification to avoid centering topics around the common collocates associated with our moderated post detection heuristic (e.g. "mod edit") We ran the original CSM, as well as an augmented version with supervision based on whether the post was moderated for selecting foreground and background topics. We set the number of sentence-level speech acts to the setting that gave the highest log-likelihood over the data (10 for unsupervised, 40 for supervised). The number of states (soft clusters of sentence-level speech acts) is set to 5 for the unsupervised model and 20 for the supervised.[5] We will primarily focus on interpreting the unsupervised model, as the final results from the regression model with supervised CSM followed the same patterns.

[5]The rest of the hyperparameters are set to: $\alpha^F = 0.1$, $\gamma^A = 0.1$, $\beta = 0.001$, $\alpha^B = 1$, $\gamma^S = 1$, $\eta = 0.85$, $\nu = 0.9$.

Speech Act	Examples
F0: Making a claim	i don't think the gender of your in-home role models matters all that much
F1: Making a counterclaim	but gender and race are linked / that is very variable by culture
F2: Expressing a personal perspective	i fully agree / i knew this too / i thought it was / i'm really surprised to see such a stink being made over this / i don't understand
F3: Correcting information	i think you're misinterpreting what's being said / missionaries serve in all places , not just college campuses
F4: Jovial side comments	it's that sort of day / ps - your ravetar is cute / i'll trade a slice of dessert pizza for one of your cupcakes
F5: Reporting personal experiences	i was coming back to the us from europe once , seated next to a mom with infant , i would guess about 8-10 months old .
F6: Exclamations and emotional outbursts	sheesh ! / thank you / le sigh / good grief / right / oy vey
F7: Statement of fact/evidence	i noted only 24 countries , all ruled at the time by white males , that preceded the us in granting women the right to vote .
F8: Probing/evaluation of perspective quality	can you explain that further ? / makes me take the article (even) less seriously .
F9: Proffering a hypothetical	if parents wouldn't buy the toys at those crazy prices , the speculators would be hit hard .

Table 3.2: Speech acts/foreground topics learned by CSM.

Identifying high-risk behaviors

After running CSM, we identified the learned speech acts most heavily associated with being moderated as our high-risk behaviors. It is difficult to interpret a speech act by examining the words with the highest weights, as is frequently done for topic models, because speech acts are highly associated with function words that reflect the style and intention of a speaker. Thus, we initially had two native English speakers separately interpret the learned speech acts for consistency by examining the 10 sentences with the highest weight for each speech act and looking for common themes and trends in user intention. After coming up with general themes associated with each of the ten learned speech acts, the two annotators conferred on their interpretations of each speech act and together, came up with a description for the major themes falling under each speech act. As a final sanity check, a third collaborator examined the descriptions generated by the two interpreting annotators to ensure that the descriptions aligned with examples falling under each speech act. Though this method has some limitations, as it is dependent on how different annotators interpreted utterances with few content overlaps, the main themes identified for each speech act were generally consistent between annotators and similar interpretation methods are commonly used for topic models and other forms of thematic analysis. The interpreted speech acts are displayed in Table 3.2.

Many of these identified speech acts are expected in a debate forum: speech acts F0: Making a claim and F1: Making a counterclaim are typical moves in argumentation. F5: Expressing a personal perspective establishes a user's credibility by relating their own experiences with the issue being discussed, while F4: Jovial side comments could be used to build social rapport with other users. Talk classified as F3: Correcting information, F7: Statement of fact through evidence, or F8: Probing/evaluation of perspective quality negotiates the reliability of information presented in the debate. On the other hand, speech acts for expressing a personal perspective (F2) and giving short exclamations (F6) are more surprising in the domain of political argumentation, as they are highly emotional in nature and primarily used to express a user's personal state, rather than make any moves towards building an argument.

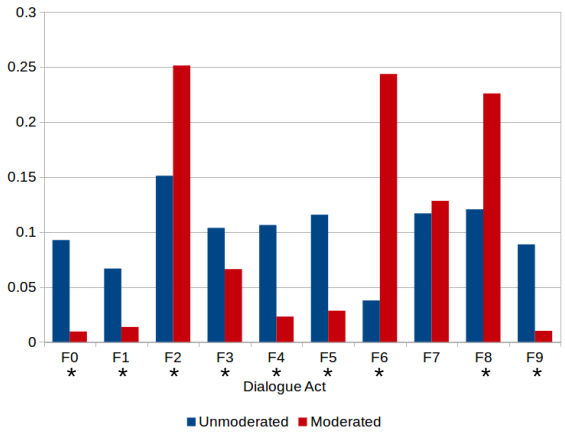


Figure 3.2: Comparison of the distributions of speech acts between moderated and unmoderated posts. Speech acts that are different with statistical significance $p < 0.05$ between moderated and unmoderated posts are marked with *.

variable	OR	Std. Err
high_risk	1.649***	0.054
mod_prev	1.328***	0.029
minority	5.685***	1.032
high_risk × mod_prev	1.000	0.006
high_risk × minority	0.915	0.056
mod_prev × minority	0.827***	0.018

Table 3.3: Logistic regression results for whether moderators are biased against users holding minority viewpoints. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

From Figure 3.2, we see that the three speech acts with the greatest difference in distribution between moderated posts and unmoderated posts are F2, F6, and F8. These three topics fit with violations of BID's moderation guidelines. F2, which contains many expressions of personal states and opinions, includes examples of harsh personal judgments that were moderated for being uncivil or attacking (e.g. "I do not for one second think you are trying to hide anything", "You would make a great politician"). F6, which is largely made up of exclamations and short comments, contains many snippy statements that could come off as being uncivil and dismissive to another user. "Le sigh", for example, sarcastically dismisses a previous comment as being beneath the author's attention, while "good grief" and "oy vey" express a speaker's irritation towards a previous utterance. As a whole, F2 and F6, which reflect more emotional acts, may be more associated with moderation, as the rules of BID espouse argumentation around the topic and not the users participating in the debate. Probing and evaluating other perspectives (F8) is inherently threatening to other users, and statements where a user immediately dismisses or questions another user's claim without reasoning often violate BID's rule against debating the person and not the topic. Though these particular speech acts and their association with moderation may be specific to the norms of BID, we argue that the unsupervised nature of CSM makes it easily applicable to identify high-risk, norm violating behaviors in other domains.

After identifying this set of high-risk speech acts, we combine their weights to create the control variable *high_risk*, which characterizes to what extent a given post has some form of high-risk behavior. Before we combine them, we normalize the weights on the three speech acts to have mean of 0 and a standard deviation of 1 across all posts to account for differences in scale between speech acts. This also allows us to measure the intensity of a speech act in terms of standard deviations from its mean. For a given post, we then take its maximum weight over the three topics as the value of the *high_risk* variable. Taking the maximum of the three topic weights allows us to indicate if at least one of the three high-risk speech acts has a high intensity in a post. Thus, *high_risk* gives us a measure of whether a post has an unusually large amount of the identified high-risk speech acts.

3.5 Findings

Table 3.3 shows the findings from our regression on which factors contribute to the likelihood of a post being moderated. Models trained using subsets of the full feature set showed the same pattern as the full feature set, which achieved the best 5-fold cross validation F1 score (70.44).

Are users with minority viewpoints unfairly moderated?

As expected, we see that the odds ratio on the *high_risk* speech acts (OR = 1.649, $p < 0.001$) has a significant positive relationship with the likelihood of being moderated. However, the *minority* variable also has a significant and stronger positive effect on being moderated (OR = 5.685, $p < 0.001$). Thus, users who consistently express minority viewpoints are more likely to be moderated than users who consistently express majority viewpoints, even accounting for behavior. In comparison, a standard deviation increase in *mod_prev*, the number of a user's posts in the last 30 days that have been moderated, has a smaller significant positive effect on the likelihood of a post getting moderated (OR = 1.328, $p < 0.001$). This lends weaker evidence that moderators are also biased against certain individuals with a history of moderation.

On the other hand, the interaction term *high_risk* $\times$ *minority* is not significant (OR = 0.915, $p = 0.148$). This means that users with minority viewpoints are moderated more even at the same level of high-risk behaviors as their majority-view counterparts. Figure 3.3, which shows the predictive margins of majority-view vs. minority-view users at different values of *high_risk* on the probability of a post getting moderated, demonstrates that this is the case.

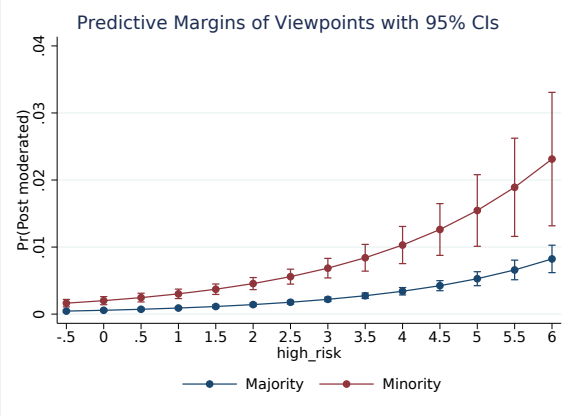


Figure 3.3: Comparison of predictive margins of minority vs. majority view users over different values of high-risk speech act use (standard deviations from mean) on the probability of a post getting moderated.

Though it is not directly relevant to our questions about viewpoint affecting moderation, the interaction term *mod_prev* $\times$ *minority* (OR = 0.827, $p < 0.001$) suggests an interactional effect that users in the minority are less likely to have a post moderated if they have been recently moderated. This interaction term, however, does not fully take into account the behaviors in the post being judged. Minority users who have been moderated in the past, for example, may actually avoid high-risk behaviors in order to avoid getting moderated again.

How strong is the effect of viewpoint on moderation?

The regression model suggests that posts by users who express minority viewpoints are more likely to be moderated than posts by users who express majority viewpoints. To get a standardized view of how much of an impact being a minority-view user has on having a post moderated, we calculated the Cohen's d between minority-view and majority-view users on whether users in the two groups are moderated. The effect size in terms of Cohen's d is 0.058, suggesting that the effect of a user's viewpoint on whether their posts are moderated is small. Thus, while there is evidence that there is some form of moderation bias against users who express minority views, the impact of bias on these users compared to the level of moderation on BID is negligible.

3.6 Discussion

From our regression analysis, we find evidence that the moderators of BID are more likely to moderate the posts of users with minority viewpoints, even after accounting for the types of behaviors that appear in the post. This suggests that moderators are somewhat biased against conservative users on BID, which supports our first hypothesis. On the other hand, we find that the effect size of the viewpoint bias is small, suggesting that the impact of the moderator bias is negligible, which supports the contrasting hypothesis that users themselves may be biased in overblown accusations of unfair moderation. As we can see, bias is present on both sides. However, the tension between the moderation team and ordinary users suggests that the perception of bias itself is a problem in political discussion forums, even if the actual bias is minimal. In this section, we discuss explanations for the actual bias we see in BID, the issues surrounding the perception of bias in political discussions, and future work to address the dual problems of actual and perceived bias on political discussion forums.

3.6.1 Sources of actual bias

In the case of BID, moderators can be susceptible to bias against certain viewpoints for a number of reasons. One of the most notable systemic reasons for bias [11] is ambiguity in how rules and guidelines can be interpreted. Users of BID explicitly raise this issue of rule ambiguity:

> *It's been said so many times I've lost count but the answer is: decide on clear, unambiguous rules; state them clearly; moderate for breaking those rules. Instead we keep going for nonsense like "be excellent" "be civil" "civil discourse".*

This type of ambiguity can make moderation susceptible to the cognitive biases of individual moderators [11] and mask subjectivity in determining who is acting in a "civil" way. When moderators are not aware of these biases and instead believe they are acting objectively, this can make moderation even more biased [113].

Specific cognitive biases that could influence moderators to moderate unfairly include the *ecological fallacy*, making assumptions about individuals based on judgments about a group [113]. In the context of BID, moderators likely recognize users who express conservative viewpoints and make judgments based on that group membership instead of individual behavior. *In-group/out-group bias* [113] may also be a factor in moderator bias. Moderators may more easily

make negative judgments about users expressing positions that differ from their own group's. Unfortunately, we cannot easily compare the ideological positions of the moderators in BID with the users they judge. Moderators do not give their names with mod edits and the current Ravelry API does not include logs of post edits, so pinpointing the specific moderator who handed down judgment is impossible. Additionally, it is difficult to determine the viewpoints of the moderation team on BID with our current approach for assigning ideology. Though moderators can in theory participate in debate threads they are not moderating, moderators in practice almost never post outside of their moderating duties. This is likely due to the high workload of the moderator role and a previous prohibition against all moderator participation in debate, which some moderators still follow.

Even without biased behavior from the moderation team, users with minority viewpoints in BID may still be more likely to be moderated if more of their posts are flagged. The moderation process in BID begins with users anonymously flagging posts as potentially violating the rules of discussion, which moderators then judge. Posts from majority-view users may be less likely to be flagged as there are, by definition, fewer users who have the incentive to flag offensive posts from majority-view users. In this case, even if moderators make fair judgments given what they see, due to imbalance in flagging they may miss posts that should be moderated from majority-view users.

3.6.2 Sources of perceived bias

Ambiguity in the moderator guidelines may also play a role in why users perceive bias against them when they are moderated. Vague rules, such as "Behave civilly" in BID, allow users to make judgments about their behavior in their own self-interest [11]. As it is in their interest not to get moderated, a user may be prone to *blind-spot bias* [113] and perceive themselves as being more civil than they actually are. If these users are then moderated, they may be inclined to believe that moderators made an unfair judgment by moderating them for their "civil" behavior. While we saw that most users viewed the main issue of censorship in BID to be false negative judgments, some users do argue that they have been moderated without cause:

> *Excuse me Pop but who did I personally attack ... Could you please explain why my post was modded?*
>
> *Again, can you explain how this post is off topic/about myself?*

Another possible source behind the perception of biased moderation from minority-view users in general is that minority users may experience a *halo effect* where their perception of the moderators are shaped by their experiences with other users within the group. Kelly et al. [118] found that in political Usenet groups, minority-view posts are overrepresented compared to the population of minority-view authors, meaning minority-view users generate more posts per person than majority-view users. We see this same pattern in BID (Figure 3.4). This pattern suggests that individual minority users must spend more effort on defending their views, as there are fewer people on their side who can help support their arguments. As a result, these minority-view users may feel like they are outnumbered and targeted by majority-view users, who can afford to spend less effort individually. These feelings of unfairness could be transferred to the

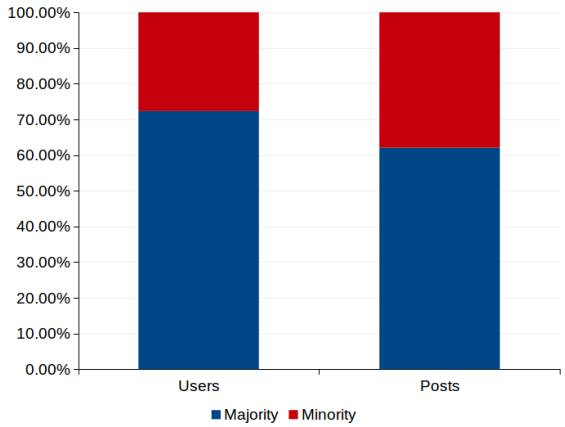


Figure 3.4: Comparison of viewpoint distributions over users vs. posts. The proportion of majority vs. minority are different between users and posts with statistical significance $p < 0.001$ by Pearson's chi-square test. Note that the distribution of viewpoints over posts is more balanced than the distribution of viewpoints over users.

moderation team, as the moderators are responsible for regulating conversations and maintaining order within the group.

3.6.3 Interventions and future work

One way of addressing the image of moderators as biased dictators is to shift both the power and burden of moderation in the group. Studying the political branch of the technology news aggregator Slashdot, Lampe et al. [136] argue for the success of a distributed moderation system in which users with highly rated comments can become moderators, who in turn are allowed to rate others' comments higher or lower. Along with a "meta-moderation" system that broadly crowdsources the review of moderator actions, they argue that this model can filter out unproductive behaviors as well as develop and pass on community norms. Such a meta-moderation system could not only counter moderator bias, but improve feelings of ownership in the moderation system for users who are not moderators. Crowdsourced moderation systems additionally allow the labor load of moderation to be distributed to the community, instead of placed in the hands of a select group of moderators. While BID has historically attempted a semi-distributed moderation system with a series of rotating volunteers performing temporary moderator duties, the addition of a meta-moderation system would allow more opportunities for ordinary users to engage in governance, especially for users who may not have the capacity to act as a moderator, even on a temporary basis. A danger of these meta-moderation systems that rely on the user base, however, is that minority-view users have fewer protections against the majority. An independent panel of judges may be helpful in protecting minority-view users from the tyranny of the majority, yet these judges should be made aware of their own biases to avoid introducing blind-spot biases [113].

Moderators accused of censorship are often criticized for providing little evidence for why a particular post is moderated while others are not. One possible intervention in these cases

is an automated system that does not directly classify posts as needing moderation, but instead provides better grounding for the discussions between moderators and those being moderated [81]. An example of such a grounding is an automated metric of inflammatory language that also provides comparisons to similar past posts that have been moderated. Making this visible to both the moderators and users could lend greater transparency and objectivity to how moderators operate, though this method would have to be safeguarded against the possibility of reproducing the bias of previous moderation.

Finally, it may be possible to address some of the sources of perceived and actual bias by working towards reducing ambiguity in how rules of proper debate are written. Most moderated discussion forums, like BID, frame their rules primarily in terms of what NOT to do (e.g. No personal attacks, don't derail the thread, etc.) Even the positively worded statement "Behave civilly" in BID is framed in terms of what not to do, as it is unclear what it means to behave in a civil manner. It instead implicitly tells users not to be uncivil. These negatively framed rules, however, are unlikely to capture the full range of offensive or inappropriate behavior, as users will try to find ways to circumvent the rules. One possible way of reducing the number of users skirting around ambiguous, negatively-framed rules is reframing rules in terms of positive discussion behaviors that users should include before they post. Encouraging political moderators to enforce rules in terms of what users should do may reduce both inappropriate behaviors and rule ambiguity by clearly defining what is expected of users.

3.7 Conclusion

Moderation in political discussion forums can be controversial, especially when claims of illegitimate censorship of specific views and individuals arise. In this chapter, we examined whether perceived unfairness against minority-view conservative users aligns with actual moderation patterns in Ravelry's Big Issues Debate forum. We found that users holding minority views are more likely to be moderated, even after accounting for levels of potentially offensive behaviors across groups. We found, however, that the effect of this bias is much smaller than how the issue is represented. Nevertheless, the perception that there is bias against certain subgroups remains an issue in political forums, as it may lead to tension and conflict over how moderation should be handled. We argue that ambiguity in how guidelines are laid out allows cognitive biases to slip in, explaining how both actual bias from the moderators and the perception of bias from users arise. We make recommendations for interventions that mitigate these biases by reducing ambiguity and increasing transparency in moderation decisions. While our study focuses primarily on Big Issues Debate, the techniques presented can easily be applied to other political debate forums and it is likely that our findings about the issue of perception of bias are not exclusive to this context.

Chapter 4

Ideological framing of policy

4.1 Introduction

In the previous chapter, we demonstrated that ideology can be used as a tool to challenge moderation decisions. In this chapter, we build on this finding and argue that a 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, especially in contexts where sensitive discussion can occur. Reddit [200] and Twitter [161], for example, have faced recent backlash for allowing racist content to remain on their platforms over concerns of bias against right-leaning viewpoints.

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 [179]. Here, we focus on a specific policy change to provide an in-depth analysis of the polarized stances users take.

The rest of the chapter is organized as follows. (1) We give an overview of related work on examining the effects of content policies and describe the recent Reddit quarantine policy update. (2) We present a general topic analysis of discussion surrounding the quarantine policy. (3) We describe how we operationalized polarization by characterizing users based on their participation across subreddits, then examine how different users frame issues within topics. (4) we discuss the implications and limitations of our work.

4.2 Content policies and their impacts

One of the primary roles of moderation in online spaces is the regulation of anti-social behaviors [121], such as spamming, cyberbullying, and hate speech. The design and best practices

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

for moderating abusive content on large social media platforms, however, is a fundamentally challenging issue [70], 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 [179], 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 [103, 194] 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. [32] investigated the impact of the 2015 Reddit hateful content ban on users who participated on the banned subreddits, while Chang and Danescu-Niculescu-Mizil [37] 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. [99] interviewed both users who used blocklists and users who have been blocked on Twitter on their insights about harassment and blocking. Myers West [158] 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. [98] 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.

4.3 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.

[2] `https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/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 thread over the course of one month using the Pushshift API [10]. 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[3] among the remaining users. Both comments by and responses to these meta-bots were removed, 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.

4.4 Topical analysis

Topic choice has been commonly used in NLP [49, 57, 207] as a proxy for *agenda-setting*, the strategic highlighting of what aspects of a subject are worth discussing [153]. Here, we first describe our preliminary topic analysis for discovering the range of topics discussed.

4.4.1 Models

We used Latent Dirichlet Allocation (LDA) [16] to construct our topics. While Structural Topic Models (STM) [177] 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.[4]

For the LDA models, we considered each comment to be a document. Comments were tokenized using SpaCy [87] 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 [181] than semantic coherence [154]. When analyzing and interpreting the topics discovered, we examined both the highest weighted words and example comments associated with each topic.

4.4.2 Results

Table 4.1 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 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

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

[4]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 [213]. 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	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 4.1: 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.

whether quarantined content will generate ad revenue. The proportion of **T0** across comments, however, is relatively low (13.6%), compared to discussion centered on 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

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

the subreddits quarantined during the quarantine wave, much of the discussion surrounding the announcement centered on 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 [115], promotion of violence and racism [198], and visibility manipulation and mobilization through bots [23, 62]. 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 on 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.

4.5 Characterizing user participation on Reddit

In order to better understand how different users highlight or *frame* particular aspects within each topic [22, 55, 163], 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 [32, 164]. 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. We then evaluate these subreddit interest "categories" and describe our method for considering users as a distribution of participation across categories.

4.5.1 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[5] 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 [164]. 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 $A_{ij} >= 5$.[6] We apply an additional user overlap threshold of 0.5 over the graph to ensure significant overlap in the users who participate on both subreddits for an edge.

4.5.2 Community detection

We use the Louvain community detection algorithm [17] 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

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{4.1}$$

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

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

[6]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 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.

(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 [134] of 1.0 and select the highest modularity partition of the dendrogram for our subreddit categories. The resulting 5 categories are shown in Table 4.2.

4.5.3 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 [36]. 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 4.2. 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 participation 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.

4.6 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 4.4 in an unsupervised manner.

4.6.1 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:

$$C_\beta(u_i) = \begin{cases} \text{left}, & \text{if } p_l(u_i) - p_r(u_i) > \beta \\ \text{right}, & \text{if } p_r(u_i) - p_l(u_i) > \beta \\ \text{neutral}, & \text{otherwise} \end{cases} \tag{4.2}$$

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

4.6.2 Polarized agenda-setting

Figure 4.1 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

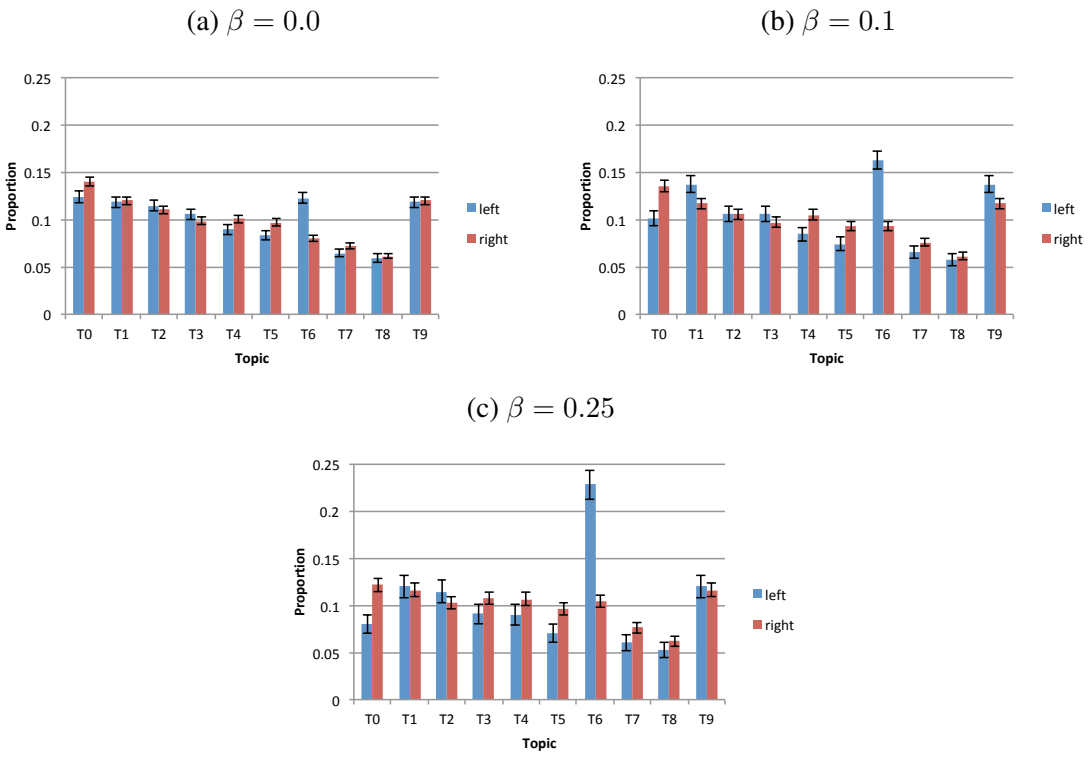


Figure 4.1: Topic prevalence across left and right-leaning users at different levels of polarization, with 95% confidence intervals.

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.

4.6.3 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 [155] as a salience

score, $\delta_w^{r(t)-l(t)}$:

$$\delta_w^{c(t)} = \log \frac{y_w^{c(t)} + \alpha_w^t}{n^{c(t)} + \alpha_0^t - (y_w^{c(t)} + \alpha_w^t)} \tag{4.3}$$

$$\delta_w^{r(t)-l(t)} = \delta_w^{r(t)} - \delta_w^{l(t)} \tag{4.4}$$

$$\sigma(\delta_w^{r(t)-l(t)}) = \frac{1}{y_w^{r(t)} + \alpha_w^t} + \frac{1}{y_w^{l(t)} + \alpha_w^t} \tag{4.5}$$

$$z(\delta_w^{r(t)-l(t)}) = \frac{\delta_w^{r(t)-l(t)}}{\sqrt{\sigma(\delta_w^{r(t)-l(t)})}} \tag{4.6}$$

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 4.6.2 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 4.6.2, 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 [159]. This suggests that left-leaning users are strongly concerned with harms potentially perpetuated by The_Donald as it is allowed to remain on the site.

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

4.7 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 harms caused by 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 [63] and moral value associations associated with the two sides [75]. 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 particular statements are strategically highlighted by different sides as a defense of a user's legitimate participation on Reddit without interference from an antagonizing group. While previous work has examined the use of free speech discourse as a defense against ego or expressive threat [214],

further exploration is needed into why the specific strategies of censorship vs. consistency and harms 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.[7] 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[8] argue that this ambiguity allows for the Reddit administration to protect popular but controversial subreddits like The_Donald.

4.7.1 Limitations and future work

This work 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. r/CringeAnarchy, r/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] `https://www.redditinc.com/policies/content-policy`

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

4.7.2 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 [59], 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,[9] we only release the user distribution across subreddit categories to prevent the user tracking across subreddits.

4.8 Conclusion

In this chapter, 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.

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

Chapter 5

Soft moderation, polarization, and community impacts

5.1 Introduction

In this chapter, we extend our investigation from the previous chapter on the moderation intervention of quarantining on Reddit. However, rather than viewing the practice of quarantining as a new platform-wide policy shift, we consider quarantining as it is applied, a community-level form of soft moderation.

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 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 [103, 104, 194].

As an alternative to bans, Reddit has proposed the use of quarantines as a softer form of community-level moderation. Quarantining, as described in the previous chapter, is a feature on Reddit that allows administrators to hide "communities that, while not prohibited, average redditors may nevertheless find highly offensive or upsetting" from being searched, recommended, or monetized. Unlike bans and deletions, however, quarantines do not directly limit the expressive power of the quarantined communities. Quarantined controversial communities can continue to operate on Reddit, and participants can still freely post content to these communities after the initial warnings and visibility restrictions. As a result of 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. Questions still remain over whether quarantines are effective in addressing toxic content while allowing communities to remain on Reddit or whether they perpetuate or even reinforce the existing controversial behaviors of these communities.

Prior work examining the impact of subreddit quarantines have suggested that they have lim-

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

ited effectiveness for mitigating toxic behavior on Reddit. Chandrasekharan et al. [35] examined how the quarantines of r/TheRedPill and r/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 [41] and Ribeiro et al. [174] 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 [56, 72] and the alt-right, whose audiences share similar values of patriarchy and cyberlibertarianism [150]. In our previous chapter, we found that redditors with varying political beliefs differed in how they discussed issues in the announcement, with right-leaning users associating quarantines with censorship and left-leaning users concerned about consistency of application and potential harms. These differences were aligned with moral values favored by liberals and conservatives [75], suggesting that users with different beliefs highlight different priorities when considering moderation issues, partly as a defense of their continued participation on Reddit. It remains unknown whether these differences may also be reflected in how different ideological groups respond to actual acts of quarantine, which partly threatens the participation experience.

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. In order 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 patterns of 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 political discussion in and out of the quarantined subreddits?

RQ1 is similar to the causal inference analyses in Chandrasekharan et al. [32] and [35], 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 [76]. we use these value associations as a proxy for changes in political polarization and attitudes expressed in these controversial political communities. Through this analysis, we examine how users carry or maintain political

linguistic practices across different communities in response to shifts in priority in response to the quarantines.

After addressing the three main research questions, we discuss how our findings about the stability of linguistic norms across subreddits may influence cross-community participation. Finally, we reflect on the implications of our analyses for platform moderation and intervention design.

5.2 Community-level content moderation on Reddit

Participation on Reddit is centered on interest-based subcommunities called subreddits. Reddit emphasizes free speech and cyberlibertarianism as central tenets of the platform [179]. 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. [32] 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.[2,3,4] 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:[5]

> *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 increase attention directed towards a controversial community. Citing the "Streisand Effect" [97], some users, both in support of and opposing the prominence of r/The_Donald on Reddit, perceived that its quarantine had actually driven up traffic to the subreddit:

> *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:

[2] https://www.reddit.com/r/TheoryOfReddit/comments/bhgcle/in_light_of_the_banning_of_rcringeanarchy_its/

[3] https://www.reddit.com/r/TheoryOfReddit/comments/cb5w2q/has_anyone_done_a_statistical_analysis_of_t_d_pre/

[4] https://www.reddit.com/r/AgainstHateSubreddits/comments/dwv7aw/update_the_donald_is_no_longer_evading_their/

[5] Quotes from redditors are lightly paraphrased for user privacy

Its still a congregating point to keep them in one spot.

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 on important social issues and discussions are influential in shaping political outcomes. Guided by these discussions over quarantines as a moderation strategy, in this chapter, 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. In the remainder of this section, we briefly describe the timeline for our two focal subreddits.

5.2.1 The_Donald

r/The_Donald was a subreddit centered on 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 [62, 149, 197], with around 750,000 subscribers at the time of its quarantine [201]. Before its quarantine, r/The_Donald was a well-known source of controversial, hateful, and violent content, and Reddit had implemented measures to prevent posts from the The_Donald from reaching the front page through collective vote manipulation [180]. The subreddit was eventually quarantined on June 26, 2019, with repeated calls for violence against Oregon police and public officials in response to a walkout by Republicans in the Oregon state Senate during a climate change vote cited as the catalyst for the quarantine. The subreddit was eventually banned in June 29, 2020, reflecting updated content standards in response to growing pressure from the Black Lives Matter movement [162].

5.2.2 ChapoTrapHouse

r/ChapoTrapHouse is a subreddit centered on the popular left-wing comedy podcast Chapo Trap House, influential in the populist "dirtbag left" movement [64]. The subreddit was reported to have around 130,000 subscribers around the time of its quarantine on August 6, 2019 [148], 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.[6] As a prominent left-leaning quarantined com-

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

The_Donald	ChapoTrapHouse
The_Durham	EscobarOpiumDen
DonaldJTrumpFanClub	ChapoTrapHouse4
The_MuellerMeltdown	LessTankieChapo
DrainTheSwamp	BlackWolfFeed
TheRightBoycott	ChapoFYM
Reddit_TDS	ChapoTrapHouse3
TheNewRedScare	EpsteinBrain
MetaCanadaTwo	CitationsNeeded
MAGAjuana	chapotraphouse2_2_2
HeadlineCorrections	FULLPOSADISM

Table 5.1: Top 10 control subreddits for The_Donald and ChapoTrapHouse. These control subreddits are communities where the total userbase has the highest percentage of users who also participate in the quarantined subreddit.

munity with controversy surrounding its quarantine, ChapoTrapHouse provides an interesting contrast to previously studied quarantined communities from the alt-right and Manosphere.

5.3 Data

Data for analysis was collected using full monthly dumps of Reddit activity ranging from May to September 2019 from the Reddit Pushshift API [10]. 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, as well as all activity by users who posted to the quarantined subreddit during the observation period. In addition to the quarantined subreddits, we extract data from related control, destination, and neighboring subreddits for analysis. In this section, we describe our procedure for finding these subreddits.

5.3.1 Control subreddits

In this chapter, 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. While these techniques cannot definitively prove that the quarantines caused a certain outcome in a subreddit, they enable us to investigate evidence of causality when randomized controlled trials are not possible. Following Chandrasekharan et al. [32], we used co-posting behavior from users who actively posted[7] 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

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

The_Donald	ChapoTrapHouse
WatchRedditDie	nfl
kotakuinaction2	CFB
gtaonline	PoliticalCompassMemes
StrangerThings	classic_wow
SubredditDrama	pan_media
YangForPresidentHQ	fantasyfootball
HongKong	fireemblem
modernwarfare	BetterEveryLoop
fantasyfootball	PresidentialRaceMemes
TheRightCantMeme	Epstein

Table 5.2: Top 10 destination subreddits for The_Donald and ChapoTrapHouse. These destination subreddits are the communities with the highest increase in posting behavior after the quarantine of the relevant subreddit.

percentage of users who were also active users in the quarantined subreddits. Control subreddits were filtered to ensure that none were quarantined or banned during the observation period. The top 10 control subreddits for The_Donald and ChapoTrapHouse are listed in Table 5.1.

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 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.

5.3.2 Destination 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 the invaded subreddits from Chandrasekharan et al. [32], we consider subreddits that had a 100% increase in posts by active users from the quarantined subreddits as *destination subreddits*. This definition gave us 33 destination subreddits for The_Donald and 36 destination subreddits for ChapoTrapHouse. Destination subreddits ordered by increase in total posting behavior after the quarantine are listed in Table 5.2.

For both The_Donald and ChapoTrapHouse, we note that many of the top destination subreddits include communities focused on interests outside of politics, such as gaming (e.g. gtaonline,

The_Donald	ChapoTrapHouse
unpopularopinion	chapotraphouse2
Conservative	BreadTube
PoliticalHumor	LateStageCapitalism
conspiracy	COMPLETEANARCHY
AskThe_Donald	ENLIGHTENEDCENTRISM
worldpolitics	unpopularopinion
Libertarian	PoliticalHumor
WatchRedditDie	socialism
hottiesfortrump	ABoringDystopia
pussypassdenied	TopMindsOfReddit

Table 5.3: Top 10 neighboring subreddits for The_Donald and ChapoTrapHouse. These neighboring subreddits are the communities with the highest percentage of users from the quarantined subreddit who also participate in that community.

modernwarfare), sports (e.g. nfl, CFB), and television (e.g. StrangerThings). As such, users may engage in drastically different behaviors in destination 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 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 5.3.

5.3.3 Estimating user ideology

Users who participate in a political subreddits may not necessarily be aligned with the beliefs and norms of that community [44, 80]. 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.

We estimate user-level beliefs by drawing from previous work leveraging participation across subreddits as a proxy for user interests or ideology [164]. 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 consid-

	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 5.4: Number of posts and users collected for The_Donald and ChapoTrapHouse

ered "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.

5.4 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 and new users that the quarantined subreddit received over time. For posts, we consider both submissions and comments on submissions made to the quarantined subreddit. We define new users as users who have never participated in the subreddit, using a 10 day buffer before our observation period to account for pre-existing users. 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 users who are or are not ideologically aligned with the main goals of the subreddit or power users, who have considerable influence over the direction of content in the quarantined subreddit. We consider the following user categories in our activity analysis:

- **Power users:** Users in the 90th percentile in terms of number of posts in the quarantined subreddit pre-quarantine.
- **Non-power users:** Users who participated in the quarantined subreddit pre-quarantine but are below the 90th percentile in 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 users identified as "neutral" under our methodology).

By our definition, power users and non-power users are already established members of the community before the quarantine. Thus, these categories were not examined separately in the new users analysis. Statistics for these users are located in Table 5.4

5.4.1 Interrupted time series analysis

To assess whether there is evidence of quarantines having an impact on the level of activity within quarantined subreddits, we ran Interrupted Time Series (ITS) analyses [12]. In ITS analysis, the goal is to determine whether there is sufficient evidence that an intervention at a known timepoint impacted a dependent variable Y. The values of Y before the intervention are used to find a slope and level for the underlying trend of the dependent variable without the intervention. This underlying trend is then used as a quasi-experimental counterfactual against a model that accounts for changes after the intervention timepoint to determine whether there is enough evidence that the intervention interrupted Y. In order to account for potential changes in both the level and slope of the dependent variable after our intervention, a quarantine, we fit the following regression model

$$Y_t = \beta_0 + \beta_1 t + \beta_2 X_t + \beta_3 t X_t \tag{5.1}$$

where X_t is a binary indicator of whether timepoint t takes place after the intervention. In this equation, β_0 and β_1 represent the level and slope of the underlying trend respectively, while β_2 and β_3 represent the changes in level and slope after the intervention. The change coefficients are then tested for significance. While the slope change coefficient β_3 was included in our regression, for all of our models, we found that β_3 either did not show a significant change or indicated a leveling off trend, with the total slope going to zero after the quarantine. Thus, we only report results on the level coefficient β_2 (hereinafter referred to as β for ITS analyses throughout the chapter).

5.4.2 Results

Table 5.5 shows the results for the level coefficient β from the ITS analysis, while Figure 5.1 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. [35], 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

	Posting Activity				New Users			
	β_{td}	p-value	β_{cth}	p-value	β_{td}	p-value	β_{cth}	p-value
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 5.5: 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. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

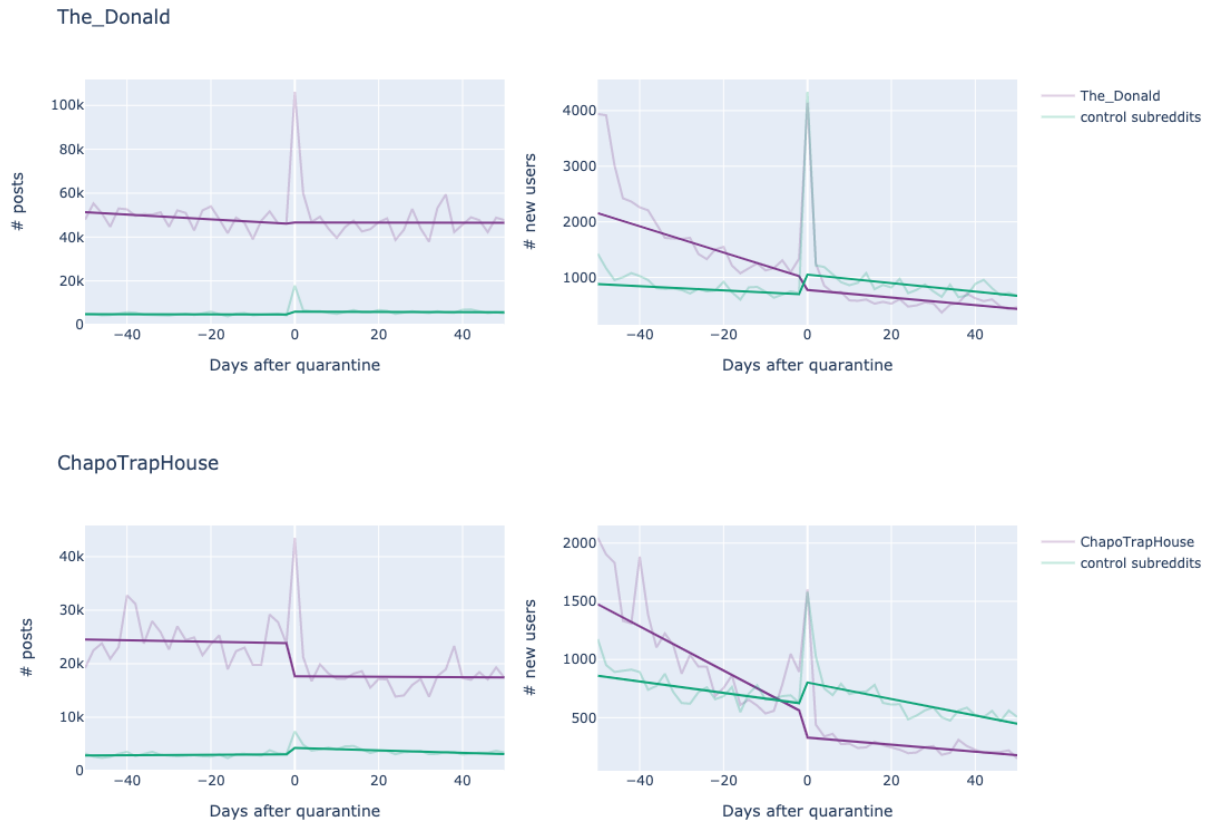


Figure 5.1: Total number of posts and new users over time in The_Donald and ChapoTrapHouse compared with aggregated control subreddits with fitted ITS regression.

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 signifi-

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

Table 5.6: Interrupted time series coefficients for percentage of cross-ideology interactions per user in The_Donald and ChapoTrapHouse. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

cant 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 actively contributing users in The_Donald may have been more strongly impacted by the quarantine, while the overall level of activity from other pre-quarantine users 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 new 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 [118, 156] 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:** A user engages in a direct cross-ideology interaction if the user replies to or is replied to by a user with a different ideology label than themselves.
- **Indirect interaction:** A user engages in an indirect cross-ideology interaction if they participate in the same comment thread as a user with a different ideology label than themselves. We define comment threads as comments within the same tree, rooted one level

below a submission. We chose to start comment threads at the level directly below a submission, as submissions are often used as links to other content sources, rather than conversation starters.

Table 5.6 contains the ITS analysis results for cross-ideology interaction. For both types of interaction and in both The_Donald and ChapoTrapHouse, we find a significant decrease in the level of cross-ideology interaction experienced by an average user in the subreddit. Thus, we find further evidence that quarantines have a homogenizing effect on the user experience within a subreddit.

Overall, our analyses of activity suggest that quarantines decreased the participation of users not ideologically aligned with the subreddit. This supports the hypothesis that by making access to subreddits more difficult, quarantines have a homogenizing effect on participation within an affected political subreddits.

5.5 RQ2: Visibility and monitoring

One claim raised in debates about quarantining was that quarantines allow controversial subreddits to be monitored from a known, centralized space. By allowing subreddits to remain on the platform, quarantines may give users who engage in the associated controversial behaviors the opportunity to continue on Reddit, rather than displacing them to self-hosted sites with less oversight. 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. Quarantines may also potentially compromise monitoring efforts by serving as a warning for affected communities to coordinate migration efforts off of Reddit [160, 174]. 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. We additionally analyze the texts associated with incidents in the quarantined subreddits tracked by monitoring subreddits to determine whether there were notable changes in the monitoring process post-quarantine.

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

Monitoring Sub	Quarantined Sub	β_{mon}	p-value
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 5.7: Interrupted time series coefficients for number of monitoring submissions mentioning the quarantined subreddit. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

- **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 from users to track individual incidents aligned with their objectives. Thus, for RQ2, we used submissions as our unit for analysis.

One notable aspect of the monitoring subreddits is that all three subreddits were shown to have an ideological lean from our subreddit labeling process (Section 5.3.3) – SubredditDrama and AgainstHateSubreddits were labeled "left" and WatchRedditDie was labeled "right". We do not argue that these monitoring subreddits are unbiased in how they operate. In actuality, the ideological biases of these subreddits are partly 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 Chapter 4. We are instead interested in seeing whether the political leanings of these subreddits led to differences in how issues were monitored in our quarantined political subreddits.

5.5.2 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 5.7). 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 on The_Donald after its quarantine. To better understand what drove this increase, we use a Sparse Additive Generative Model [54], or SAGE, to compare distinctive terms in WatchRedditDie submissions

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

Table 5.8: Distinctive terms in WatchRedditDie before/after the quarantine of The_Donald and ChapoTrapHouse obtained using SAGE.

before and after the quarantine (Table 5.8). 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, site-wide content issues, and antagonism towards the Reddit administration. This suggests that discussion around The_Donald on WatchRedditDie shifted from specific examples of content removal in the subreddit to issues surrounding its quarantine. 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 the quarantine of ChapoTrapHouse, where an opposing subreddit was moderated. This difference in reaction suggests that, to some degree, monitoring of moderation issues in WatchRedditDie is motivated by personal ideology.

While not significant, differences between β_{mon} for The_Donald and ChapoTrapHouse for the other monitoring subreddits suggest that there may be some ideological effects in those subreddits. In order to examine whether there may be similar trends for the other two monitoring subreddits, we repeat our SAGE analysis on SubredditDrama (Table 5.9) and WatchRedditDie (Table 5.10). We, however, see that the major shifts in top terms in submissions after the quarantine for these two subreddits are primarily centered on how the quarantined subreddits discussed specific political subjects, such as Charlottesville, Israel, Andy Ngo, etc., rather than broader moderation issues on Reddit. Nevertheless, we found some submissions high in ChapoTrap-

Quarantined Sub	Time Period	Top Terms (SubredditDrama)
The_Donald	after	sin, jericho, comment, electric, months, post, submission, banned, top, message
	before	bans, community, meltdown, thread, moderator, children, stories, people, account, issues
ChapoTrapHouse	after	againsthatesubreddits, thread, mods, drama, israel, hate, ironic, justiceserved, auto-mod, modified
	before	cops, dumbass, people, stonetoss, tankies, over-upset, level, threads, rogan, accusation

Table 5.9: Distinctive terms in SubredditDrama before/after the quarantine of The_Donald and ChapoTrapHouse obtained using SAGE.

Quarantined Sub	Time Period	Top Terms (AgainstHateSubreddits)
The_Donald	after	furry, antifa, charlottesville, misinformation, defend, watchredditdie, car, bullying, violent, rally
	before	moderators, racist, blatant, group, black, banned, admins, triggered, brigading, upvotes
ChapoTrapHouse	after	find, quarantined, glorifies, walk, nice, india, chapotraphouse3, ridicule, dictator, ocean
	before	celebrating, people, andy, police, attack, death, violent, killed, ice, terrorist

Table 5.10: Distinctive terms in AgainstHateSubreddits before/after the quarantine of The_Donald and ChapoTrapHouse obtained using SAGE.

House post-quarantine terms suggesting that some users in AgainstHateSubreddits question or oppose the quarantine:

> *AgainstHatesubreddits can't stand that CTH has been quarantined for hate. 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 the quarantined communities.

5.6 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 the amount of toxic content produced by a community. By limiting outside influence or intervention, for example, political ideas could be reinforced and radicalized in more isolated communities. Thus, in RQ3, we examine the impact of quarantines on an additional linguistic phenomena tied to political argumentation, the association of political issues with moral values. We examine the impact of quarantines on these linguistic features within quarantined subreddits, as well as related communities that may be affected by quarantine events By modeling how users' language change as they participate across communities over time, we examine to what extent these linguistic features are carried by users across subreddit boundaries.

5.6.1 Measuring toxicity

We use Perspective API[8], a popular machine learning system from Google for evaluating the impacts of texts, in order to estimate the amount of toxic content produced by a subreddit. While Perspective API suffers from limitations, such as scores being manipulable [88, 96], bias against minority groups [38, 185], and lack of consideration of more subjective forms of toxicity, its general purpose nature provides some advantages for our case studies. For our subreddits of interest, the exact reasons behind ChapoTrapHouse's quarantine remain unclear, unlike previous case studies examining communities associated with particular forms of racism and misogyny [35, 41, 174]. In our case, Perspective API can provide grounding for toxic behaviors without having to specify known targets of hate. Additionally, we focus our analysis on texts that were already produced, rather than texts targeted to deceive the system. As such, Perspective API can be widely applied to estimate toxicity across a wide variety of Reddit communities. Thus, for the quarantined subreddits, as well as our control, destination, 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. [187]. 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. [32] and Chandrasekharan et al. [35], 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.

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

5.6.2 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 [76] provides a framework for describing basic moral values held across human cultures and has commonly been used in political studies to describe differences in moral associations drawn by liberals and conservatives [75]. 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:** This foundation is concerned with caring for others and being sensitive to others' suffering. It is associated with evolutionary attachment systems.
- **Fairness/Cheating:** This foundation is concerned with issues of fairness, equality, and justice. It is related to the process of reciprocal altruism.
- **Loyalty/Betrayal:** This foundation is concerned with solidarity and self-sacrifice with one's in-group. It is related to human history of tribalism and shifting coalitions.
- **Authority/Subversion:** This foundation is concerned with respect for legitimate authority, social roles, and tradition. It is likely shaped from historical hierarchical social interactions.
- **Sanctity/Degradation:** This foundation is concerned with avoidance of the impure and taboo and likely shaped by the psychology of disgust and contamination.
- **Liberty/Oppression:** This foundation is concerned with the desire not to be restricted by a dominating power. While not one of the original 5 moral foundations, it was proposed [94] to differentiate the heavy emphasis that libertarians place on freedom and liberty, in contrast to issues of proportionality fairness.

Table 5.11 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 [106]. 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 5.12). 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 2,100 comments.

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 dic-

	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 5.11: 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/Cheating	67.65	11.72	49.14
Loyalty/Betrayal	49.32	27.06	58.87
Authority/Subversion	63.77	2.75	42.73
Sanctity/Degradation	18.48	5.47	64.68
Liberty/Oppression	63.41	10.18	47.38

Table 5.12: 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}).

tionary from Graham et al. [75] to account for more Reddit-specific invocations of moral foundations using pointwise mutual information [39]. Using both the original dictionary and our set of annotated posts to identify an initial set of posts for each moral foundation, 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 [184] 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

Feature	β_{td}	p-value	β_{cth}	p-value
Toxicity	-0.552	0.100	-0.062	0.883
Care/Harm	-0.702	0.100	-0.019	0.963
Fairness/Cheating	-0.455	0.255	-0.563	0.165
Loyalty/Betrayal	0.697	0.085	-0.266	0.521
Authority/Subversion	-0.209	0.610	0.146	0.737
Sanctity/Degradation	-0.467	0.080	0.187	0.649
Liberty/Oppression	0.219	0.612	0.461	0.293

Table 5.13: Interrupted time series coefficients for the value of the linguistic feature in The_Donald and ChapoTrapHouse. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

training/seed set for 10 epochs. We use AdamW [142] as our optimizer, setting the learning rate to $2\text{e}{-5}$.

Table 5.12 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.

5.6.3 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 5.13 gives the interrupted time series coefficient results for the linguistic features within each of our quarantined subreddits. Overall, we found no evidence that the quarantines caused a change in either the toxicity or the moral associations expressed in either The_Donald or ChapoTrapHouse. We see similar results when running ITS analysis over the control, destination, and neighboring subreddits for both The_Donald and ChapoTrapHouse. To illustrate how linguistic trends compare between the quarantined subreddit and its control, destination, and neighboring subreddits, Figure 5.2 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 destination and neighboring subreddits, which represent priority shifts and 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 descriptive 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.

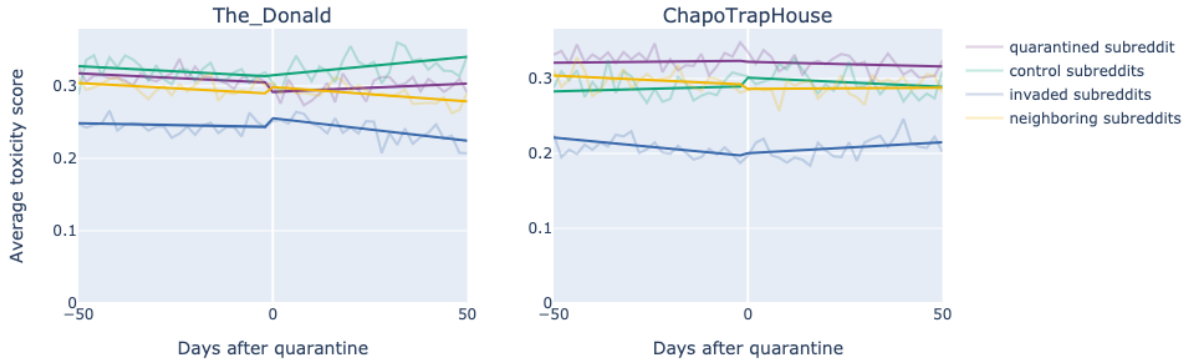


Figure 5.2: Average toxicity scores over time in The_Donald and ChapoTrapHouse compared with aggregated control, destination, and neighboring subreddits with fitted ITS regression models.

5.6.4 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 [77]. 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:

$$Y_t = \sum_{i=1}^{n} \alpha_i Y_{t-i} + \sum_{j=1}^{m} \gamma_j X_{t-j} \tag{5.2}$$

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 modification is used to ensure that the calculated prior linguistic tendencies of users and subreddits are independent of each other when making the prediction for the current timepoint.

Table 5.14 shows the results of the Granger analysis. Due to sampling limitations for Perspective API, we only run this analysis over the moral foundation categories, which we labeled over our entire dataset. 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

	Destination 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/Cheating	0.305	**0.627**	0.405	**0.437**	0.339	**0.687**	0.381	**0.666**
Loyalty/Betrayal	0.404	**0.572**	0.391	**0.511**	0.442	**0.621**	0.363	**0.696**
Authority/Subversion	0.248	**0.672**	0.390	**0.467**	0.291	**0.709**	0.380	**0.657**
Sanctity/Degradation	0.317	**0.603**	0.320	**0.625**	0.384	**0.640**	0.375	**0.642**
Liberty/Oppression	0.274	**0.616**	**0.399**	0.387	0.279	**0.746**	0.427	**0.614**

Table 5.14: Granger causality regression coefficients for linguistic feature values of an author in destination 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 $p < 0.001$.

linguistic tendencies than their own. Overall, this suggests that subreddit linguistic norms are quite stable and users adjust to these norms when participating in a subreddit. As such, quarantines, with their lack of true restrictions on participation, may be limited in their ability to address content issues within communities.

5.7 Quarantines and cross-community participation

In the previous section, we examined linguistic features related to toxicity and political moral value associations in both the focal quarantined subreddits and related destination and neighboring subreddits. We found limited evidence that quarantines had an impact on the language of either the quarantined subreddits or the related subreddits, likely due to strongly entrenched community linguistics norms across Reddit overall.

One simplifying assumption that we made in our previous Granger analysis, however, is that all changes in community participation were treated similarly for the prediction task, within each subreddit category (i.e. destination or neighboring subreddits). However, there are many possible reasons for why users may shift how they participate across communities. Table 5.15 gives another view of example destination subreddits with a substantial increase in participation[9] from users in The_Donald and ChapoTrapHouse after their respective quarantines. We break down these destination subreddits into three key categories for how they relate to their respective quarantined subreddit.

Unsurprisingly, one category of destination subreddits is subreddits ideologically aligned with the original quarantined subreddit, such as r/Conservative for The_Donald and r/MoreTankieChapo for ChapoTrapHouse. These subreddits may act as spaces where users can interact with ideological content similar to the original subreddit, without the restrictions of the quarantine.

[9]In this section only, we also include (1) subreddits that received an at least 1000 post increase in posts to account for larger subreddits and (2) destination subreddits defined within one week before and after the quarantine in order to account for subreddits that received a short-term increase in participation post-quarantine.

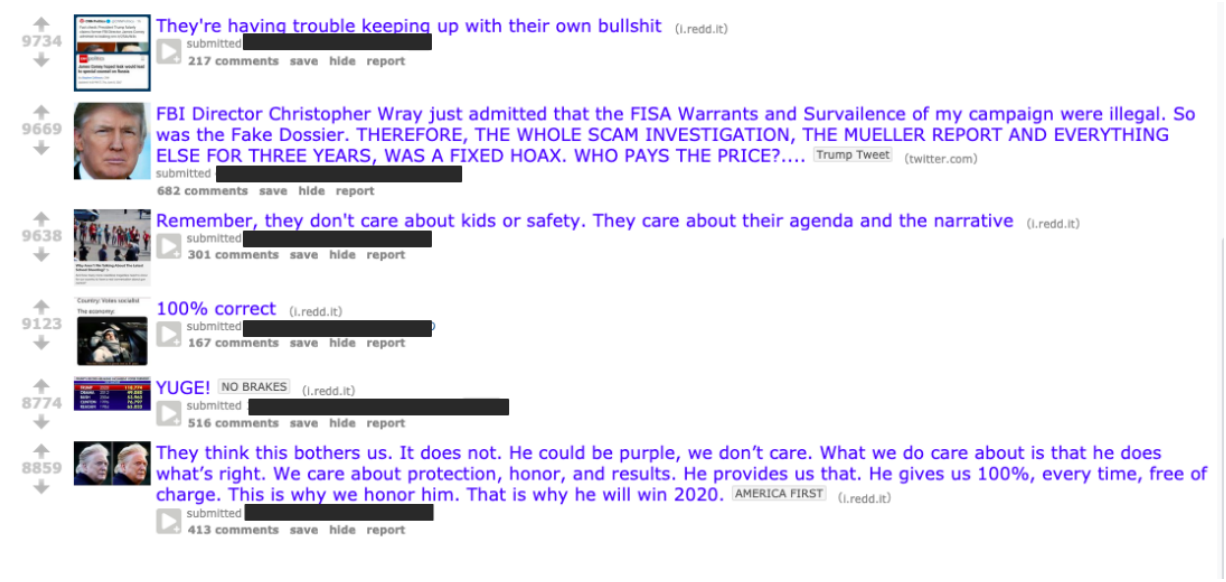
r/The_Donald Destinations	r/ChapoTrapHouse Destinations
Conservative	MoreTankieChapo
The_Donald_CA	LeftWithoutEdge
Republican	chapotraphouse_2_2_2
HillaryForPrison	ContraPoints
conservatives	chapotraphouse2
SubredditDrama*	SubredditDrama*
OutOfTheLoop*	OutOfTheLoop*
YangForPresidentHQ	JordanPeterson
Trumpgret	JoeRogan
ToiletPaperUSA	conspiracy*

Table 5.15: Examples of three categories of destination subreddits for The_Donald and ChapoTrapHouse: (1) quarantined subreddit to ideologically aligned subreddits, (2) quarantined subreddit to monitoring/informational subreddits, and (3) quarantined subreddit to ideologically unaligned subreddits. * denotes short-term only increase.

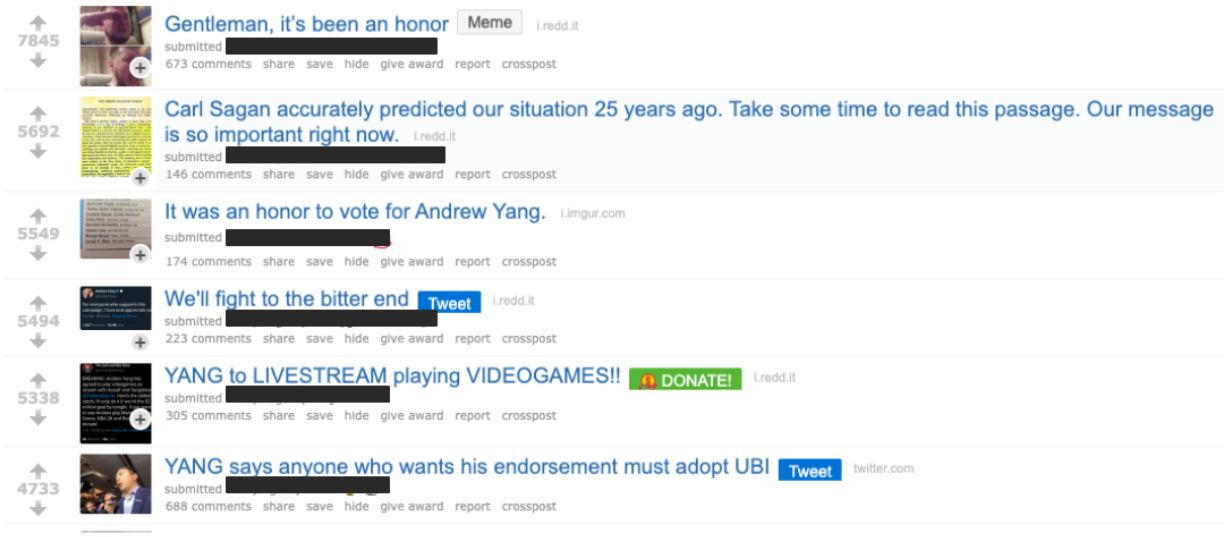
In the second category, for both The_Donald and ChapoTrapHouse, r/SubredditDrama and r/OutOfTheLoop, a subreddit dedicated to answering clarifying questions about major events both on and off Reddit, received a substantial short-term increase in participation from users from the quarantined subreddits. The increase in participation in these monitoring or informational subreddits suggests that users have some informational need, likely questions about why a subreddit was quarantined, directly after the quarantine. These subreddits, however, did not have a long-term increase in participation from users from The_Donald or ChapoTrapHouse.

A third, less intuitive pattern we see, however, is that there are some cases where there is a substantial increase in participation in subreddits that seem ideologically dissimilar or even opposed to the original quarantined subreddit, such as r/YangForPresidentHQ for The_Donald or r/JordanPeterson for ChapoTrapHouse. One possible explanation behind this increase in participation is that the participation that we see here is primarily antagonistic; in response to the quarantine, users may troll opposing subreddits as a way of venting against the quarantine. However, another explanation for this behavior is that there may social, stylistic, or value-based similarities between these subreddits that may appeal to users beyond the ideological content. Figure 5.3 shows examples of popular submissions from both The_Donald and YangForPresidentHQ, which demonstrate some degree of social and value-based similarity between the two subreddits – both subreddits position themselves in opposition to mainstream political discourse, value grassroots coordination on social media, and emphasize familiarity with memes and internet culture. This may suggest that the desire to substitute the social elements of the original quarantined subreddit may in fact override the ideological components of community participation. These kind of interactions, in turn, may provide a potential path for addressing some of the homogenization issues introduced by quarantines or deescalating toxic content produced by users.

In order to gain more insight into these cross-community participation patterns, in this section, we present further analyses into how users' behaviors across subreddits changed in response



(a) The_Donald



(b) YangForPresidentHQ

Figure 5.3: Examples of popular submissions from The_Donald and YangForPresidentHQ. Notable similarities between the two subreddits are opposition against mainstream political discourse and frequent references to meme or internet culture.

to the quarantines. We analyze broad shifts in user participation across subreddits in response to quarantines to examine how quarantines affected users' participation across communities. We also begin to investigate whether there is evidence that linguistic and social practices play a role in how users choose to participate across communities.

5.7.1 Trajectories in cross-community participation

Defining the proportion of total activity that a user spends in a particular subreddit as the *focus* of a user towards that subreddit, our goal is to find common patterns of user focus for The_Donald and ChapoTrapHouse over time and relate these trends to other subreddits that a user may participate in. For each active user in The_Donald or ChapoTrapHouse, we construct a time series of their focus in the quarantined subreddit, bucketed over spans of 5 days to smooth short-term sparsity or burstiness for an individual user's participation. This resulted in 23,758 user trajectories for The_Donald and 12,509 trajectories for ChapoTrapHouse.

We then perform clustering over these trajectories using soft dynamic time warping [42] as a measure of trajectory distance. Dynamic time warping (DTW) enables us to compare two time series with different lengths or speeds by allowing one-to-many mappings between points in different time series. Each point in the compared time series is mapped to at least one point in the other time series in a monotonically increasing manner such that the sum of Euclidean distances between matched points is minimized. We use soft-DTW, a differentiable variation of DTW that replaces the non-differentiable minimization with a soft-min, as our distance metric for k-means clustering over trajectories.

2,000 user trajectories were used for each quarantined subreddit to fit their respective trajectory clustering models. Using the elbow method over model inertia for k from 3 to 8, we found that 5 trajectory clusters for The_Donald and 6 for ChapoTrapHouse gave us the best fit. The resulting models were used to label all remaining users for each quarantined subreddit. In addition to the cluster assignments, we also label each user with the *transition* between the subreddit they were most active in before and after the quarantine.

Figure 5.4 shows the trajectory cluster centroids for The_Donald and ChapoTrapHouse, ordered by number of users assigned to each cluster, while Tables 5.16 and 5.17 give the user proportions, top user transitions, and average change in focus within each cluster for The_Donald and ChapoTrapHouse respectively. Overall, for both The_Donald and ChapoTrapHouse, most users engage in sustained low-level participation in the quarantined subreddit and relatively few users increased their focus on the quarantined subreddits after the quarantine. Comparing the two subreddits, we found that a smaller proportion of users in the ChapoTrapHouse maintained high focus after the quarantine (**C3:** High Sustained Focus, 10.37%) than in The_Donald (**C1:** High Sustained Focus, 21.77%). We also see that the overall percentage of users who were assigned to a cluster with a decreasing focus trend was greater in ChapoTrapHouse (**C1 + C4**, 35.19%) than in The_Donald (**C3:** Decreased Focus, 14.77%). This seemingly aligns with a finding in Section 5.4, where we noted a notable overall decrease in activity in ChapoTrapHouse but not in The_Donald post-quarantine.

Breaking down each cluster by user-level transitions, we see that users rarely made a substantial change to their distribution of participation across subreddits after the quarantine. For all trajectory clusters in The_Donald (Table 5.16), the most common transition for users within that cluster was The_Donald $\rightarrow$ The_Donald, meaning The_Donald was the subreddit that the user participated in the most, both before and after the quarantine. We see a similar pattern for ChapoTrapHouse (Table 5.17), where for all but one cluster (**C0:** Low Sustained Focus), ChapoTrapHouse $\rightarrow$ ChapoTrapHouse was the most prevalent transition among users. Users who had a transition where the top subreddit before and after the quarantine were the same made up 58.86%

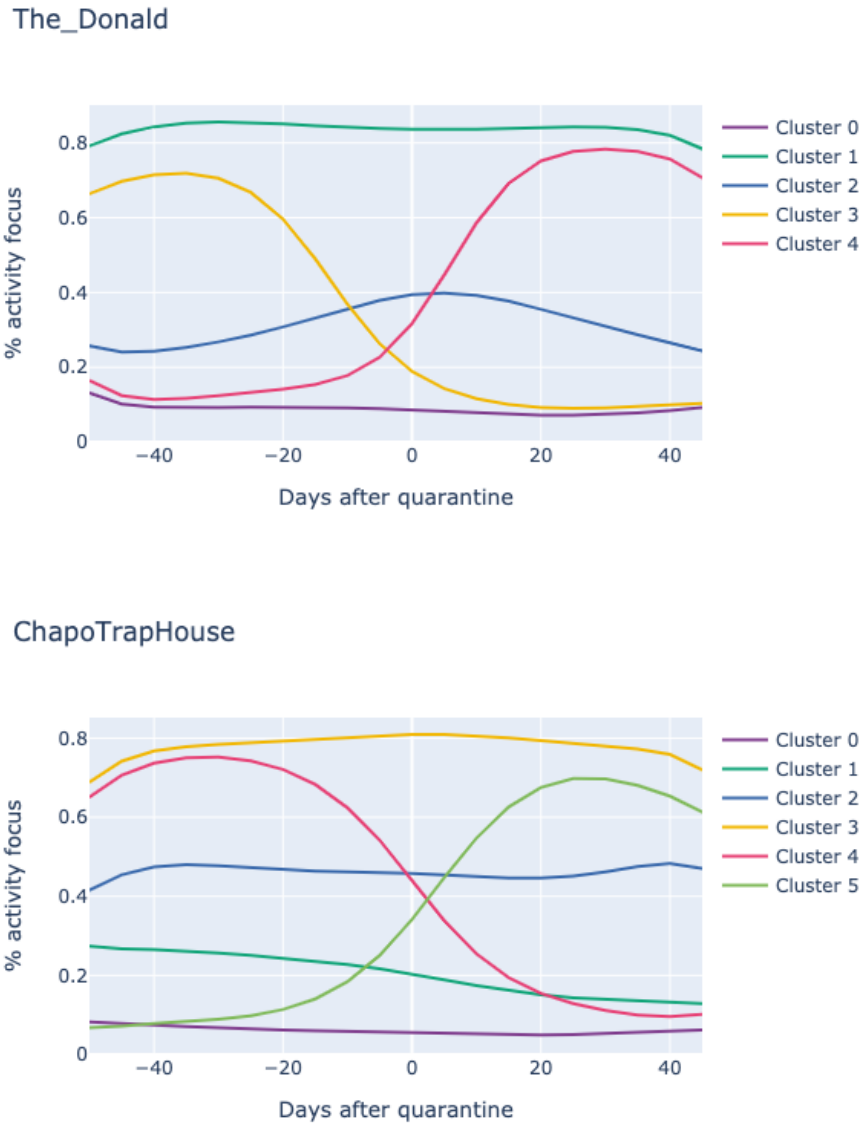


Figure 5.4: Trajectory cluster centroids for user focus (percentage of total activity) over time for The_Donald and ChapoTrapHouse. Clusters are ordered by decreasing number of users assigned to that cluster.

Trajectory Cluster	Top Transitions	Δ
C0: Low Sustained Focus (33.15%)	The_Donald → The_Donald (9.02%) The_Donald → Ø (7.47%) Ø → The_Donald (4.97%) unpopularopinion → unpopularopinion (1.66%)	-0.0523
C1: High Sustained Focus (21.77%)	The_Donald → The_Donald (98.74%)	0.0005
C2: Moderate Sustained Focus (19.35%)	The_Donald → The_Donald (62.40%) freefolk → The_Donald (0.96%) The_Donald → conspiracy (0.89%) unpopularopinion → The_Donald (0.78%)	0.0299
C3: Decreased Focus (14.77%)	The_Donald → The_Donald (44.94%) The_Donald → Ø (20.03%) The_Donald → Conservative (1.94%) The_Donald → conspiracy (1.34%)	-0.4040
C4: Increased Focus (10.96%)	The_Donald → The_Donald (55.05%) Ø → The_Donald (19.05%) freefolk → The_Donald (1.65%) Conservative → The_Donald (0.85%)	0.3542

Table 5.16: Top user transitions and average change in focus before and after quarantine within each identified participation trajectory cluster for The_Donald. Clusters are ordered by decreasing number of users assigned to that cluster. Soft dynamic time warping was used to assign 23,758 users to clusters.

of all users in The_Donald and 51.47% of all users in ChapoTrapHouse, with the next most prevalent transition (leaving Reddit entirely) only accounting for 5.45% of users in The_Donald and 3.72% in ChapoTrapHouse.

Overall, despite the observed increase in participation by quarantined subreddit users from the perspective of destination subreddits, this analysis suggests that individual users tended to still prioritize the same subreddits before and after a quarantine. In actuality, the most common shift in focus at the user-level is users from The_Donald or ChapoTrapHouse leaving Reddit entirely after their quarantines, rather than broad shifts towards one of the destination subreddits. This suggests that the impact of the quarantines on destination subreddits may be driven by a few users, slight overall increases in behavior, and/or notable only due to the relative sizes of the original quarantined subreddit and the destination community. We see some evidence of this, as destination subreddits average only 9 users per subreddit for The_Donald and 15 users per subreddit for ChapoTrapHouse that make a full transition to that subreddit post-quarantine. By running analyses at different levels of impact, such as for users, subreddits, and broader interconnected communities on Reddit, we can get a more comprehensive picture of the impact of quarantines on the participatory experience in Reddit from multiple perspectives.

Trajectory Cluster	Top Transitions	Δ
C0: Low Sustained Focus (36.25%)	ChapoTrapHouse → Ø (4.68%) ChapoTrapHouse → ChapoTrapHouse (3.73%) Ø → ChapoTrapHouse (2.98%) stupidpol → stupidpol (0.99%)	-0.0508
C1: Low Decreased Focus (27.29%)	ChapoTrapHouse → ChapoTrapHouse (32.63%) ChapoTrapHouse → Ø (4.57%) ChapoTrapHouse → MoreTankieChapo (1.11%) ChapoTrapHouse → chapotraphouse2 (1.05%)	-0.1376
C2: Moderate Sustained Focus (10.99%)	ChapoTrapHouse → ChapoTrapHouse (85.89%) ChapoTrapHouse → CFB (0.51%)	-0.0038
C3: High Sustained Focus (10.37%)	ChapoTrapHouse → ChapoTrapHouse (99.07%)	-0.0129
C4: High Decreased Focus (7.90%)	ChapoTrapHouse → ChapoTrapHouse (56.58%) ChapoTrapHouse → Ø (9.82%) ChapoTrapHouse → chapotraphouse2 (2.43%) ChapoTrapHouse → MoreTankieChapo (1.42%)	-0.3285
C5: Increased Focus (7.20%)	ChapoTrapHouse → ChapoTrapHouse (50.06%) Ø → ChapoTrapHouse (14.65%) nba → ChapoTrapHouse (2.11%) BlackPeopleTwitter → ChapoTrapHouse (0.78%)	0.3216

Table 5.17: Top user transitions and average change in focus before and after quarantine within each identified participation trajectory cluster for ChapoTrapHouse. Clusters are ordered by decreasing number of users assigned to that cluster. Soft dynamic time warping was used to assign 12,509 users to clusters.

5.7.2 Dynamics of cross-ideological community participation

While overall, we see that users tended to continue to prioritize the same subreddit before and after the quarantine, users who did shift their focus often sought out communities not ideologically aligned with the original subreddit. 72.5% of users in The_Donald and 47.8% of users in ChapoTrapHouse who shifted from the quarantined subreddit to another subreddit after the quarantine ended up transitioning to a subreddit not ideologically aligned with the original quarantined subreddit. Over 90% of these cross-ideology transitions comprise of shifts from the quarantined subreddit to "neutral" subreddits not necessarily related to politics, such as interest or hobby-oriented communities. This suggests that users who do choose to move from a quarantined subreddit tended to shift away from political discussion after their preferred political community was moderated. Nevertheless, we still see some common transitions between the quarantined subreddit and political subreddits with opposing ideology. Some of these shifts, such as The_Donald → YangForPresidentHQ and ChapoTraphouse → JoeRogan, may actually

	The_Donald		ChapoTrapHouse	
Feature	$\Delta \rightarrow$ **neutral**	$\Delta \rightarrow$ **left**	$\Delta \rightarrow$ **neutral**	$\Delta \rightarrow$ **right**
Care/Harm	0.21*	-1.07	0.45	-1.00
Fairness/Cheating	-0.05	-1.29	0.24	-0.65
Loyalty/Betrayal	-0.24	-1.03	-0.19	-0.72
Authority/Subversion	-0.32	-0.87	-0.44	-0.98
Sanctity/Degradation	-0.18	-0.89	-0.47**	-0.56
Liberty/Oppression	-0.16*	-0.82	-0.47***	-0.53

Table 5.18: Percent difference between linguistic features for cross-ideology transition users and matched same-ideology transition users from The_Donald and ChapoTrapHouse. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

be reflective of less intuitive similarities between political communities, where users may share social norms or specific values. To gain more insight into what may drive these cross-ideology shifts, in this section, we present an analysis into whether the moral foundations features may be able distinguish users who make cross-ideology transitions.

We focus on two different cross-ideology transition patterns: quarantined subreddit (i.e. The_Donald or ChapoTrapHouse) $\rightarrow$ "neutral" subreddit and quarantined subreddit $\rightarrow$ subreddit with the opposite alignment (i.e. "left" for The_Donald, "right" for ChapoTrapHouse). For each user that belongs to either pattern, we match that user with a quarantined subreddit $\rightarrow$ same alignment subreddit user (including the quarantined subreddit itself) who does not participate in the destination subreddit for the original user. We use the soft-DTW metric over user focus trajectories from our clustering approach (Section 5.7.1) to define similarity between two users, then greedily matched similar users without replacement.

The matched users were then compared using the moral foundation categories to determine whether these categories may have some predictive signal for someone making a cross-ideology transition. Linguistic features for users are calculated only from pre-quarantine posts, excluding posts from the destination subreddit for the cross-ideology transition users. For each of our two cross-ideology transition patterns, we use the Wilcoxon signed-rank test to test whether there are distinctive differences between the matched users for each linguistic feature. While ideally, we would also run this comparison on a subreddit-by-subreddit basis to account for subreddit-level differences in transitioning users, most individual subreddits have very few instances (< 20) of users making a full transition from The_Donald. Thus, we limit our analysis to a single transition category for a particular destination ideology.

Table 5.18 gives the average percent difference between linguistic features between cross-ideology transition neutrals and their matched users. For users transitioning between a quarantined subreddit and an opposite alignment subreddit, we did not find a significant difference in the use of moral foundations compared to the matched users. However, we do find significant differences between users transitioning between the quarantined subreddit and a neutral subreddit for **Care/Harm** and **Liberty/Oppression** in The_Donald and **Sanctity/Degradation** and **Liberty/Oppression** in ChapoTrapHouse. Notably, for both quarantined subreddits, users

who transition to a neutral subreddit after the quarantine tended to be less likely to use the **Liberty/Oppression** foundation. Due to Reddit's emphasis on cyberlibertarianism [179], this may suggest that differences in the prevalence of moral foundations may be a proxy for overall political engagement, rather than a distinguishing feature of ideological. Overall, however, user-level tendencies for moral foundations do not seem predictive of full user transitions after the quarantines. Nevertheless, due to the relative infrequency of full user priority shifts across subreddits, the linguistic practices of communities may still have an impact on cross-community participation on Reddit.

5.8 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. This supports the hypothesis that quarantines had a homogenizing effect on participation within political subreddits. However, we found no evidence that quarantines impacted the visibility of issues within quarantined subreddits (though the ideological biases of the monitoring subreddits may have impacted how they discussed quarantined subreddits) or language in the general Reddit ecosystem. We similarly found few differences in overall effects of the quarantine 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.

5.8.1 Implications for platform moderation

Prior work examining the impacts of quarantines suggested that they had limited effectiveness for addressing toxicity in online spaces [35, 174]. Our results support these previous findings, but also provide more insight into why quarantines were ineffective at addressing toxic content. We find evidence that the linguistic norms of subreddits are strongly entrenched and that users adjust to the linguistic practices of the communities they participate in. Similarly, analysis of cross-community participation trends before and after the quarantine demonstrated that users who choose to remain on Reddit after the quarantine continued to maintain their participation in their highest focus subreddit. As this was usually the quarantined subreddit itself, this suggests that users still preferred to engage in political discussion within the quarantined subreddit over other alternatives. We see more evidence of this from users who *did* shift to another subreddit post-quarantine, who actually preferred to move away from political discussion subreddits entirely, rather than seek out alternative political subreddits to participate in. This suggests that quarantines, as an intervention that allows ongoing participation within a community but does not directly address the content production experience, may be ineffective as strategy for mitigating offensive content on Reddit. Users who wish to participate in the original quarantined communities can still freely participate in the original communities, with few incentives to address initial behavioral concerns or seek out alternative communities. Similar to our reflection in Section 3.6.3, we suggest that interventions that allow controversial communities to remain on a platform should reward improvements in that community, rather than solely act as a punishment for toxic behavior.

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 homogenizing effect within the quarantined subreddits, limiting exposure to users and content with different beliefs in what remained the favored communities of most users post-quarantine. 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 and celebration after the quarantine of ChapoTrapHouse. As discussion of political and social issues tend to be sensitive and personal for users, disrupting participation in political discussion communities may have a particularly strong impact on user experience within a platform. Thus, in exploring effective alternatives to deplatforming political content, further research considering political polarization as a consequence is needed.

5.8.2 Limitations and future work

The linguistic insights in this chapter are based on labels from two ML systems for detecting toxicity and invocations of moral foundations in text. While these labels allow us to get a general sense of linguistic trends over time, 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 that might be prevalent in political communities on Reddit. Similarly, while our DistilBERT classifier was finetuned, then trained on Reddit political discussion data, our sample of annotated comments likely does not fully cover the full spectrum of how moral foundations can be invoked on Reddit. Finally, while we found that quarantines had a limited impact on toxicity and moral value associations within quarantined subreddits and related communities, they may have led to more subtle shifts in language within these communities not captured by our models, such as a shift towards discussion of Reddit content issues.

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 very high-profile political subreddits quarantined around a similar timeframe. While the similarities between our focal subreddits may be useful for isolating much of the interaction between quarantines and ideology, we may see different effects for quarantined subreddits with different properties. Lesser-known subreddits and subreddits centered on other controversial subjects, such as gore or eating disorders, may provide more general insights into quarantine impacts not covered by this or other prior work.

In this chapter, 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.[10] This is indeed the case with our two subreddits of in-

[10] `https://www.reddit.com/r/reclassified/`

terest, 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.

5.9 Conclusion

In this chapter, we investigated the quarantines of two ideologically distinct political subreddits, The_Donald and ChapoTrapHouse. Taken as a whole, our analyses reinforce previous findings suggesting that quarantines are ineffective at their intended goal of addressing toxic content, while also showing that quarantines increase polarization in political spaces. We found evidence that due to quarantines not directly addressing the participation experience within a quarantined subreddit, user behaviors within and across different communities ultimately remain stable. Overall, this suggests that quarantines may ultimately be ineffective at addressing the issue of offensive content, while also introducing other unintended and potentially negative consequences for political engagement. We highlight that future research into alternatives to removals and bans should examine the issue of entrenchment in existing communities and its implications for intervention design.

Chapter 6

Evaluating differences in political community norms

6.1 Introduction

In the previous chapter, we examined the impacts of the quarantines of The_Donald and ChapoTrapHouse on broader political discussion and engagement on Reddit. While we found that quarantines had some impact on levels of user activity within the quarantined subreddit, especially for ideologically non-aligned users, we found little evidence that quarantines had an effect on language usage within political subreddits. Instead, our analyses supported the idea that when users post content to a subreddit, they adjust to the linguistic tendencies of that subreddit.

Prior research on community norms often focuses on the processes for how community norms are learned and maintained. Lampe and Johnston [135], examined how three different mechanisms – transfer from other experiences, observation of the community, and feedback from existing members – affect how new users learn to participate on Slashdot, finding that all three influence user participation patterns. Rajadesingan et al. [173] examined different processes for maintaining toxicity norms in political subreddits, finding that pre-entry observation contributes the most to newcomer conformity. Our findings from the previous chapter partly align with these findings by suggesting, to some extent, that users develop a conception of the norms of a community through observation. While not focused on newcomers, we see that users adjust to the linguistic norms of the subreddits they participate in, outweighing their own distinctive linguistic tendencies. Through this process, political communities on Reddit are able to maintain stable linguistic tendencies. The question remains, however, as to how users recognize the salient and distinctive linguistic differences across the variety of political communities they can choose to participate in. Thus, in this chapter, we present an exploration of how to evaluate what is distinctive about political language across different communities. We focus our investigation on evaluating differences in *descriptive norms* or typical behavioral tendencies in these communities, rather than *injunctive norms* that are explicitly encouraged or punished by community standards [40].

In the first part of this chapter, we investigate to what extent annotators can recognize ideological distinctions between how left-leaning and right-leaning communities discuss political

entities on Reddit. In Reddit, where political subcommunities are focused around supporting certain issues or candidates, recognizing the ideologies behind a community plays an important role in how learning community norms. Thus, as a starting point for evaluating norms, we want to understand how people recognize ideological differences between content from different political subreddits. We introduce an entity-based paired ideology ranking task that specifically asks annotators to make distinctions between individual texts from left-leaning subreddits and right-leaning subreddits. Through this task, we examine the impact of experiential factors on annotator agreement and analyze what elements of these texts may make it difficult for humans to distinguish viewpoints across ideologically distinct subreddits.

Using insights from our analyses from the annotation experiments, we then propose a framework for categorizing political assertions and associations in order to capture more nuanced language differences across political subcommunities. We then use this framework as a lens for examining differences in assertion usage across different political communities. We first measure differences in assertion usage across political subreddits to examine what contributes to different linguistic norms within political subreddits that may be ideologically similar. We then evaluate an embedding-based model trained to capture linguistic tendencies maintained within subreddits on its ability to capture the fine-grained linguistic distinctions from our previous analysis.

6.2 Perception of ideological labels

Successful interaction in online political discussion communities often requires users to recognize and identify the ideologies behind texts. Being able to perceive the beliefs of interlocutors allows users to more carefully tailor their comments to more effectively engage with their political audience. In the context of Reddit, where political communities are centered around support for particular candidates or social issues [197], understanding the ideologies behind content in a subreddit plays a role in how users learn and conform to the discussion norms within that community. Thus, an important starting point for evaluating community linguistics norms is understanding how individuals distinguish the ideologies behind texts from different political communities.

The task of identifying the ideologies behind texts has been widely studied in NLP research, as the viewpoints reflected in political discussions can provide insight into the partisanship of beliefs [155] or the persuasive strategies used by different ideological groups [207] for their particular policy goals. One issue from directly drawing from prior NLP work on the relationship between text and ideology, however, is that the interactive, community-centric nature of participation in Reddit may rely on a different set of assumptions about normative behaviors in groups. The primary paradigm in language technologies for associating linguistic behaviors with group labels is based on *variationist sociolinguistics* [204]. Variationist sociolinguistics concerns itself with the study of structured variation in language between groups and takes an essentialist position on the influence of labels on language. In the context of ideology, with the variationist view, a person with left-leaning beliefs speaks, writes, or behaves in a consistently and observably left-leaning way. Under this paradigm, an observer or annotator examining a post from a left subreddit could then pick up on the distinctively left-sounding aspects in the text to infer that it was produced in a left-leaning context.

An alternative paradigm for how aspects of ideology are reflected in the language of a group or community is a perspective based on *interactional sociolinguistics* [205]. Under this view, there may be behaviors that are associated with certain ideologies, and people engage in behaviors to strengthen or weaken associations between these ideologies and themselves. Thus, beliefs are not inherently conveyed through the text itself but in the interaction between the author and an audience that they are contextually adjusting for. As a result, there may be certain contexts where authors may not need to strongly indicate their ideology in what is directly said to convey it. We hypothesize that this interactional view plays an especially prominent role in political subreddits, where aspects of the beliefs and values of the community are known and internalized by participants – subreddits are centered around known interests, and users can infer the beliefs of other users and the community around them through repeated interaction and feedback. A typical annotator for an NLP task, however, is generally presented with texts removed from the context where the content is generated. Thus, they may rely on a different set of expectations for what linguistic behaviors are associated with labels. As a result, different individuals may interpret the same text differently based on their own background world knowledge and how it interacts with the domain of interest.

To explore these issues related to identifying or perceiving ideology within political subreddits, in this section, we describe an experiment investigating whether annotators could identify distinctive, systematic differences between left-leaning and right-leaning subreddits. We describe the construction of an annotated corpus of posts from political subcommunities on Reddit, then use this corpus to explore how experiential factors play a role in how annotators perceive the ideologies behind texts. We focus our analysis on the perception of ideology of different political subreddits, as identifying the exact subreddit that a post comes from is likely infeasible for annotators. Building upon prior work investigating annotation bias [69, 108, 183, 188, 210, 221], we not only ask annotators to give ideology labels to posts from political subreddits on Reddit but also incorporate additional contextual information about the annotators making ideology judgments.[1]

While previous work [108] has shown that source-side contextual features, such as user profiles and previous tweets, can influence label quality in ideology annotation, we focus our analyses on contextual factors on the side of annotators to understand what factors influence human perception of ideology. Most similar to our work, Carpenter et al. [24] and Carpenter et al. [25] examine the impact of an annotator's identity and openness on their ability to accurately assess author attributes, including political orientation. We, however, focus on examining factors that we believe to be specific and salient to political participation on Reddit, namely the annotator's political beliefs, political knowledge, and familiarity with Reddit.

6.2.1 Dataset

As in Chapter 5, our dataset is drawn from the monthly dumps from May to September 2019 from the Reddit Pushshift API [10]. In addition to all the submissions and comments collected in Section 5.3, however, we augment this set with additional submissions and comments from

[1]This annotation study was approved by the institutional review board at Carnegie Mellon University.

the top political subreddits focused on U.S. politics[2,3] by subscriber count.[4] Subreddits in this augmented set were manually labeled as left or right, based on the subreddit description and top posts, to ensure that ideological distinctions between subreddits of interest were more consistent than the original quarantine dataset. As preprocessing steps, full submissions and comments were broken down into paragraphs, then tokenized using Stanford CoreNLP [146]. We consider these paragraphs as our unit of text representing a post throughout the rest of the chapter to constrain the amount of length variation between texts. Paragraphs consisting only of URLs were filtered out to limit content to Reddit-native text.

6.2.2 Paired ideology ranking task

For the construction of our annotated corpus, we introduce a paired ideology ranking task that makes it possible for workers from a wider variety of backgrounds to annotate ideology with less training overhead. Prior work on annotating the viewpoints of a text [7, 95] generally presents annotators with texts in isolation to label with an ideology of interest. One drawback of this approach is the high degree of political expertise that annotators are required to have to recognize that a text matches an ideology. Given a post to annotate on a particular political issue, such as economic policy, an annotator would have to recognize the issue referenced in the post, detect what the stance expressed towards the issue in the post is, then make a judgment as to whether that stance fits within the boundaries of a certain belief system in isolation. When extended to more general political discussion, such as our Reddit corpus, this process requires the annotators to be able to draw distinct ideological boundaries over a vast array of political issues. We present annotators instead with a paired ideology ranking task to reduce the amount of overhead in recruiting and training political ideology annotators. Rather than examining texts in isolation, annotators are shown two related texts and asked to select the text that is more likely to be authored by someone with a particular ideological perspective. By considering two related posts, the annotators can use contextual clues from both posts, as well as the comparison itself, to make a relative judgment of ideology, rather than needing to determine what specific ideology category a post belongs to. This enables us to use annotators who are reasonably familiar with U.S. politics and/or Reddit but are not necessarily political experts to make ideology judgments.

For the setup of our annotation task, our goal is to pair a post from a left-leaning subreddit with one from a right-leaning subreddit. In order to ensure that texts from a particular subreddit adhere to community norms, we require that the texts must be authored by an user labeled as "left" or "right" by the heuristic procedure in Section 5.3.3 and have a non-negative karma score. While this procedure is likely to give us posts that are aligned with the general U.S. definition "left" or "right", due to users on Reddit primarily engaging with pro-social home communities [44], we emphasize that we do not assume that these are inherently "correct" ideology labels for texts. Instead, we use these labels primarily to create a basis of comparison between texts from

[2] https://www.reddit.com/r/redditlists/comments/josdr/list_of_political_subreddits/

[3] https://www.reddit.com/r/neoliberal/comments/9195w2/what_are_the_biggest_political_subs_on_reddit/

[4] r/politics was not included due to its initial history as a default subreddit contributing to its high subscriber count.

	Selected subreddits
Left	r/LateStageCapitalism, r/SandersForPresident, r/democrats, r/socialism, r/Liberal, r/VoteBlue, r/progressive, r/ChapoTrapHouse, r/neoliberal, r/esist, r/YangForPresidentHQ, r/The_Mueller
Right	r/The_Donald, r/Libertarian, r/Republican, r/Conservative, r/JordanPeterson, r/TheNewRight, r/Anarcho_Capitalism, r/conservatives, r/ShitPoliticsSays, r/POLITIC, r/AskTrumpSupporters, r/AskThe_Donald

Table 6.1: Subreddits included in the entity extraction corpus and their ideological alignments.

	Selected entities
People	*Donald Trump*, *Joe Biden*, *Bernie Sanders*, *Barack Obama*, *Hillary Clinton*, *Robert Mueller*, *Nancy Pelosi*, *Kamala Harris*, *Alexandria Ocasio-Cortez*, Andrew Yang, Elizabeth Warren, Pete Buttigieg, Jeffrey Epstein, Bill Barr, Ben Shapiro, Vladimir Putin
Ideologies	*conservatives/conservatism*, *liberals/liberalism*, libertarians/libertarianism, socialists/socialism, capitalists/capitalism, communists/communism, centrists/centrism, left, right, Antifa, Islam
Organizations	*Republican Party/Republicans*, *Democratic Party/Democrats*, Congress Reddit, Google, Twitter, Facebook, Youtube, Amazon, CNN, Fox News, Department of Justice/FBI, Supreme Court, NRA, ISIS
Locations	Russia, China, Israel, Iran, Europe, Mexico, North Korea, Syria

Table 6.2: Selected entities included in the construction of the dataset. Italicized entities are also included in the screening set.

ideologically distinct subreddits and separate posts that are likely to align with subreddit norms vs. ones that do not.

To ensure that the text comparison actually helps annotators to perceive ideological differences, we want to avoid presenting annotators with two unrelated texts that are essentially considered in isolation. Instead, we want to show annotators paired posts that are similar in content. As a first step for generating comparisons with similar content, we require paired texts to mention the same entity, since political discussions are primarily centered on the politicians, organizations, and geopolitical entities influencing policy decisions. We use Stanford CoreNLP [146] to extract occurrences of people, locations, organizations, and ideologies over a corpus of the 12 most subscribed left-leaning and 12 most subscribed right-leaning political subreddits on Reddit (Table 6.1). We limit entities under consideration to those that have occurred at least 300 times in our corpus and are easy to disambiguate (i.e. common first names in isolation, such as Bill or Joe by themselves are excluded in the count). The considered entities are shown in Table 6.2.

In order to limit the impact of confounds, such as topic or entity salience, when comparing

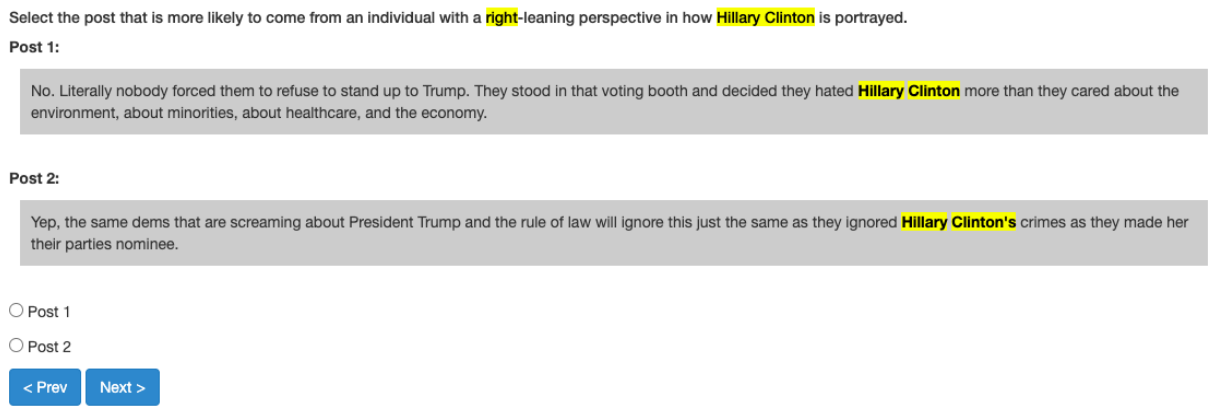


Figure 6.1: Screenshot of a question in the paired ideological annotation task. Annotators are presented with two texts discussing the same highlighted entity in a similar context, one from a left-leaning subreddit and another from a right-leaning subreddit. Annotators are asked to select which of the two texts is more likely to be authored by someone with the highlighted ideology.

posts with the same entity, we use propensity score matching [182] to match each left-aligned post with a right-aligned post that discusses the same entity in a similar context. We use logistic regression probabilities over averaged BERT embeddings [50] over the posts as our propensity scores, then greedily matched texts. A subset of 173 pairs was manually curated to use as *screening* set questions to ensure that workers had a baseline knowledge of U.S. politics. These screening pairs were selected to be easier than the main task pairs in certain aspects – they are more limited in which entities are discussed and express more explicit and/or extreme attitudes towards the highlighted. The entities select for the screening set are italicized in Table 6.2. All of the other remaining pairs are considered as part of the *main task* set for the annotation experiments.

6.2.3 Annotation task details

Given a pair of texts discussing the same political entity, we ask annotators to determine which of the two posts is more likely to have been written by someone from a left-leaning or right-leaning perspective. The political entity of interest is highlighted in both posts being compared. Pairs are also randomized to determine whether the question asks for the comparison from a left-leaning or right-leaning perspective. The annotation task interface is shown in Figure 6.1.

Annotators were instructed to use as many contextual cues as possible to form an impression of the political views behind the texts. To provide some guidance to annotators for what cues to consider, we prime workers to consider the following features in the instructions:

- **Attitude:** evaluation in favor of or against the highlighted entity. Ex: The statement *I trust **Bernie*** indicates that the author is someone who favors Bernie Sanders (left).
- **Positioning:** situating one's viewpoint with respect to the entity's. Ex: *Listen to **the Dems*** refers to Democrats specifically as "The Dems", highlighting them as an out-group relative to the author (right).

- **Jargon:** use of speciality in-group vocab. Ex: ***Trump*** *GEOTUS!* – GEOTUS stands for "God-Emperor of The United States", an abbreviation specifically used by Trump supporters (right).

We recruit workers from Amazon Mechanical Turk (MTurk) to complete our paired ideological ranking task in July 2020, shortly before the U.S. general election. Each worker was asked to annotate 21 pairs from our main task set and answer 5 screening questions, which were scattered throughout the assignment as an attention check. For each main task pair, we assign up to 5 workers to complete the question. We restrict the worker pool to the U.S. and filter out workers who scored less than a 80% on the screening questions. Overall, we collect annotations for 3,003 non-screening pairs over 50 entities.

6.2.4 Annotator background post-survey

After completing the annotation task, workers were asked to complete a survey (Appendix B) to assess their political beliefs, exposure to U.S. political news, and familiarity with political discussion on Reddit. Answers to the survey were manually inspected to assign annotators to groups within three identifier categories:

- **Political ideology:** This category indicates an annotator's self-identified political ideology. Annotators are labeled as *left*, *center*, or *right* along this category based on their answers about their ideology and affiliation with U.S. political parties.
- **News access:** This category indicates the annotator's exposure to and familiarity with news related to U.S. politics. Annotators are labeled as *news* or *non-news* based on how frequently they access news about the 2020 U.S. presidential election.
- **Reddit familiarity:** This category indicates the annotator's familiarity with participation in political discussion on Reddit. Annotators are labeled as a *redditor* or a *non-redditor* based on their level of participation on Reddit in the past year. Redditors are further subdivided into *political* and *non-political* redditors based on their familiarity with the political subreddits included in our corpus.

Worker ids were replaced with random ids after matching survey responses to the corresponding annotation results to protect worker privacy.

6.2.5 Dataset statistics and analysis

Annotator demographics

Of the 744 workers initially recruited for the task, 158 were discarded for answering fewer than 80% of the screening questions correctly, giving us a final pool of 586 annotators. Table 6.3 illustrates the distribution of the remaining workers across labels within the three categories. Labels do not appear to be correlated across categories (mean variance inflation factor = 1.302).

Agreement results

We use Krippendorff's α [131] to evaluate annotator agreement on our task to account for different worker pools for each question. Despite a high degree of agreement established across the

		# workers	α
Overall	-	586	0.2920
Ideology	left	346	0.3329
	right	153	0.2420
	center	87	0.3312
News	news	502	0.3029
	non-news	84	0.1784
Reddit	redditor	418	0.3200
	–political redditor	335	0.3126
	–non-political redditor	83	0.3419
	non-redditor	168	0.2362

Table 6.3: Number of workers and Krippendorff's α agreement within the annotator groups over the full non-screening set.

pool of screening questions ($\alpha = 0.7070$), the overall agreement across annotators in our general, non-screening set is low ($\alpha = 0.2920$), suggesting that much of the discussion on Reddit consists of more subtle, non-obvious, or implicit indicators of attitudes and ideology that may be open to interpretation.

We also calculate agreement for workers within each of our annotator groups (Table 6.3) in order to examine whether annotators with similar backgrounds are more likely to perceive ideology similarly. Overall, in-group agreement remains similarly low as the general task, ranging from 0.17-0.34 within each group. However, an interesting pattern across annotator labels is that workers who are less likely to be familiar with the expression of political ideology on Reddit – non-redditors ($\alpha = 0.2362$) and people who do not frequently read political news ($\alpha = 0.1784$) – have lower in-group agreement compared to other groups in the same category. This suggests that familiarity with the norms of political discussion on Reddit may contribute to a more consistent perception of ideology for Reddit texts. Unlike our results from a more limited set of political subreddits [193], however, annotators who identified with the right had the lowest in-group agreement ($\alpha = 0.2420$) among the ideology categories. One possible explanation for this is that compared to the worker pool from the original analysis, the right-leaning annotators from the current recruitment pool were significantly more likely to be non-redditors (41.17%) than the original pool (27.9%) or the overall rate in the current (28.67%) and original pools (27.84%).

We additionally use McNemar's chi-squared test over pairwise comparisons of annotator groups under the same category to examine whether annotators with different backgrounds differ in their judgments. To ground the comparison, we evaluate annotator groups based on whether the majority of workers in the group gave the same answer as the alignments of the subreddits the paired posts were from. For example, for questions that pair a post from The_Donald with a post from ChapoTrapHouse, we consider how often each annotator group judges the post from The_Donald as more right/less left than the post from ChapoTrapHouse. These counts are primarily used as a basis of comparison, allowing us to specifically quantify differences in judg-

Group comparison	% mismatch
right/center	35.79
non-political (r)/non-redditor	31.81
political (r)/non-redditor	30.90
left/right	30.32
left/center	30.21
news/non-news	29.79
redditor/nonredditor	28.98
political (r)/non-political (r)	28.82

Table 6.4: Comparison pairs with highest percentage of questions where the majority gave different answers.

ments between groups, rather than a true gold-standard. We find that for all comparison pairs within an identifier category, groups differ significantly in their answers over the same questions. In our pairwise comparisons, we see that the ideology of the annotator contributes heavily to variability in annotator judgments. The two groups with the highest percentage of questions with mismatched answers are center and right annotators, and 3 of the top 5 comparison pairs with the most mismatched answers are between ideology groups (Table 6.4). We also see that non-redditors were likely to make different judgments on our task compared to either of the redditor annotator groups.

We additionally investigate whether specific aspects of the texts themselves were likely to contribute to between-group variations in judgment by running salience [155] analyses for mismatched question pairs between ideology annotator groups. For left and right-leaning annotators, we found that annotators were less likely to select a post that expresses explicit abuse or non-policy oriented insults towards an opposing entity as being authored by someone with the same political views as themselves. For example, a right-leaning annotator was less likely to consider a post calling Biden a "pedophile" as right-leaning compared to liberal annotator. Similarly, a left-leaning annotator was less likely to consider a post mocking Trump's weight and intelligence as left-leaning compared to a right-leaning annotator. This may suggest that social desirability bias [132], may have an impact on decision-making, even when the task is not directly related to collecting data about the annotator themselves. Annotators in the center, on the other hand, were less likely to identify posts containing ideology-specific associations as belonging to their specific ideology, such as the use of "neoliberals" as a denigrating term for centrist Democrat entities or criticism of Pete Buttigieg around policing by users on the left. This suggests that annotators who are not strongly affiliated with a particular ideology may make judgments that differ from more ideologically-oriented annotators due to unfamiliarity with ideology-specific terminology and issues.

Overall, our results provide evidence that experiential factors related to participation in political discussion on Reddit influence the consistency of political ideology judgments made by annotators. Our analyses as a whole suggest that in language technologies, there is a greater need for targeted recruiting of annotators, especially those that are familiar with and contextualized to the domain being annotated.

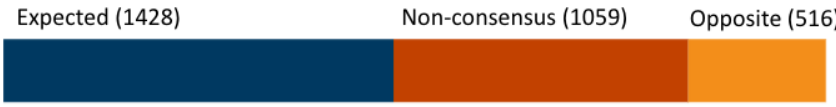


Figure 6.2: Distribution of questions across the three agreement pattern categories.

6.3 Question variation in ideology perception

Although we found evidence of systematic differences in how annotators perceived ideology, overall, we found that agreement between annotators was low, both within and across annotator groups. This suggests that differences in judgment across annotators did not primarily arise from cross-group differences in perception. Instead, the low overall agreement on the main task provides evidence that distinguishing ideology on Reddit may be inherently difficult and thus, linguistic norms in political subreddits may be subtle or open to interpretation. In this section, our goal is to investigate different factors in the questions themselves that may contribute to why ideology on Reddit is often difficult to perceive. To do so, we break the questions down into three distinct categories based on agreement patterns among annotators:

1. **Expected consensus:** This category contains questions where at least 75% of the annotators assigned to the question select the same answer. Additionally, that answer must agree with the expected ideology label, based on the relative ideological alignments of the subreddits that the paired posts originated from. For example, if a question pairs a post from The_Donald with a post from the ChapoTrapHouse and asks which post is more likely to be written by an author from a left-leaning perspective, at least 75% of the annotators assigned to the question must select the post from ChapoTrapHouse to belong to this category. 1,428 questions belong to this category.
2. **Non-consensus:** This category contains questions where fewer than 75% of the annotators assigned to the question select the same answer. 1,059 questions belong to this category.
3. **Opposite consensus:** This category contains questions where at least 75% of the annotators assigned to the question select the same answer. That answer, however, does not align with the expected ideology labels based on the relative ideological alignments of the subreddits that the paired posts originated from. 516 questions belong to this category.

Based on these three categories of question types, we discuss key properties of these categories that may contribute to the different observed agreement patterns.

6.3.1 Expected consensus questions

Table 6.5 gives the entities with the highest proportion of questions in each agreement pattern category. For the expected consensus category, the top entities tend to be entities that are more likely to draw a high degree of antagonism from one end of the political spectrum compared to the other. Among the top entities for this category are Mexico, the left in general, Antifa,

	Top entities
Expected	Mexico (70.37), the left (61.97), Antifa (58.49), Kamala Harris (58.33), Islam (56.52), Alexandria Ocasio-Cortez (56.16), Barack Obama (55.91), Elizabeth Warren (53.85), Department of Justice/FBI (53.57), Donald Trump (50.91)
Non-consensus	Israel (48.53), North Korea (47.06), Jeffrey Epstein (46.67), libertarians/libertarianism (44.07), Facebook (44.00), Reddit (43.69), Youtube (43.48), Hillary Clinton (42.19), Iran (41.46), China (40.00)
Opposite	ISIS (30.43), Amazon (30.00), Vladimir Putin (30.00), libertarians/libertarianism (25.42), Andrew Yang (25.00), Syria (25.00), centrists/centrism (25.00), the right (24.47), Iran (24.39), Facebook (24.00)

Table 6.5: Top entities and percentage of questions for that entity in each agreement pattern category.

and several progressive Democrats/Democratic politicians of color, who tend to be heavily criticized and attacked in right-leaning communities on Reddit. Similarly, Donald Trump tends to be viewed very favorably by users on the right but heavily criticized by the left. While some of the entities antagonized by the right are widely supported or defended by left-leaning communities on Reddit (namely the left in general and Mexico), others (e.g. specific politicians) may be more contentious on the left. Nevertheless, due to the high degree of antagonism towards these entities expressed by the communities on the right, these top entities tend to be the entities with a wide gap in perception between the left and right on Reddit.

Table 6.6 shows some illustrative examples[5] of questions that belong to the expected consensus category. Aligned with our observations for entities that are likely to contain questions in this category, these questions tend to have the post on the opposite side of the spectrum make a very strongly negative association with the key entity (e.g. the right calling Joe Biden a pedophile and associating liberalism with mental illness, the left connecting the right with Nazis). When both posts in a pair express a general negative attitude towards the highlighted entity, one common distinction is that the more obviously negative post tends to attack the character of the entity, while the less negative post tends to focus more on policy-oriented actions and beliefs associated with the entity. For example, in the comparison of the Bernie Sanders posts, the left-leaning post focuses on Bernie's actions as a career politician, while the right-leaning post directly accuses Bernie as a fraud and uses a derogatory term for a communist in an almost slur-like manner.

6.3.2 Non-consensus questions

The entities with the highest proportion of question pairs belonging to the non-consensus categories are listed in Table 6.5). These entities primarily consist of foreign countries (e.g. Israel, North Korea, Iran, China), social media companies (e.g. Facebook, Reddit, Youtube), and third

[5]Lightly paraphrased for user privacy.

Entity	Left post	Right post
Joe Biden	One annoyance from this poll is that apparently **Biden** lapped the field in locked support	True, either way I dislike **Biden**, the establishment full of pedophiles
liberals	And statistically speaking, conservatives have a much bigger amygdala than **liberals**. They give into fear mongering much easier than most.	I think this proves **liberals** suffer from various mental disorders.
Bernie Sanders	**Bernie** is a career politician the same as all the corrupt reps and dems out there. He doesn't care about you or me only himself.	**Bernie** is one of the biggest hypocritical frauds in DC. He should be removed from office the commie fuck
the right	Acting like bro nazis and radical **right wing** assholes are being denied their free speech is asinine at best and a boon to extremism at worst	Oh, yes. I cant turn on the news without seeing yet another **right wing** group killing people en masse here in the US lol... stupid leftists will be stupid leftists
Donald Trump	Except I'm literally saying that what **Trump** said makes **Trump** a racist, not anyone else.	You cherrypick things **Trump** says to fit your narrative. If it doesn't then he's a liar and if it does then it's proof that Russians got Trump elected

Table 6.6: Example questions from the expected consensus category.

parties (e.g. libertarians/libertarianism). In the context of the 2020 election, these entities tend to have limited day-to-day impact on U.S. political news.

We speculate that there are a number of reasons why annotators may struggle to reach a consensus on questions centered around these entities. First, because many of these entities are not central to U.S. politics, the relationship between these entities and the left and right in U.S. politics may be beyond general political knowledge of our annotators. In the Europe example in Table 6.7, annotators need to (1) be somewhat familiar with the debate over Brexit to recognize that the left-leaning post opposes the Leave campaign while the right-leaning post expresses anti-EU sentiment and (2) be able to relate these different views towards the EU and Brexit to general attitudes in U.S. politics towards Europe in order to distinguish these two posts. These distinctions about foreign entities may be beyond the scope of many annotators, even if the annotators are regularly exposed to U.S.-centric political news. Related to this issue, because certain entities are less likely to be understood with baseline knowledge of U.S. politics, they often have to be defined and clarified in discussion. Thus, these entities may commonly be mentioned in contexts that are not strongly polarized, such as the statements in the communism example in Table 6.7. Some entities may also draw similar associations across the political spectrum. Hillary Clinton,

Entity	Left post	Right post
Europe	I'm not saying everyone who voted to leave the **EU** is a massive racist, but I'm fairly certain that every racist voted to leave.	It is so sad to see the once great UK whimpering to be "released" from the **EU**
communism	**Communism** itself is a classless society where everything is for all and everyone does their part in upholding the wellbeing of themselves and others.	**Communism** is when a society is completely classless and moneyless.
Hillary Clinton	You're right. **Hillary** didn't get my vote for many reasons but one of the big ones is that she is a career politician and a piece of shit.	Bullshit. If **Hillary** had won, we'd be in a confiscation war right now. Don't be myopic.
the left	It looks like supporters of the one credible **left wing** party don't enjoy crust punks telling them they're wasting their time and should join the Communist party instead.	She's a Blue Dog Democrat, most got wiped out in the 2010 midterms. I hope they make a come-back and wipe out the radical **left**. Far more reasonable.

Table 6.7: Example questions from the non-consensus category.

for example, is commonly portrayed antagonistically by both left-leaning and right-leaning communities on Reddit as a corrupt and power-hungry (Table 6.7). Finally, for many question pairs, there may not be enough contextual information from the comment in isolation to determine the political ideology behind the author. Due to the interactional nature of participation on Reddit, it may not always be contextually relevant for a user to strongly indicate their ideological affiliation in a discussion, especially when the conversation is situated within a particular political subreddit. Thus, sometimes posts may be inherently difficult to distinguish ideologically.

6.3.3 Opposite consensus questions

The entities with the highest proportion of question pairs belonging to the opposite consensus category are listed in Table 6.5. While there are few obvious patterns among the top entities for the opposite consensus category, one notable observation is that "liberal/liberalism" has a higher proportion of opposite consensus questions (16.87%) than other left-leaning organizations or ideologies, such as "the Democratic party" (9.02%) or "the left" (5.63%). Within more far left-leaning communities on Reddit, such as ChapoTrapHouse, the term "liberal" has a different denotation than the general conception of "liberal" in U.S. politics. Rather than referring more generally to people on the left in U.S. politics, "liberal" in these spaces refers to neoliberals and the establishment left in contrast to "progressive" or "leftist". Thus, a ChapoTrapHouse user would likely not consider themselves a "liberal" while an outsider annotator more attuned

Entity	Left post	Right post
liberalism	Real talk though, **liberalism** was useful at some point and might be again. It isn't great right now.	Is there one anymore? **Liberalism** is what we used to call the left. Equality, tolerance and free speech.
Barack Obama	For liberals, **Obama** is a strategic choice. They'll dismiss destabilizing poor nations if they get their pride parades and weed. That's as far as they've thought.	I agree Presidents deserve vacation, but given the times Trump's criticized **Obama** for the same thing he's doing now, it just comes off as very hypocritical.
Andrew Yang	Yeah **Yang** is bullshit. The CENTRAL part of his UBI proposal is gutting benefits/welfare. He literally says that the funding for his yangbucks comes from people giving up benefits for it.	I see **Yang** is laser focused on solving issues that actually matter like poverty, automation, and mental health, instead of identity politics. That's why he has my support.
Republicans/ Republican Party	The point is conservatism leads to fascism, not the **GOP**.	They were stolen , but you are correct that the **GOP** establishment was complicit in the theft.
Joe Biden	If **Biden** gets the nomination, Trump will absolutely destroy him. As addled as Trump is, **Biden's** brain disease might be worse.	Pretty much confirms **Biden's** running against Trump
Democrats/ Democratic Party	**Democrats** are more corrupt than Republicans, because Republicans campaign on "I'm going to take money from everybody." **Democrats** just lie about it.	It's fine when Republicans do it, but it's horrible when **Democrats** say they're going to do it too.

Table 6.8: Example questions from the opposite consensus category.

to the general U.S. conception of "liberal" may, leading to some potential contextual scoping discrepancies between annotators and Reddit users. Table 6.8 gives an example illustrating this observation, with the post from the left-leaning subreddit being more critical of "liberalism" (i.e. neoliberals) than the post on the right discussing the left more generally.

Other examples from Table 6.8 show that many of the opposite consensus question pairs consist of comparatively negative attitudes expressed toward the highlighted entity from a subreddit aligned with the entity on the left-right spectrum. For example, many of the left-leaning posts tend to be more strongly negative towards left-leaning organizations and politicians than their paired right-leaning posts. There, however, appear to be consistent, systematic associations for why users or communities would express strongly negative attitudes towards an entity one would expect them to side with within the opposite consensus category. For example, there

is a tendency in left-leaning posts in this category to express disdain towards the Democratic party/establishment for maintaining the status quo instead of pushing for progress and suppressing more progressive voices. We see a similar tendency in right-leaning posts that support Donald Trump, where the posts criticize establishment Republicans by questioning their loyalty to Trump and ability to achieve Trump's policy goals. Left-leaning users who support specific Democratic primary candidates also consistently attack specific aspects of their rivals, such as Joe Biden's competence or Bernie Sanders' loyalty. One notable assertion among opposite consensus question pairs for Andrew Yang is that the right-leaning posts often express strong support for Yang as a Democratic candidate due to his focus on issues like automation and mental health over more identity-oriented issues. This aligns with our observations of cross-community participation in the previous chapter where there was a large increase in participation in YangForPresidentHQ by users from The_Donald after its quarantine, possibly in part due to shared values.

Overall, our analyses suggest that the usage of specific associations and assertions within posts may contribute to the different patterns of agreement we see across our annotators. Certain assertions made in reference to the highlighted entities may be expected or stereotypical to an annotator's conception of left or right, such as right-leaning users and communities being significantly more antagonistic and/or abusive to left-leaning entities than left-leaning users and communities. Others assertions, such as the belief that the Democratic Party is corrupt, may be commonly held by certain left-leaning communities on Reddit but not align with the image of the left held by outside annotators. As a result, when these types of assertions are used, annotators may fail to come to a consensus or perceive the posts as the opposite ideology from the actual communities they came from. Thus, understanding the specific associations and assertions that are commonly drawn with entities in different political subreddits may provide both crucial insight into how content in these communities is produced and perceived.

6.4 Political assertions framework

In the previous section, we noted that specific assertions and associations drawn with entities may contribute to differences in how content from political subreddits is perceived. Certain assertions may be considered typical for a specific political community but may seem unusual and counter-intuitive based on one's stereotypes about behaviors on the left and right. In order to formalize our observations from Section 6.3 and create a lens for finding subtle and unusual linguistic tendencies in political subreddits, our objective in this section is to develop a framework for categorizing common assertions and associations that are invoked when making value judgments towards political entities. Because these assertions are used in order to strategically highlight favorable and unfavorable aspects of entities in discussion, we design our framework as an extension of existing theories of value judgments. Inspired by Martin and White's framework of attitude [147], which revolves around a three-way system of meaning centered around emotion, ethics, and aesthetics, we divide our assertions space into three related dimensions of value judgments:

- **Affect:** Similar to the *affect* dimension in Martin and White, this dimension is concerned with emotional expressions made by a speaker or author. When directed towards entities in the context of political discussion, affective associations are primarily centered around

the expression of positive or negative sentiment directed towards the entity by the author of a comment. Thus, this dimension is primarily concerned with overall sentiment directed towards the key entity (*positive*, *neutral*, *negative*).

- **Policy-Oriented Judgments:** In Martin and White, the *judgment* dimension is concerned with how people praise or commend the behaviors of others. In the context of political discussions, the important behaviors to consider are related to beliefs and policy decisions associated with the entity of interest. Thus, this dimension focuses on associations attributing value to policy-oriented aspects of the highlighted entity. Martin and White initially break down this dimension into associations dealing with *social esteem* (ones social worth) and *social sanction* (how one should behave based on morals/ethics/edicts). For our purpose of categorizing different political entity associations and the strategies behind making these associations, however, we break down these associations into more fine-grained categories based primarily on Moral Foundations Theory [75] with some elements from the original subcategories from Martin and White:
  - **Care/Harm:** This category is concerned with either the benevolence of or the physical/material/emotional harms caused by the entity.
  - **Fairness/Inequity:** This category is concerned with issues of equality, fairness, and retributive justice carried out by or experienced by the entity.
  - **Integrity/Dishonesty:** This category is concerned with issues of honesty and transparency vs. corruption and deception related to the entity.
  - **Loyalty/Betrayal:** This category is concerned with the allegiance, loyalty, or adherence of the entity to their in-group and their beliefs and values. Groups of interest include political parties, coalitions, alliances, nations, or families.
  - **Authority/Subversion:** This category is concerned with the entity's adherence assigned social roles, participation in social contracts, or deference to social hierarchy or tradition. This includes discussion about sources of authority or influence related to the entity.
  - **Sanctity/Degradation:** This category is concerned with issues of purity and cleanliness vs. taboo in relation to the entity.
  - **Liberty/Oppression:** This category is concerned with rights and restrictions placed on or by the entity by others.
  - **Capacity/Incompetence:** This category is concerned with qualities related to the entitys competence or perceived competence at achieving their goals/success. This includes discussion of whether or not the entity was able to achieve their goals.
- **Appreciation:** In Martin and White, the *appreciation* dimension is concerned with the process attributing worth to something. In our framework, we center this dimension on associations that attribute inherent value to political entities without necessarily being tied to policies and views. Budesheim and DePaola [21] refers to these non-policy oriented aspects of political candidates as their "image".
  - **Physical Attributes:** This category includes associations related to physical traits of

Category	κ_H
Affect	65.25
Care/Harm	50.62
Fairness/Inequity	70.55
Integrity/Dishonesty	63.17
Loyalty/Betrayal	59.19
Authority/Subversion	46.37
Sanctity/Degradation	34.57
Liberty/Oppression	75.34
Capacity/Incompetence	50.62
Physical Attributes	66.17
Personality Traits	25.74

Table 6.9: Cohen's κ agreement results for the political assertions framework by humans (κ_H).

the highlighted entity, such as appearance and voice.

- **Personality Traits:** This category includes associations related to the personality or culture surrounding the highlighted entity, such as agreeableness.

Given a text with a highlighted entity, the assertion framework can be used to categorize what types of assertions are invoked in relation to the highlighted entity. In the **Affect** dimension, we indicate the polarity of the affective judgment the text makes towards the highlighted entity (e.g. positive, negative, neutral). Along the policy-oriented judgments and appreciation dimensions, the framework is used to indicate whether a post invokes an assertion or association belonging to each specific sub-category (e.g. **Care/Harm**, **Physical Attributes**).

6.4.1 Annotating political assertions

Based on this framework, two annotators, both familiar with discussion of U.S. politics on Reddit annotated 90 posts for whether each comment made an association with the highlighted entity belonging to each assertion type category (see Appendix C for details). Inter-annotator agreement was calculated using Cohen's κ (Table 6.9). Overall, annotators were able to achieve moderate agreement across most of the assertion type categories. As in Section 5.6.2, however, **Sanctity/Degradation** remained relatively low in agreement, due to differences in what was considered taboo in political contexts. Similarly the new category of **Personality Traits** had low overall agreement, due to differences in interpretation for whether certain associations commented on the culture surrounding a highlighted entity. After deliberation, the two annotators agreed to focus primarily on personality traits directly attributed to the key entity in a post The two annotators then annotated a set of 2,120 posts under the political assertions framework. All posts in the set of 2,120 were used during the paired ideology ranking task

6.4.2 Semi-supervised assertion labels

We explore two different approaches for propagating the human political assertion labels to a larger dataset for analysis and evaluation – a lexicon-based approach and a classifier based on a fine-tuned DistilBERT model. We take 2,000 of the labeled posts as a training/seed set and 100 of the remaining posts each as validation and test sets. For the lexicon-based approach, we additionally use a sample of 485,000 posts from our full set of political subreddits as unsupervised data for expanding the initial seed lexicons.

Lexicon

We use VADER [92] as our lexicon-based baseline for labeling the **Affect** dimension. VADER is a simple rule and lexicon-based approach for general sentiment analysis of social media content. As such, we expect it to have reasonable out-of-the-box performance on recognizing the sentiment of Reddit comments. For the remaining political assertion categories, however, we need to build category-specific lexicons for labeling. We begin with a set of seed words for each non-Affect political assertion category. For the policy-oriented judgments based on moral foundations, we use the original moral foundations dictionary from Graham et al. [76] as our seed set, splitting seed words in the original **Fairness/Cheating** foundation into those more closely tied to equality and justice and those more focused on corruption into **Fairness/Inequity** and **Integrity/Dishonesty**. For the **Capacity/Incompetence** category, we start with seed words from Martin and White [147] related to *capacity* (how capable something is) and *tenacity* (how reliable something is). We additionally augment the **Capacity/Incompetence** category, as well as the **Physical Traits** and **Personality Traits** categories with words from ESL learning resources for appearance and character.[6,7]

Using these seed words, we annotate the unsupervised sample with political assertion categories if there is an occurrence of a seed word for a category within that post. With this extended labeled set, we then calculate the pointwise mutual information [39] between each word in our labeled set and posts containing a particular assertion category C. We discard words that occur in more than 80% of posts and fewer than 0.01% of posts from consideration. After this filtering, the 100 words with the highest PMI for each political assertion category C that were not an original seed word in category C are then added to the lexicon for that category.

DistilBERT

Similar to the DistilBERT moral foundations classifier used in Section 5.6.2, we fine-tune a DistilBERT [184] pre-trained model to label paragraphs with political assertions categories. We use the same base language model from Section 5.6.2, which has been fine-tuned on a sample of r/politics posts from May to August 2019 using the masked language model objective. However, because the political assertions framework relies on detecting assertions associated with a key entity, the classifier training for the political assertions framework task is slightly different. For classification training and inference, in addition to the raw text of the post, we append at the

[6] `https://usefulenglish.ru/vocabulary/appearance-and-character`
[7] `https://www.ieltsspeaking.co.uk/ielts-vocabulary/`

Category	κ_L	κ_{DB}
Affect	17.21	**28.11**
Care/Harm	41.79	**66.55**
Fairness/Inequity	30.20	**52.77**
Integrity/Dishonesty	28.71	**48.81**
Loyalty/Betrayal	19.61	**59.15**
Authority/Subversion	-0.67	**45.09**
Sanctity/Degradation	-1.91	**52.36**
Liberty/Oppression	35.47	**42.03**
Capacity/Incompetence	14.29	**45.85**
Physical Attributes	**79.48**	0.00
Personality Traits	**38.97**	0.00

Table 6.10: Cohen's κ agreement results for the political assertions framework using assertions lexicons (κ_L) and a fine-tuned DistilBERT model.

beginning of each post a special tag, $[e]$, where e is the entity of interest for the post. Not all posts that we may wish to label with political assertion types, however, may have easily identifiable key entities. To account for posts where our Stanford CoreNLP pre-processing was unable to identify any entities within a post, during training, the key entity tag can be randomly switched to [NONE] for training instances. We use an entity drop rate $\lambda = 0.2$ and AdamW [142] with learning rate $2\text{e}{-5}$. Classification training was run for 10 epochs with early stopping.

Results

Table 6.10 shows the results of the two labeling approaches on the test set, using Cohen's κ as the evaluation metric. Overall, we found that both methods did not perform particularly well on labeling the **Affect** dimension. One possible explanation for the poor performance on **Affect** is that the category requires the labeling method to distinguish the polarity of a particular association, rather than just the presence of that association. Analysis of labeling errors from both approaches suggests that the polarity of a post is difficult to detect due to the heavy use of sarcasm and mockery on Reddit. Detecting the use of sarcasm on political Reddit may require additional contextual knowledge of the social norms on Reddit and the violation of expectations of the key political entities being discussed [109].

For the policy-oriented judgment categories, the DistilBERT model consistently outperforms the lexicon-based approach and achieves moderate agreement over all categories. However, the DistilBERT model performed very poorly on the appreciation categories. The lexicon-based approach, which tends to have higher precision but lower recall, however, greatly outperformed the DistilBERT model on identifying posts invoking the **Physical Attributes** and **Personality Traits** of key entities. This is likely due to the relative sparsity of posts invoking associations related to an entity's physical or character traits – fewer than 2% of the labeled posts invoked an association with an entity's physical traits and fewer than 7% with an entity's character traits. Thus, we use

	Expected		Opposite		Non-Consensus	
	left	right	left	right	left	right
Affect	**-0.304**	-0.492	-0.450	**-0.288**	**-0.355**	-0.462
Care/Harm	0.291	**0.309**	**0.283**	0.243	**0.374**	0.363
Fairness/Inequity	0.191	**0.193**	**0.215**	0.205	**0.267**	0.237
Integrity/Dishonesty	0.146	**0.229**	0.145	**0.217**	0.160	**0.252**
Loyalty/Betrayal	**0.397**	0.339	0.330	**0.348**	**0.351**	0.290
Authority/Subversion	0.231	**0.256**	0.255	**0.263**	0.259	**0.271**
Sanctity/Degradation	0.176	**0.181**	**0.140**	0.133	0.137	**0.176**
Liberty/Oppression	0.196	**0.216**	**0.230**	0.185	0.198	**0.206**
Capacity/Incompetence	0.286	**0.299**	**0.285**	0.267	0.221	**0.225**
Physical Attributes	**0.020**	0.013	0.040	0.040	0.019	**0.023**
Character Traits	**0.095**	0.080	**0.068**	0.030	0.061	**0.076**

Table 6.11: Average value of a political assertion category for left-leaning or right-leaning posts within the three question types based on annotator agreement patterns from the paired ideology ranking task. Bolded numbers indicate the higher average value for an assertion category between left-leaning and right-leaning posts for a question type.

the DistilBERT model to obtain semi-supervised labels for policy-oriented judgments and our constructed lexicons to label physical and character traits.

6.5 Analysis of political assertions

In this section, we use show how the political assertions framework can be used to analyze fine-grained linguistic distinctions in political communities. In Section 6.6.1, we investigate the relationship between the use of different assertion types in texts and how annotators perceive that text, with the goal of validating our observation that certain systematic associations may contribute to differences in agreement patterns when annotating ideology. In Section 6.6.2, we measure the usage of different assertion types across left-leaning and right-leaning subreddits to gain more insight into linguistic tendencies and norms within political communities on Reddit. Finally, in Section 6.6.3, we use our observations of assertion category usage across different subreddits to evaluate embedding-based models designed to capture subreddit linguistic norms on their ability to detect strategic value distinctions across communities.

6.5.1 Assertions across question types

In order to validate the use of political assertions framework as a lens into more subtle distinctions in content from political communities, we first investigate the relationship between the use of assertions in a text and how it may be perceived by annotators. In Section 6.4, we analyze different patterns of agreement for questions in the paired ideology ranking task, identifying three

key categories of question types. Based on our analysis of these question types, we speculated that there may be distinctive but systematic associations that may make posts from left-leaning and right-leaning subreddits more difficult to distinguish or even appear as the opposite side to an annotator. To investigate whether differences in assertion category usage may play a role in how questions were perceived in the paired ideology ranking task, we examine the differences in the distribution of assertion usage across question types. Because they were obtained for posts used in the main political ideology ranking task, we focus this analysis only on posts with human labels.

Table 6.11 shows the average value of a political assertion category for a post on the left compared to a post on the right across our three question types. Bolded numbers indicate whether left-leaning or right-leaning posts have a higher average value for a particular assertion category within a question type. One interesting pattern to note is that between the expected consensus questions and the opposite consensus questions, whether left-leaning posts or the right-leaning posts has a higher average value for a category often switches. For example, in expected consensus questions, the post on the right tends to be more negative than the post on the left. However, for the opposite consensus questions, on average, the post on the left is more likely to be negative than the post on the right. This suggests that posts in the opposite consensus category invoke assertions or associations in a fundamentally different way than most annotators' conception of left and right. Annotators may assume that people on the right are more likely to be negative, and thus, may reach the opposite conclusion when presented with a question where the left-leaning post is more negative. This pattern holds for seven out of the eleven assertion categories, suggesting that authors tend to rely on their own conceptions of what sounds left vs. right and thus, often select the opposite answer when the question violates their stereotype.

Certain assertion categories, however, remain consistently higher for one side vs. the other across all three question types. Notably, assertions related to **Integrity/Dishonesty** and **Authority/Subversion** are more commonly invoked in right-leaning posts, regardless of question type. This indicates that there are stable ideological tendencies for certain types of associations, such as right-leaning users/communities consistently valuing and invoking authority more than those on the left. This, however, does not necessarily mean that these assertion categories play no role in how ideology is perceived. As we saw in Section 6.4, opposite consensus questions commonly had the left-leaning post express negative attitudes towards entities on the left through associations with corruption (**Integrity/Dishonesty**) and establishment politics (**Authority/Subversion**). While the overall use of **Integrity/Dishonesty** and **Authority/Subversion** may still lean toward posts from right-leaning subreddits, left-leaning posts that invoke those categories may be more negative in the opposite consensus set.

Overall, we find some evidence that differences in assertion usage may contribute to differences in how annotators perceive the ideology of texts. In particular, there are notable differences in how assertion categories are used in posts from opposite consensus questions compared to expected consensus questions, particularly in regards to **Affect** towards the highlighted entity. This supports our observations from Section 6.4 suggesting that there may be some systematic differences in questions that may change how they were perceived by annotators. Our analysis in this section, however, provides additional evidence that annotators' own stereotypes about the political left and right play a role in how they interpret different questions.

6.5.2 Assertions across subreddits

In the previous section, we noted that differences in how assertions were used in left-leaning and right-leaning posts may impact annotator judgments on the paired ideology ranking task, we also found evidence of stable ideological tendencies for certain types of associations. For example, we found evidence that across the board, right-leaning posts were more likely to make assertions related to the integrity and dishonesty of key entities than left-leaning posts. However, in our previous analysis over questions from our ideology annotation task, we made the simplifying assumption that all subreddits with the same ideological leaning on the left-right political spectrum are actually similar to each other. This was due to our annotation task being centered around perceiving the ideology behind a text, rather than the specific community it came from. Identifying the subtle distinctions in language use across different political subreddits would likely be infeasible for human annotators, even for someone who regularly participates in political discussion on Reddit. Thus, in this section, we use the political assertions framework to perform large-scale analyses into nuanced differences between a wide variety of political subreddits. Using the semi-supervised assertion labels described in Section 6.5.2, we examine differences in how these assertion types beyond a binary left-right distinction. From this analysis, we can determine whether certain subreddits have an unusual expression over our assertion categories compared to other subreddits with similar ideological leanings.

In order to get a sense of differences in assertion usage across different subreddits, we rank political subreddits in our corpus by the percentage of posts in that subreddit labeled with a particular assertion category. Figure 6.3 then plots the percentage of posts in each subreddit against its rank, with blue points representing subreddits on the left and red points representing subreddits on the right. One interesting observation we see across the assertion categories is that relatively few of them are dominated by a single ideology. Although we see evidence that left-leaning subreddits are more likely to invoke the **Loyalty/Betrayal** and **Goal-Oriented Traits** categories and right-leaning subreddits are more likely to invoke Integrity/Dishonesty and **Liberty/Oppression**, most other assertion type categories are used at different levels by both left-leaning and right-leaning subreddits. This suggests that the relevance of certain assertions is not tied to ideology in and of itself. Instead, the range of variation in how different subreddits use assertions suggests that there may be certain community contexts where specific political assertions are more relevant. Ideology plays an important role in shaping the norms of political subreddits, but there may be other factors or practices in these communities contributing to differences in how assertion types are used.

In order to understand what may contribute to differences in assertion usage, Table 6.12 lists the top subreddits for each assertion type by percent usage. From this, we can see that assertion usage is heavily linked to the specific goals and interests of a certain subreddit. For example, in the **Care/Harm** category, we see that the left-leaning subreddits that most commonly invoke the category are focused on discussing the harms committed by police officers (Bad_Cop_No_Donut) or capitalist systems (ABoringDystopia, LateStageCapitalism). In **Fairness/Inequity**, three of the top subreddits in question commonly discuss cases of racism (FragileWhiteRedditor, SelfAwarewolves, beholdthemasterrace). The remaining subreddit, ENLIGHTENEDCENTRISM, on the other hand, is a left-leaning subreddit critical of centrists who portray themselves as impartial or unbiased while often aligning with right-wing views.

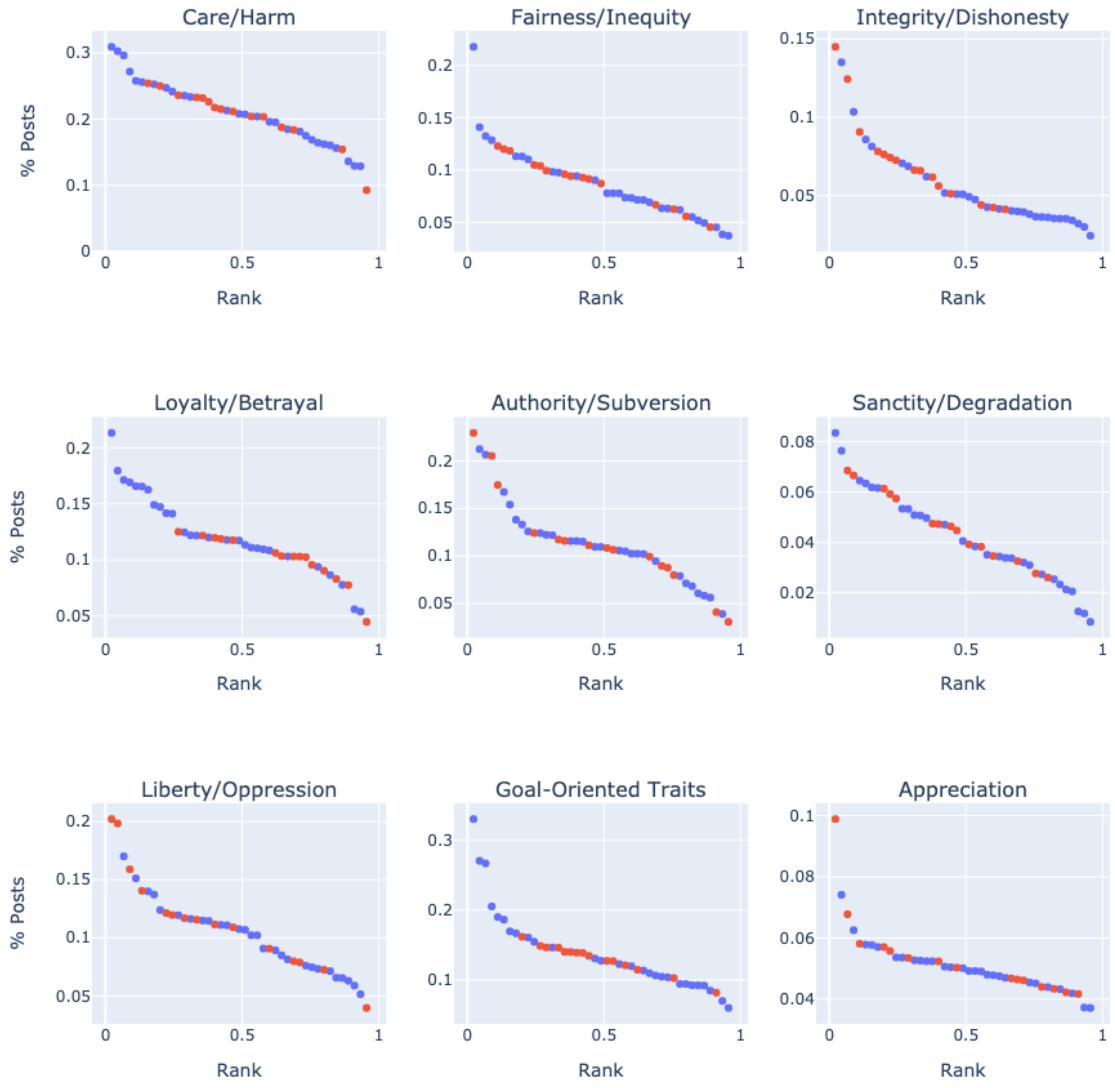


Figure 6.3: Relative usage of assertion categories for key subreddits. Each point represents a particular subreddit, with the percentage of posts using the assertion category in that subreddit plotted against its rank in assertion use. Blue points represent subreddits on the left, while red points represent subreddits on the right.

Subreddits aligned with libertarianism (Libertarian) and anarchism (Anarcho_Capitalism, Anarchism) are among the top subreddits in both **Authority/Subversion** and **Liberty/Oppression**. As these communities are focused on opposing government authority, discussion of personal liberty and criticism of authorities/regulations likely contribute to these assertion categories being prominently used in these subreddits. The top subreddits in terms of **Physical Attributes** usage are similarly unsurprising. hottiesfortrump and beholdthemasterrace, for example, are

	Top Subreddits
Care/Harm	Bad_Cop_No_Donut, ABoringDystopia, LateStageCapitalism, socialism
Fairness/Inequity	FragileWhiteRedditor, SelfAwarewolves, ENLIGHTENEDCENTRISM, beholdthemasterrace
Integrity/Dishonesty	conspiracy, The_Mueller, POLITIC, Liberal
Loyalty/Betrayal	ENLIGHTENEDCENTRISM, SandersForPresident, BreadTube, ShitLiberalsSay
Authority/Subversion	Anarcho_Capitalism, socialism, Anarchism, Libertarian
Sanctity/Degradation	beholdthemasterrace, AntifascistsofReddit, pussypassdenied, WatchRedditDie
Liberty/Oppression	Anarcho_Capitalism, Libertarian, Anarchism, AskThe_Donald
Capacity/Incompetence	VoteBlue, YangForPresidentHQ, SandersForPresident, democrats
Physical Attributes	hottiesfortrump, beholdthemasterrace, pussypassdenied, The_Donald
Character Traits	JordanPeterson, Liberal, BreadTube, progressive

Table 6.12: Top subreddits by usage of assertion categories.

subreddits centered around the physical appearance of female Trump supporters and white supremacists respectively, while pussypassdenied is a primarily right-leaning subreddit centered around discussion of women using their gender and often appearance for their advantage.

Some patterns in assertion usage at the subreddit level, however, may be more subtle. For the **Capacity/Incompetence** category, for example, the top subreddits, which are all left-leaning, are either centered around specific candidates (YangForPresidentHQ, SandersForPresident) or the general Democratic strategy (VoteBlue, democrats). In the subreddits centered around specific candidates, **Capacity/Incompetence** is commonly invoked to compare the capabilities of the preferred candidate with other Democratic primary candidates. In particular, **Capacity/Incompetence** was commonly used to criticize Joe Biden, who was both viewed as the front-runner and incompetent or senile at the time. For the subreddits more focused on the general Democratic strategy, on the other hand, **Capacity/Incompetence** was commonly invoked when discussing whether candidates were likely to be able to beat Donald Trump during the general election. We see a similar pattern in the **Loyalty/Betrayal** category, where the top subreddits all are left-leaning subreddits more critical of centrist and/or establishment Democrats. As a result, much of the discussion in these subreddits is focused on drawing boundaries between their communities and other communities who may consider themselves left.

One key element highlighted by our analysis is that political communities on Reddit are not part of ideological monoliths. Different subreddits on the left and the right serve different purposes, and thus, the language seen in these subreddits is distinctive to that community, rather than

its underlying political ideology. Some subreddits may choose to focus on certain issues, such as policing and economics, which lead to certain values and associations being highlighted in discussion. Other communities may focus on community-building and social interaction, leading to more commentary on non-policy oriented aspects of political engagement. Communities may also try to differentiate themselves from other factions that may appear to be aligned with them under a binary left-right distinction. By using the political assertions framework as a lens into fine-grained linguistic differences across subreddits, we can get a more nuanced picture of communities in the Reddit political landscape.

6.5.3 Evaluating subreddit norm models

In this section, we use the political assertions framework as a basis for evaluating how well an embedding-based model for detecting subreddit norms can capture subtle distinctions in language usage between different communities. Embedding-based techniques have been growing in popularity in use for comparative text analyses [3, 58, 67, 84, 140, 220], as they are believed to capture more abstract, higher-order aspects of semantics and social meaning. As such, we propose a lookup model for learning a shared representation for texts from the same subreddit in order to encode common linguistic practices for each subreddit. We then analyze whether there is evidence that the learned subreddit representations capture assertion usage tendencies across political communities in Reddit.

Model description

Our approach for learning subreddit embeddings is inspired by a technique in multilingual machine translation [107], where representations are learned for artificial language tokens added to the input to control for target language in generation. Given a set of texts with some label, the goal of the lookup model is to learn a representation for each label, which is shared between texts with the same label. With parameter constraints, the model is encouraged to store features shared between texts with the same label in the label representation. In our case, the subreddit lookup model tries to learn an embedding representation for each subreddit a text can come from, which will ideally store information about that subreddit's distinctive linguistic norms in comparison to other subreddits. We use an autoencoder to learn representations for individual posts, but during the decoding/reconstruction process, we augment the post-specific representation with the embedding representing the subreddit in which the comment takes place. Figure 6.4 illustrates the architecture for learning the subreddit representation space $\mathcal{S}$.

Given the text of post $\boldsymbol{p}_i$, the goal of an autoencoder is to encode and compress $\boldsymbol{p}_i$ into a vector representation such that a decoder can reconstruct the text from the representation. We use a bidirectional LSTM with attention to construct an embedding $\boldsymbol{v}_{i(full)}$ representing post $\boldsymbol{p}_i$. In order to provide some constraints so that the subreddit embeddings contain information that is useful for augmenting the post representation, we further compress $\boldsymbol{v}_{i(full)}$ using a linear transformation $\boldsymbol{P}_c \in \mathbb{R}^{f \times d}, d < f$, giving us a compressed post text embedding $\boldsymbol{v}_i$:

$$\boldsymbol{v}_i = \boldsymbol{P}_c^\intercal \boldsymbol{v}_{i(full)} \tag{6.1}$$

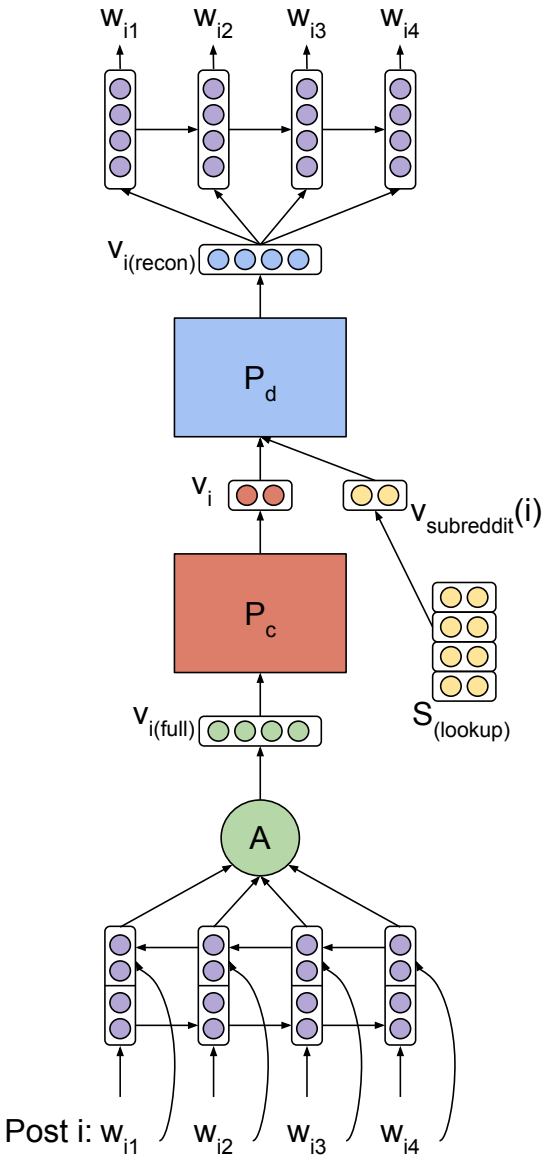


Figure 6.4: Architecture for the subreddit-based lookup autoencoder.

By constraining the size of the post-specific embedding $\boldsymbol{v}_i$, we force the model to try to store common linguistic features shared between texts from the same subreddit in $\mathcal{S}$. At decoding time, our model performs a lookup to find $\boldsymbol{v}_{subreddit}(i) \in \mathcal{S}$, the representation for the subreddit that post $\boldsymbol{p}_i$ was found in. We combine v_i with $v_{subreddit}(i)$, then apply another linear transformation $\boldsymbol{P}_c \in \mathbb{R}^{d \times f}$ to get our decoder input:

$$\boldsymbol{v}_{i(recon)} = \boldsymbol{P}_d^\intercal [\boldsymbol{v}_i \oplus \boldsymbol{v}_{subreddit}(i)] \tag{6.2}$$

The post embedding and the subreddit embedding can be combined either through addition (which requires the two representations to have the same size) or concatenation. We consider

both combination approaches, which we refer to as LOOKUP-ADD and LOOKUP-CONCAT respectively. $\boldsymbol{v}_{i(recon)}$ is then passed to an LSTM for decoding.

Validating reconstruction performance

To ensure that the subreddit embedding lookup models are able to maintain their ability to reconstruct texts, we compare the reconstruction performance of the subreddit lookup models with two baselines:

- **Basic autoencoder:** The basic autoencoder uses the same architecture (bidirectional LSTM with attention encoder) to construct a compressed text representation $\boldsymbol{v}_i$ for post $\boldsymbol{p}_i$. This compressed representation is then fed directly to the decoder as input:

$$\boldsymbol{v}_{i(recon)} = \boldsymbol{v}_i \tag{6.3}$$

  We use the basic autoencoder to give us a sense of the expected reconstruction performance of the lookup autoencoder under similar model parameter constraints.

- **Variational autoencoder (VAE):** Rather than representing each text as a fixed embedding, with a variational autoencoder [124], each post $\boldsymbol{p}_i$ is represented probabilistically in a k-dimensional latent space. The encoder outputs two vectors $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ representing mean and variance $\boldsymbol{p}_i$ in the latent space. The decoder then samples from the Gaussian latent space distributions during reconstruction:

$$\boldsymbol{v}_{i(recon)} \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \tag{6.4}$$

  We use the variational autoencoder to compare reconstruction performance on another model architecture that recognizes some latent shared elements between posts. Unlike the lookup models, however, these latent elements are not specifically tied to some known characteristic we are interested in, such as what subreddit a post came from.

All models use a 2-layer LSTM with a hidden dimension $f = 512$ as the encoder and decoder. We set the compressed size of the post embedding and the subreddit embeddings to 256. For the VAE baseline, we set the number of latent variables $k = 50$. We apply 20% dropout at training to prevent overfitting. All models are trained using Adam [123] with a learning rate of $1\text{e}{-3}$ for the first 8 epochs before switching to SGD with learning rate=0.1. Models are trained to convergence with an early stopping threshold of 5.

Table 6.13 shows the results of the reconstruction experiments. Both LOOKUP-ADD and LOOKUP-CONCAT achieved performances comparable to the basic autoencoder baseline. LOOKUP-ADD, however, had the best reconstruction performance, with a 1.52 BLEU improvement compared to the basic model. Thus, we focus our subreddit embedding analysis on LOOKUP-ADD for the rest of this section, though we see similar results for LOOKUP-CONCAT.

Subreddit embedding interpretability

To investigate whether there is evidence that the learned subreddit embeddings $\boldsymbol{v}_{subreddit} \in \mathcal{S}$ capture information related to political assertion usage tendencies, we adapt a procedure from

Model	BLEU	Perplexity
Basic	49.07	5.325
VAE	26.04	13.250
Lookup-Add	**50.59**	**4.555**
Lookup-Concat	49.17	4.757

Table 6.13: Autoencoder reconstruction results for our subreddit lookup models compared to basic autoencoder and VAE baselines.

	R^2(*k*=1)	R^2(*k*=3)	R^2(*k*=5)	R^2_{post}(*k*=5)
Care/Harm	0.1145	0.2890	0.3820	0.7951
Fairness/Inequity	0.1636	0.3600	0.5612	0.7614
Integrity/Dishonesty	0.0781	0.2278	0.2777	0.6073
Loyalty/Betrayal	0.1667	0.3397	0.4809	0.7081
Authority/Subversion	0.0542	0.2519	0.4169	0.8593
Sanctity/Degradation	0.1645	0.3298	0.4102	0.7137
Liberty/Oppression	0.1106	0.2004	0.3113	0.7482
Capacity/Incompetence	0.0717	0.2829	0.3369	0.8463
Physical Attributes	0.1172	0.2124	0.2880	0.5952
Character Traits	0.1342	0.2617	0.3664	0.6473

Table 6.14: R^2 values showing how the top k PCA factors in subreddit embeddings correlate with the assertion usage in that subreddit. We compare the R^2 values from using learned subreddit embeddings with R^2_{post}, the R^2 value from generating subreddit embeddings by averaging post-only embeddings from the same subreddit.

Shi et al. [195] for tracking whether LSTM units correlate with some feature of interest. Given a subreddit embedding, our goal is to predict the value of the subreddit's usage of a political assertion category using a weighted, linear combination of dimensions from the subreddit embedding. For our application, however, the number of subreddits examined and the number of dimensions in our subreddit embedding are close in scale to each other. Thus, running linear regression over the full subreddit embedding is likely to lead to overfitting. In order to address potential issues with overfitting, we first perform dimensionality reduction over the subreddit embeddings using principle component analysis (PCA) [89]. We consider the top components that explain up to 90% of variance in the subreddit embeddings. We then find the top k PCA components contributing to explaining assertion usage by selecting factors greedily in terms of how much they increase the R^2 value of the regression.

Table 6.14 shows the resulting R^2 values between the subreddit embeddings and level of assertion category usage in that subreddit for different values of k. Overall, the R^2 values for the political assertions categories are low, even at higher levels of k, with **Fairness/Integrity** as the only assertion category with more than 50% of the variance explained using the subreddit embeddings. This suggests that the learned subreddit embeddings by themselves are not tracking

information related to the usage of assertions within a subreddit. One possible explanation for this is that the lookup model assumes that there is some consistent and learnable difference between one label and another, due to the "lookup" embedding being shared across all posts with the same label. For example, in the multilingual translation model that the subreddit lookup model was based on, the lookup model essentially learns a representation for each target output language. Outputs from the same target language will have consistent surface-level differences from output from other target languages in the lookup, which allows the model to learn how to successfully generate texts over multiple languages. In our domain of political subreddits, however, the use of a particular assertion type may be tied to certain contexts, such as discussion around a certain candidate or issue, that might not always be relevant to the overall goals of a particular community. Thus, information about the use of assertions may not be stored in the shared subreddit embedding but instead be captured by the post-specific embedding $\boldsymbol{v}_i$

In order to examine whether information related to assertion usage within a subreddit is instead stored in the post-specific embeddings $\boldsymbol{v}_i$, we repeat the same regression and component selection process over the post-specific embeddings. We generate subreddit embeddings from the post-specific embeddings by averaging $\boldsymbol{v}_i$ over all posts p_i from the same subreddit. As seen in Table 6.14, the averaged post-specific embeddings have a much higher correlation with the assertion category usage in a subreddit, suggesting that the subreddit encoder model *does* capture some aspects of assertion usage but in the post-specific embeddings, rather than the subreddit embeddings. Overall, this seems to support our interactional perspective that while subreddits have certain tendencies towards how they make assertions about political entities, the way in how associations are invoked may be specific or unique to certain contexts, rather than shared across the entire community. Although we may see differences in assertion usage at the broader subreddit level, such as in Section 6.5.2, we should not expect these kinds of descriptive norms to be universally reflected by all content from the same community. From the perspective of modeling subreddit norms, this analysis provides evidence that capturing more subtle linguistic distinctions between communities requires more sophisticated approaches than sharing a representation of a community across all its content. Forcing all content from the same community to share a single representation may in fact wash out contextual variations out of the shared community representation. In future work, we hope to build models that better recognize when specific contexts should be shared to more effectively capture relevant linguistic phenomena.

6.6 Discussion

Throughout this chapter, our findings from examining differences in linguistic norms and tendencies across political subreddits push back against the predominant variationist view of political language. In our paired ideology ranking task annotation experiment, we found evidence that annotator perceptions of the ideology behind content from political subreddits was influenced, sometimes in systematic ways, by experiential factors related to political participation in Reddit. By incorporating insights from this annotation experiment, we develop a framework for categorizing different types of political assertions and associations, then use this framework to measure and model the use of assertions across different political subreddits. Overall, our analy-

ses highlight that linguistic variation always exists within how we define community boundaries. Subreddits with similar ideologies, for example, have different linguistic norms because they serve different purposes and interests. Similarly, linguistic distinctions between different subreddits may not be reflected in all the content produced by the community but rather only occur in certain contexts or situations. For NLP research, these findings highlight that future work needs to be more aware of the trade-offs between categorization and variation, from knowing the limitations of annotator generated labels to recognizing when and why simplifying assumptions about categories and labels are made.

6.6.1 Limitations and future work

Due to the relatively strict screening requirements and political sensitivity of the annotated content, we were only able to recruit a relatively small number of workers for our annotation study on the paired ideology ranking task. As a result, the number of annotators within each group and how many of each group were assigned to a particular question may vary. We leave a more controlled study of annotator variation with targeted recruiting across our annotator groups for future work.

In our analyses using our political assertion category labels, we rely on two different approaches to propagate our human annotated labels to a larger set. Both methods of label propagation have certain advantages and disadvantages compared to the other, with neither method covering all the ways one can invoke a particular assertion category. The DistilBERT model generally had moderate-to-good performance over most of the assertion categories but failed on the sparser appreciation categories. On the other hand, the lexical approach was more precise for these sparser categories but may not capture other ways of invoking those categories outside of what is already in the lexicon. Overall, we found that getting high agreement on the task of annotating how political assertions are invoked was difficult for both human annotators and our label propagation approaches, as the task requires understanding both political concepts in U.S. politics and Reddit-specific social cues. In future work, we hope to develop both clearer guidelines and more sophisticated label propagation approaches in order to improve our analyses.

6.6.2 Ethical considerations

As this chapter focuses on linguistic evaluation of political discussion on Reddit, our work involves handling potentially sensitive information about the political participation of Reddit users. All usernames were removed from our data in order to limit the impact of our work on users' privacy [59]. Text examples throughout this chapter are also lightly paraphrased to prevent posts from being traced back to users.

Due to the sensitive and potentially offensive content used throughout this chapter, the paired ideology ranking task was posted on MTurk as a HIT containing potentially containing adult content. Workers were paid $5 for completing our task, resulting in a $10/hour based on an estimated average time of completion from an initial pilot study. We were required to use worker ids in order to match worker answers on the paired ideology ranking task with their responses in the post-survey. After this matching process, however, worker ids were replaced with a random numeric id.

6.7 Conclusion

In this chapter, we present an examination of how to evaluate differences in descriptive linguistic norms in political communities. We first present an investigation of how the ideologies underlying political communities are perceived through an MTurk annotation experiment over a paired ideology ranking task. Using insights from our analysis of the annotation results, we develop a framework for categorizing types of assertions and associations that participants in political subreddits draw with key entities. We then use this framework to analyze fine-grained linguistic distinctions in political communities in three ways. Our analyses throughout the chapter push back against a variationist view of political language by providing evidence arguing against (1) the existence of an inherent ground-truth in annotation, (2) the simplification of Reddit political subcommunities as falling under the umbrellas of "the political left" and "the political right", and (3) the modeling of variations between subreddits as inherent and ubiquitous in that community. We want to emphasize that future work in language technologies should be more cognizant of similar issues in language variation within and across different communities. While NLP as a field requires us to make simplifying assumptions for modeling, we should not expect that the ultimate goal of the field is to sacrifice linguistic variation for label validity.

Chapter 7

Conclusion

> *A genuinely political society, in which discussion and debate are an essential technique, is a society full of risks.*
>
> –Moses Finley

In this chapter, I summarize the contributions of this thesis. I then discuss broader implications of my work for platform design and language technologies and reflect on future directions in both disciplines that may contribute to addressing abusive language in online spaces.

7.1 Summary of contributions

The goal of this thesis is to bridge the gap between language technologies and platform design when considering issues in abusive language moderation in two directions. Under the **evaluation** paradigm, I used techniques from language technologies to evaluate the social impacts of moderation in three case studies. In **Chapter 3**, I examined whether there was evidence of moderator bias in the Big Issues Debate group on Ravelry. Using a transition-based speech act model to account for more implicit, community-specific norm violations to control for user behaviors, I introduced a regression-based framework for investigating whether there was evidence that the moderators showed bias against users with minority viewpoints and or past antagonistic interactions with the moderation team. In **Chapter 4**, I used techniques from framing analysis to examine differences in user responses to announced policy changes regarding the quarantine feature on Reddit. Using a topical approach to agenda setting and framing, I analyzed how users on the left and right highlighted different priorities with regards to moderation policy, such as censorship, consistency, and harms. In **Chapter 5**, I directly examined the impacts of quarantines of two ideologically distinct political subreddits, r/The_Donald and r/ChapoTrapHouse. In addition to markers of activity and toxicity, I analyzed the visibility and discussion of issues within the quarantined subreddits, as well as the stability of linguistic norms within the quarantined and related subreddits, through the lenses of toxicity and moral value associations.

Under the **contextualization** paradigm, my goal was to reconsider assumptions regarding the relationship between identity labels and language found in computational social science work in NLP. Using insights from social theories and the study of online communities, my objective was to gain a better understanding of language issues that may play a role in defining what norma-

tive and abusive language looks like in online political discussion. Based on an interactional linguistics perspective, in **Chapter 6**, I first investigated how individuals perceive the ideology behind content from different political subreddits. I introduced a paired ideology ranking task, then analyzed what experiential factors on the side of the annotators and linguistic aspects of the paired posts contribute to differences in judgments. From these analyses, I developed a framework for categorizing common assertions and associations drawn with political entities by extending existing frameworks for value judgments to associations salient to political discussion. This framework was then used to analyze differences in assertion usage between political subreddits on the left and the right, challenging the notion of a monolithic left and right in Reddit political participation. Based on the findings of the subreddit-level assertion analysis, I evaluated an embedding-based model designed to capture common linguistic patterns within subreddits on its ability to distinguish differences in assertion usage across subreddits.

7.2 Insights and future directions for platform design

In this section, I first summarize the major themes derived from the insights from our evaluation case studies. I then discuss potential directions to consider for designing and applying moderation strategies in light of the implications from our analyses.

7.2.1 Reactive moderation and censorship

Throughout the evaluation case studies in this thesis, one overarching theme across the different strategies and communities was the association between moderation interventions and concerns over censorship. In Chapter 3, tensions arose in BID between minority viewpoint users and the moderation team over perceived unfairness in how their content was judged in accordance with the rules of the community. In Chapter 4, I found that while right-leaning users were more likely to highlight issues of censorship and free speech when discussing the quarantines of controversial subreddits, users on both the left and the right considered these issues as some of the most prominent concerns with quarantines, even as a non-removal-based intervention. In Chapter 5, my findings suggested that the quarantine of The_Donald led to increased accusations of censorship towards the Reddit administration in the right-leaning monitoring subreddit r/WatchRedditDie. While followup work [103, 104] has formalized how to evaluate whether politically biased decisions are occurring in moderation, my findings suggest that concerns and accusations over censorship are a common, if not inherent response to reactive moderation strategies in political discourse. Interventions that respond to behaviors after they occur are often perceived as a punishment towards users for their actions. In the context of political discussions, this is often interpreted as sanctioning users for expressing their voices. While this type of backlash is most commonly associated with ideological conservatives in both my work and the work of others [103, 104, 114], similar concerns were brought up by users on the ideological left in both Chapters 3 and 4, often tied to the censorship of individuals, rather than viewpoints.

In contrast with the reactive interventions in my case studies, proactive interventions may partly address some of the tensions from perceiving moderation as a form of punishment. The goal of a proactive intervention is to address potential issues with user behavior before the user

has a chance to post and be penalized for their produced content. Some research on proactive strategies for moderation has looked into identifying early markers of hate [82, 224] in conversations and communities, but the effectiveness and reception of proactive interventions remains relatively under-explored compared to reactionary strategies. One potential direction for a proactive moderation intervention, for example, could be giving explanations to a user describing how their content may be considered abusive or norm-violating, before the content is even posted. Content-removal explanations have been shown to reduce the odds of future content removal by providing feedback to users learning community norms [101], though the impact of explanations on deliberately abusive behaviors is unknown. Jhaver et al. [101] additionally found that there was no significant advantage to using human moderators to generate content removal explanations over bots, suggesting a place for the deployment of an automatic proactive explanation tool. An interesting experiment then would be to examine whether proactive detection and explanation of abusive behaviors may be effective at reducing norm violations within a subreddit. Another proactive intervention could be to use certain interface elements to prime users towards producing more positive and reflective content [191]. More research into how to design prompts and interface options that encourage reflection, however, is needed. A major disadvantage of using proactive interventions, however, is that they can be seen as more disruptive than helpful, such as in the case of Microsoft's Clippy, where proactive alerts by an automated system were seen as annoying [143]. These proactive approaches, thus, may be more useful as an optional intervention for well-intentioned users who wish to learn how to comply with community norms, rather than an overall check on (sometimes deliberately) abusive behaviors.

Another possibly fruitful direction in addressing negative responses to reactive moderation may be to draw upon different frameworks of justice in how to respond to online abuse [15]. Reactive moderation can be seen as a form of *retributive justice*, in which individuals who have committed an abusive act knowingly are punished in proportion to their moral wrongdoing. A moderation paradigm based on *restorative justice*, in contrast would focus on mediating conflicts and uplifting and empowering victims of abuse and harassment. Work on generating counter-narratives to hate speech [151] can be seen as falling under this paradigm. However, effective generation of counter-speech often relies on trained experts to craft counter-narratives and directly respond to sources of abusive content. As a result, moderation through counter-narratives may run into similar issues of scope and psychological toll as more commonly explored retributive approaches. While there is growing interest in NLP research in defining, collecting, and generating counter-speech and counter-narratives [206], similar to the retributive moderation strategies embodied by most current NLP approaches, more exploration is required into understanding whether counter-narratives can effectively address abusive behavior. While the contact hypothesis suggests that cross-group interaction can improve the relations and break down barriers between groups under particular circumstances, the unrestricted nature of participation online can potentially lead to increased hostility between groups [2], Gallacher et al. [65], for example, found evidence that online interaction between political with from opposing ideologies was linked to increased physical violence in offline protests planned online. Thus, further research is necessary to understand the situations where cross-group interaction and counter-narratives can be effectively used to mitigate abusive behavior.

An alternative direction for considering more restorative justice-oriented approaches to moderation could be to explore more non-hierarchical, community-centered initiatives to respond to

abusive language and uplift victims of harassment. These more decentralized, participatory efforts aim to empower individuals, including targets of abuse and their allies, by giving ordinary users more influence as to how to manage and respond to disruptive content. Interventions that aim to grant more decision-making power to regular users on how to respond to abuse can take many forms. Prior work, for example, has shown that individuals often rely on their peers to help mediate what kind of content they consume [128], and Mahar et al. [144] explicitly explores the use of friendsourcing to help victims of email harassment manage future email usage. One potential future direction, then, is to expand on these individual-oriented mediation approaches to help communities as a whole provide support to targets of abuse. Users who have experienced harassment within a subreddit, for example, could be assigned to community mediators or paired with community volunteer allies who help them navigate threads and submissions that may contain toxic content similar to the abuse they have experienced in the past. These types of strategies would ideally take place in an environment where platforms are willing to provide affordances for users to help each other curate and manage content.

Platforms could also potentially introduce affordances for alternative governance structures that give ordinary users the power to define and maintain the norms of their communities. Crowdsourced governance has been deployed in the past with the Tribunal system in League of Legends [129] and the Community Management Center in Weibo [130], where flagged content is judged by committees of users who have met certain behavioral standards within the community. For a potential Reddit analogue, one can imagine allowing users with high karma across communities to make judgments on reported content from different subreddits. A committee of these users who participate in different subreddits, then, could potentially make community-level intervention decisions, such as quarantining or banning subreddits where moderators refuse to enforce Reddit's content guidelines. Kou et al. [130], however, highlighted key concerns with crowdsourced moderation, including transparency issues with how the underlying systems work and lack of user trust towards crowdsourced judges and their decisions. We noted similar concerns in our analysis of BID, where users in the "Ask The Mods" thread commonly argued with the moderators over the transparency and objectivity of moderator decisions, despite having increased access to the moderators and their moderation processes through the thread itself. Thus, more research is needed in order to understand how to develop credibility for crowdsourced judges while still giving the opportunity for regular users to participate in or influence governance. An additional limitation of crowdsourced governance is that democratizing participation in moderation uncritically directly grants power to the majority, and thus, minority users may feel more disenfranchised under these systems, similar to the accusations of moderator bias in BID. Problematic norms and practices may also be reinforced in communities where controversy and abuse are the standard. As a result, future investigation of crowdsourced moderation systems should examine tradeoffs in how to weigh minority opinions within communities and platforms. An interesting future direction for non-hierarchical moderation that may empower minority voices, for example, is to explore whether crowdsourced users can perform more complex moderator roles than simply voting on moderation decisions. If crowdsourced users can be incentivized to participate in conflict mediation, consensus building, or other more transactive interactions [61] instead of making snap judgments on moderation outcomes, minority opinions may be able to have a more benevolent influence in a community.

Restorative justice efforts that challenge abusive language within a community and empower

the targets of harassment may aid in addressing abuse at scale in a less punitive manner than the retributive approaches explored in this thesis. Ultimately, however, interventions based on restorative justice, while intended to mediate conflict between harassers and their victims, must be carefully designed to not place the full burden of responding to harassment on victims themselves. We leave evaluations of the social impact of restorative justice moderation efforts to future work.

7.2.2 Deplatforming, linguistic entrenchment, and culture change

In two of the evaluation case studies in this thesis, I focused my analysis on user opinions towards and political impacts of quarantines, a community-level intervention in Reddit proposed as an alternative to the removal or *deplatforming* of a community. The effectiveness of deplatforming as a moderation strategy has come under recent scrutiny [1, 102], as growing concerns about the rise of alternative platforms and self-hosting [41, 174] have raised questions as to whether deplatforming can radicalize impacted users and communities. Users that have been deplatformed from conventional platforms like Facebook, Twitter, or Reddit, for example, are commonly believed to move to alternative sites, such as VK and Gab, where anger towards the deplatforming authorities is commonly expressed [78]. Early work into the large-scale impacts of deplatforming suggest that there are trade-offs at the community level [1] – deplatforming may increase both the activity and toxicity of displaced users but also decrease their ability to reach a wide community audience. As such, alternatives to community-level deplatforming, such as quarantining, may still have a place as a moderation strategy. From the analyses from Chapter 5, however, I found evidence that quarantines remain ineffective at addressing controversial content on Reddit, likely due to the stability of community linguistic norms on Reddit. Without any incentive to change existing practices within a community, quarantined subreddits are likely to continue carrying out the behaviors they displayed before the quarantine. Thus, when considering alternatives to deplatforming, more work is required in understanding how to institute or incentivize cultural changes within controversial communities.

One area where we can potentially derive insights into how to change the norms of a community is the field of organizational behavior studies. From studies of culture change in organizational behavior, culture change in a workplace is often driven by drastic changes in leadership, with more radical changes in leadership often necessary to address strong resistance to change [116, 176]. Using this analogy on Reddit, an equivalently top-down approach to changing the leadership in a workplace could be to enforce replacing the moderators in a controversial subreddit that do not adhere to Reddit's community standards. However, because moderators are often the most visible and influential users in a subreddit, the Reddit administration stepping in and requiring that a subreddit replace its moderators will likely also be seen as repressive [179]. Subreddits, however, are often open to drastic shifts in leadership without top-down intervention. Reddit users, for example, note that subreddits with relatively inactive moderation teams have been the target of coordinated alt-right takeovers through gradual replacement of moderators.[1,2] Research into the coordination efforts behind these takeovers can provide insight not only

[1] https://www.reddit.com/r/AgainstHateSubreddits/comments/facauf/far_right_takeover_of_runitedkingdom_that_the/

[2] https://www.reddit.com/r/AgainstHateSubreddits/comments/bbpor4/ysk_the_

into how to prevent moderator change mechanisms from being abused but also whether a similar process could be used to encourage communities to take on influencers that adhere to Reddit community standards. Thus, examining these moderator takeovers may be useful for considering how to institute leadership-driven changes subreddit culture.

Another potential direction for institutionalizing culture change within a controversial community is by changing the relationship between the user and how normative content standards are maintained. In Reddit, for example, users can influence the visibility of content based on its upvote and downvote system and users build karma based on how other users value their content. Thus, users attuned to community standards are able to reinforce existing and potentially controversial linguistic norms through both their own content and how they vote on content within a subreddit. Because of the link between upvotes on content and a user's karma score then, users are in fact incentivized to adjust their content to what is popular and accepted within a community. As a result, Reddit's voting system and karma scores have received criticism for not actually encouraging high quality, reflective content. Richterich [175], for example, found that users perceived that Reddit's econometric system seemed to inhibit innovative content, with particular criticism directed towards *karmawhoring*, the submission of low quality content that appeals to the lowest common denominator. Potential interventions that redefine the relationship between users, content, and Reddit's system of value attribution, then, may be useful for not only increasing the visibility of non-normative content but also incentivizing users to submit higher quality content. Some strategies that could disrupt this influence, then, could be to adjust the sorting algorithms available for displaying content for certain subreddits or place user-level limitations on voting behavior to encourage more reflective engagement with content. These potential design frictions, however, may lead to user frustrations with participation functionality within a targeted community and thus, unintentionally drive users away from the platform anyway. More insight into how social votes are interpreted by users [165] and how sorting systems can be used to reinforce or weaken normative behaviors may allow for more finely tuned interventions for encouraging culture change.

A major caveat with using cultural change efforts as an alternative to deplatforming is the question of whether or not a subreddit *can* even change its dominant culture. There may, in fact, be communities that so consistently violate the content guidelines of a platform such that the best action to take is to deplatform that community. Various factors may influence whether or not a given community is likely or able to change its normative behaviors vs. adhering to existing malicious norms. While outside factors, such as threats to existence, may successfully encourage some communities to change their culture, others may be more resistant to change. We can see this, to some extent, with the case study in Chapter 5, if we interpret quarantines as a warning to a subreddit that could potentially be banned in the future, as neither The_Donald nor ChapoTrapHouse showed significant changes in linguistic tendencies post-quarantine. In the case of our two subreddits of interest, this may suggest that subreddits centered around cults of personality or subreddits that hold strong anti-establishment values are less receptive to quarantines as a threat against existence. Communities where the moderation team regularly do not moderate behaviors widely considered unacceptable based on Reddit's content policy may also be less likely to be receptive to oversight from the Reddit administration. In the recent incident

takeover_of_rgenz_by_the_racists_is/

where the controversial gaming subreddit r/TheLastOfUs2 was shown to have engaged in systematic, unchecked harassment against a well-known streamer[3] for example, the moderator team continued to engage in coordinated harassment and obstruction after the subreddit came under increased scrutiny.[4] Thus, it may be worth examining factors that contribute to a community's willingness to change, from cultural factors, like value attribution to authority vs. liberty, to structural factors, such as how hands-on or influential the moderation team is.

7.3 Insights and future directions for language technologies

In terms of insights and future directions for NLP and language technologies, I will revisit three common assumptions made in computational social science with regards to the relationship between language and identity. For each assumption, I will state the assumption, describe how parts of this thesis challenge the assumption, then provide some recommendations for how to address issues with the assumption.

7.3.1 Reconsidering definitions

A broad assumption in much of the work in abusive language detection is that definitions of abuse and other related linguistic phenomena can be held universally across communities. While different definitions and subcategories of abusive language, such as cyberbullying, hate speech, and interpersonal abuse, have been discussed in recent years [19, 112, 212], there remains a push within NLP to develop models that generalize these different linguistic phenomena over a variety of domains and communities, with weak cross-domain performance seen as a major failing of current models [117, 203]. Indeed, some of the analyses in this thesis also relied on general purpose definitions of linguistic phenomena, such as the use of Perspective API to measure the prevalence of toxicity across different subreddits in Chapter 5.

However, throughout this thesis, I also noted issues with relying on generalized definitions of linguistic phenomena. In the analysis of BID in Chapter 3, for example, I found that BID's unique community norms made it difficult for us to apply conventional approaches for abusive language detection to estimate high-risk behaviors in BID. From BID's irreverent, informal environment that encourages profanity to more subtle, implicit forms of norm violating behavior related to debating practices, such as "debating the individual, not the topic", the types of offending behaviors on BID align poorly with general purpose definitions of abusive language. As a result, I relied on a specialized intent-based transitional model to operationalize the unique definition of abuse in BID for the analysis. Similarly, in Chapter 6, the perception of ideology behind content from Reddit was influenced by the differences in the culture of political discussion on Reddit in comparison to general formal conceptions of the political ideology in U.S. politics. On Reddit, where the predominant left-leaning communities were further left than the general conception of left in U.S. politics, for example, antagonism towards the Democratic party, the predominant

[3] https://www.reddit.com/r/SubredditDrama/comments/oqsok9/rthelastofus2_goes_private_after_a_user_is/

[4] https://www.reddit.com/r/SubredditDrama/comments/or9quq/rthelastofus2_mods_do_their_best_damage_control/

left-leaning party, was commonplace and expected. Understanding the unique norms, beliefs, and governance practices of the community where conversation is taking place, then, is crucial for the development of tools that can effectively aid moderators in real-world online settings [60].

While the goal of mitigating generalized abuse across different spaces on the Internet is well-intentioned, human labor in moderation is centered around a wide variety of norm violating issues that can vary across different platforms and communities. My recommendation for future work in NLP, then, is to explore and broaden the space of what can fall under the definition of various linguistic phenomena across different online communities, such as abusive language or ideology. One direction for addressing community-specific language and norms for abusive language detection purposes could be to draw from observations of the typology of norm categories from Chandrasekharan et al. [33]. Under this framework, norms across Reddit can be classified as *macro norms* or norms that are almost universally shared across communities, *meso norms* or norms that are shared across related communities, and *micro norms*, norms that are specific to certain communities. While current general purpose models of abusive language detection may eventually be useful for addressing macro norm violations, which tend to align with more explicit, severe forms of abuse with potentially greater traumatic impact, more research within NLP could be directed at how to handle meso and micro norm violations, which make up a substantial portion of a moderator's day-to-day tasks. Because meso and micro norms are less well-studied in NLP and not all signals of moderation activity are immediately apparent, there is less observational data for these types of violations to use for model training. Thus, domain adaptation and finetuning approaches that take into account shared elements between related clusters of subreddits, such as topic, formality, or belief systems in political communities, may be useful as a first step into understanding the meso and micro norms of community. Guided by this norm typology, then, we could aim to build models that could be finetuned to accommodate different definitions of abuse based on community properties, such as rules or moderation signals from neighboring communities. These types of adaptable models, which can adjust to the variety of norms held across different communities, may more successfully be able to address abuse across online spaces.

7.3.2 Reconsidering distinctions

In Chapter 6 of this thesis, I discussed two different views for considering the relationship between language and identity labels in sociolinguistic contexts. NLP as a discipline has primarily relied on a variationist or essentialist view of the relationship between labels and texts. Under this view, we assume that there are consistent, structured differences in how different identity labels sound. This perspective is often reinforced by the limitations of the commonly used modeling processes. As a field, NLP has to make simplifying assumptions about what categories to consider and how variation occurs across categories in order to fit issues of language and identity into existing machine learning paradigms. This process inherently involves figuring out where to draw the distinction boundaries between identity labels, and as a result, we must, in some form, place limitations on what the key variations in identity are in relation to the language issues we are interested in. While there is no way to avoid the fact that we have to make reductionist assumptions in order to model language, NLP as a discipline should be more aware of the limitations of these assumptions, especially with models for more socially-oriented, real-world

applications.

A common assumption in NLP that could be more carefully addressed regarding the relationship between identity labels and language is the assumption that the categorical distinctions we define onto a domain are inherently truthful and socially relevant for the issues we want to study. Work on partisanship [155], ideology prediction [169], and framing and agenda setting [207], for example, often makes the simplifying assumption that the important identity difference in the political domain is a binary distinction between the left and right. While we relied on a simple left-right distinction in our paired ideology ranking task in Chapter 6 as an easier distinction that we hoped third-party annotators could pick up on, overall, we argue against the idea that this simple left-right distinction is inherently truthful and useful in the context of Reddit political discussion. The low annotation agreement results on the paired ideology ranking task may suggest that the left-right distinction is not a particularly salient distinction in most political communities on Reddit. Additionally, in the analysis of political assertion usage in Chapter 6, we found key differences in how subreddits with the same "left" or "right" label used assertion types. Assertion usage on Reddit was tied to the specific goals and interests of a subreddit, rather than its ideological affiliation. Under the assumption that the main political difference on Reddit is between a monolithic "left" or "right", we may have missed important distinctions in interests held by different subreddits that seemingly belong to the "same side", such as the specific ways different left-leaning subreddits supported or attacked different candidates or issues based on their main goals. In the case of Reddit, the subreddits themselves were a socially relevant level of distinction, likely due to the fact subreddits are created when there is enough demand for a community with different goals or interests than what is already available. As a result, it may sometimes be misguided to come in with our own labels as a notion for what is a socially relevant identity distinction within a particular community or domain.

While some work specifically in the political domain has considered more fine-grained political distinctions by incorporating moderates and third parties as identity labels of interest [95, 172, 196], these analyses still rely on the assumption that these categories are discrete and reflect an social distinction of interest – everything within a category is assumed to be similar to each other and distinct from other categories in a linguistically or socially relevant way. Imposing our own distinctions on a community, however, may lead to biases towards a researcher's own conceptions about how identity is shaped in a community, rather than what is truly important to the community itself. One possible future direction for improving on how we define identity labels, then, could be to find methods for discovering what the important social distinctions are in a given context. Bottom-up, unsupervised approaches may be useful for limiting preconceived notions for the important distinctions in a conversational context. Darwish et al. [43], for example, proposed an unsupervised approach for detecting stances on Twitter over six controversial topics based on dimensionality reduction and clustering. Distinctions found through these unsupervised methods, however, are often difficult to interpret and label, as they can reflect potentially any underlying difference patterns between instances in the data. Thus, current unsupervised approaches for finding relevant distinctions and categories may possibly be refined by also integrating ideas from social theories about deriving social meaning from distinction [18, 20] when categorizing instances in the clustering process. Another possible direction would be to work with data with naturalistically labeled distinctions, such as abusive language datasets aligned with actual moderation decisions [28] or identity labels based on self-presentation [220].

By doing this, we rely less on categories defined by researchers, who may be outsiders in a community or domain, and more on labels that are marked as important by actual stakeholders in a community.

7.3.3 Reconsidering identity

A related assumption in NLP under the variationist or essentialist position is that given different identity categories of interest, it is often assumed that the distinctions between these categories are consistently and inherently reflected in their language. In the context of political discussion, for example, KhudaBukhsh et al. [120] explored a simple translation-based approach for detecting differences between comments on CNN and Fox videos based on the assumption that viewers of both channels consistently spoke different political languages from each other. While the paper was able to find some systematic differences between their left-leaning and right-leaning sources, the strict assumption led to the method primarily finding obvious, surface-level differences between the left and right, such as swapping presidential candidates and organizations.

The concept of linguistic agency [73] challenges the idea that identity labels, such as one's ideological beliefs, are predictably and consistently presented in text. Based on an author's social goals for participating in discussion, it may not always be contextually relevant to project a strong impression of their identity labels. For example, users that consistently interact with each other may develop an impression of each other's beliefs and worldviews, such that it is no longer necessary to strongly project one's political ideology to engage in political discussion. We see this in BID, where users are able to identify each other as liberal or conservative in threads centered around moderation, rather than political or social issues. This occurs even without the use of identity markers in a user's profile or forum signature, as users build an image of other users over time based on previous debates they have engaged in. In the context of Reddit, the specific subreddit that a conversation takes place in also can reveal a lot about a user's views without them needing to specifically write out their beliefs. On the other hand, there are often more subtle, less consistent distinctions between categories of interest that may only occur in specific contexts. For example, debates between supporters of different Democratic primary candidates may reveal subtle differences in values between users that may not be as visible in a thread dedicated to general opposition against Trump. Operating under the assumption that language features must be consistent and inherent within a category, however, may limit us to detecting the most obvious language differences between identity labels, while more subtle distinctions are obscured or seen as noise.

From a more interactional perspective, I argue that linguistic variation outside of expected parameters, such as identity label boundaries, is not necessarily noise, especially for the analysis of language tied to social interaction. When considering issues of "noise" when analyzing model results, then, we need to be more aware of cases of systematic variation that are more subtle than our initial assumptions, especially in regards to different social contexts in which conversations take place. Thus, when analyzing modeling results, we must be more aware of what simplifying assumptions we make in the model and how that affects the range of linguistic phenomena that could potentially be detected. From a modeling perspective, then, one potential next step to address some of these issues may be to build models that are aware of different situational contexts where the presentation of labels may be more or less important. Incorporating additional

contextual information, such as user history or threaded conversational structure, explicitly into a model architectures may help pick up on more subtle or context-specific patterns of distinction that may be missed under an essentialist view of language and identity. Ultimately, however, as a discipline, NLP must recognize the inherent simplifying limitations in tracing the relationship between language and identity as the different and diverse contexts in which conversation can take place will change the nature of this relationship.

Appendix A

Paired ideology ranking task instructions

In this task you will be asked to make judgments about the political viewpoints of people who make statements about political entities (e.g. candidates, demographics, organizations, and ideologies). You will be presented with two texts that discuss the same political entity (highlighted). One of these texts discusses the entity from a left-leaning perspective while the other discusses the entity from a right-leaning perspective. Your task is to identify which of the two posts sounds like it more closely matches the particular perspective asked about in the prompt. You will not need to provide justification for your judgments, but we provide them in the examples to highlight some examples of cues that may help you in making a decision. For this task, you will be presented with 26 pairs of posts.

Example 1

Select the post that is more likely to come from an individual with a **left-leaning** perspective in how **Bernie Sanders** is portrayed.

Post 1: **Bernie Sanders** is the only candidate who we can trust to address this.
Post 2: **Bernie** has no idea what he 's talking about.

Answer: Post 1
Justification: Post 1 expresses an openly positive attitude towards Bernie, especially compared to Post 2.

Note that not all posts are easily separated based on attitudes expressed towards the entity. For these difficult cases, read the posts to yourselves out loud, and to the best of your ability, using as many contextual clues as you can to imagine what kind of person would say the things in each post. We provide some examples of these more difficult pairs to highlight some other cues that may be useful in differentiating posts.

Example 2

Select the post that is more likely to come from an individual with a **right-leaning** perspective in how **Donald Trump** is portrayed.

Post 1: I agree that a lot of **Trump's** policies have been terrible, but they haven't been the end of the world like the Dems suggested.
Post 2: **Trump's** policies have been nothing but cruel and criminal.

Answer: Post 1
Justification: The speaker in Post 1 references "the Dems" as a group they are not part of, positioning themselves in opposition to a Democratic/left-leaning perspective.

Example 3

Select the post that is more likely to come from an individual with a **right-leaning** perspective in how **Bernie Sanders** is portrayed.

Post 1: **Bernie** didn't invent 99% of the stuff you give him credit for. .
Post 2: If you believe a guy that promises "MUH FREE STUFF" like **Bernie Sanders**, you're a moron.

Answer: Post 2
Justification: Post 2 takes on a more informal, direct tone, similar to Trump's tweets and speeches.

Appendix B

Post-survey questions (paired ideology ranking task)

B.1 Political ideology

1. Please indicate where you identify on the liberal-conservative spectrum.
    - Liberal
    - Somewhat liberal
    - Moderate
    - Somewhat conservative
    - Conservative
    - I don't know
2. Please indicate how strongly you identify with the following U.S. political parties.
    - Parties
        - Democratic Party
        - Republican Party
        - Libertarian Party
        - Green Party
        - Constitution Party
        - Democratic Socialists of America
        - Reform Party
    - Responses
        - I do not identify with this party
        - Somewhat identify
        - Identify
        - Strongly identify

    - I don't know

B.2 News access

1. On average, how often did you check the news related to the 2020 presidential election in the U.S. in the past year?
    - Never
    - Less than once a month
    - A few times a month
    - Once a week
    - Several times a week
    - Once a day
    - Several times a day

B.3 Reddit familiarity

1. On average, how often have you visited Reddit in the past year?
    - Never
    - Less than once a month
    - A few times a month
    - Once a week
    - Several times a week
    - Once a day
    - Several times a day
2. On average, how often have you posted content to Reddit in the past year?
    - Never
    - Less than once a month
    - A few times a month
    - Once a week
    - Several times a week
    - Once a day
    - Several times a day
3. Please indicate your familiarity with the following subreddits (listed in Table 6.1).
    - I have never heard of this subreddit
    - I have heard of but never accessed this subreddit

- I have accessed or posted on this subreddit at least once
- I sometimes access or post on this subreddit
- I often access or post on this subreddit

Appendix C

Political assertions framework annotation guidelines

C.1 Main instructions

Goal: For this task, you will be presented with a comment mentioning a highlighted entity. Your goal is to identify and categorize the assertions the comment makes towards the highlighted entity. The general workflow for our task is as follows (more detailed descriptions, examples, and boundary cases are given in Section C.2):

- **Affect:** What is the sentiment that the comment expresses towards the highlighted entity? (-1 for negative, 0 for neutral/ambiguous, 1 for positive)
- **Assertion types:** Identify the specific assertions and value judgments the author makes either directly or indirectly towards the highlighted entity (e.g. Trump forces people to sign NDAs, Clinton is corrupt). From these assertions, identify whether they invoke the specific category of assertion towards the highlighted entity. For each column, enter 1 if you answer Yes to any of the questions for that assertion type and 0 if otherwise. Note that these categories can be invoked by either pole for that dimension (e.g. Enter 1 for Care/Harm if the comment references either the care or harm caused by the entity). If it is unclear/ambiguous if the comment invokes the category, enter 0:
  - **Care/Harm:** Does the comment invoke the compassion and benevolence of the entity or the physical, material, or emotional benefits of their actions or beliefs towards others? Does the comment discuss the physical, material, or emotional harm caused by the entity, their actions, or beliefs towards others?
    - **Examples:** *understanding, concerned, personally invested, sensitive, benefits, helps, cruel, violent, killed, stole, fucked over, murderer*
  - **Fairness/Inequity:** Does the comment reference fairness and/or unfairness experienced by or carried out by the entity? Does the comment discuss issues of discrimination towards certain groups (e.g. racism, sexism, xenophobia) related to the entity? Does the comment discuss retributive justice carried out by or experienced by the entity?

- **Examples:** *fair, just, impartial, punished, reciprocal, indictment, prejudiced, biased, exclusionary, bigoted favoritism, racism, homophobia, xenophobia*

- **Integrity/Dishonesty:** Does the comment discuss the honesty or straightforwardness of conduct of the entity? Does the comment discuss issues of dishonesty, corruption, or deception related to the entity?
  - **Examples:** *honest, transparent, corrupt, bribing, underhanded, criminal, collusion, conspiracy, liar, cheater, backroom deals, secrecy, blackmail*
- **Loyalty/Betrayal:** Does the comment define the boundaries of an entitys group membership (e.g. entity X is/is not a member of group Y, entity X has the following members Y)? Does the comment discuss the entitys allegiance, loyalty, or adherence to their in-group and their beliefs and values? Common groups include political parties, alliances, nations, or families.
  - **Examples:** *loyal, member, solidarity, patriot, traitor, treason, shill, considers themselves a libertarian, patriot*
- **Authority/Subversion:** Does the comment discuss the entitys assigned social roles, participation in social contracts and exchange, or deference to authority and tradition? Does the comment discuss violations of social hierarchy committed by or towards the entity or explicit usage of the entitys power/influence? Does the comment discuss the source of the entitys authority/influence or how the entity gives authority/influence to others? Does the comment discuss the entitys obedience towards or from another entity?
  - **Examples:** *lawful, defer, fall in line, employee, authorities, party leadership, business contract, appeasement, collaboration, pushy, insurgent, dissenting, protest, refusal*
- **Sanctity/Degradation:** Does the comment invoke purity or cleanliness, especially in regard to issues of religion, sex, or drug use? Does the comment associate the entity with an undesirable individual or group such that the taboo nature of the undesirable entity is specifically emphasized? Does this comment use metaphors of disgust, uncleanness, or contamination for the entity?
  - **Examples:** *sacred, holy, chaste, civilized, profane, taboo, defile, pedophile, rapist, Nazi, commie, scum, pro-life, junkie, slavery, cancer, vermin*
- **Liberty/Oppression:** Does the comment discuss what rights are available to or provided by the entity? Does the comment discuss behavioral restrictions/control placed on or by the entity from or towards other entities?
  - **Examples:** *free, libertarian, pro-choice, rights, authoritarian, banned, silenced, gun control, oppression, tyrant, despot, censorship*
- **Capacity/Incompetence:** Does the comment discuss qualities related to the entitys competence or perceived competence at achieving their goals/success? Does this comment discuss whether or not the entity was able to achieve their goals?
  - **Examples:** *intelligent, strong, experienced, electable, versatile, independent,*

popular, creative, stable, reliable, responsible, consistent, motivated, resilient, more successful, won more votes, stupid, slow, feeble-minded, insane, hypocrite, erratic, cowardly, weak-willed, sleepy, lost

- **Physical Attributes:** Does the comment discuss the physical traits (e.g. appearance, voice) of or associated with the entity?
  - **Examples:** *tall, short, ugly, skinny, dresses well, orange man, balding, whiny voice*
- **Character Traits:** Does the comment discuss the personality or culture associated with the entity?
  - **Examples:** *warm, extroverted, likeable, welcoming, values discussion, eager, narcissistic, bunch of snowflakes, smug, creepy*

C.2 Additional details

Assertion types: Our assertions of interest can be divided into two groups. You will not be annotating for these two groups, but they may be useful in distinguishing some of the assertion types. Under these two groups are the 11 assertion types of interest you *will* be annotating for:

- **Policy-oriented judgments**: These categories are invoked in order to express a judgment on the beliefs, opinions, policies, and political actions associated with an entity.
  - **Care/Harm:** The original moral foundation of Care is concerned with kindness and compassion, while Harm is concerned with cruelty and aggression towards others. It is believed to have been derived from attachment systems and dislike of the pain of others. This assertion category contains posts that invoke either of the two dimensions in relation to physical, material, and emotional benefits or harms caused or experienced by the highlighted entity.
  - **Fairness/Inequity:** Derived from the original moral foundation of **Fairness/Cheating**, Fairness is primarily concerned with issues of justice and equality, while Inequality is concerned with issues of injustice and unfair treatment. References to discrimination towards groups, such as racism, sexism, xenophobia, etc., and application of retributive justice fall under this category.
  - **Integrity/Dishonesty:** Related to aspects of the **Fairness/Cheating** moral foundation, this category focuses more on issues of honesty, transparency, and straightforward conduct vs. issues of underhandedness, corruption, and secrecy. We separate this category out from the original moral foundation due to its high prevalence in entity associations, especially in regards to the conduct and character of specific political candidates.
  - **Loyalty/Betrayal:** This category is concerned with defining the boundaries of an in-group vs. an out-group, as well as how members of a group adhere to their in-groups goals and values. Groups can be all manner of coalitions, such as political parties, nations, alliances, and families. A comment that either defines group boundaries in relation to a highlighted entity (X is/is not a Y) or talks about a highlighted entitys

allegiance to a group also invokes this category.

- **Authority/Subversion:** Similar to the original moral foundation, this category is concerned with social roles and social hierarchy and is concerned with leadership, followership, and adherence to social roles, contracts, and exchanges. This encompasses issues involving deference to legitimate authority, respect towards traditions, role assignment, and violations of social roles and hierarchy.
- **Sanctity/Degradation:** This category is concerned with purity, sacredness, and taboo. It is shaped by the psychology of disgust and contamination and commonly underlies religious notions of what is pure/moral/clean vs. immoral/taboo. As such, issues involving sexual issues, drug use, and metaphors of dirtiness are likely to invoke this category. In politics, this is also commonly used to draw associations with undesirable groups, such as Nazis and communism in the U.S. One thing to note, however, is that simply referencing an undesirable group does not necessarily invoke this category. A comment that discusses facts about communism or even supports communism does not invoke this category. However, if a comment references or invokes the taboo nature of communism in U.S. politics, this category is invoked.
- **Liberty/Oppression:** This category is concerned with the desire not to be dominated by others/restricted from acting in certain ways. It often overlaps heavily with the **Authority/Subversion** category but is more concerned with specific discussion about rights and restrictions, rather than social hierarchy and roles. In cases where there is overlap between these categories, mark both categories with a 1. For examples where there is a distinction between the two categories:
  - **Authority/Subversion-only:** There will always be **socialist** rebels.
  - **Liberty/Oppression-only:** **Biden** does not care about minority rights.
- **Capacity/Incompetence:** This category is concerned with whether an entity is capable of carrying out their goals. It focuses mostly on statements evaluating the competence of the entity and whether the entity is able to successfully achieve their aims and objectives. Assertions in this category can include elements related to either *capacity* (how capable something is) and *tenacity* (how reliable something is) from Martin and White [147].

- **Appreciation:** These categories are related to characteristics associated with the highlighted entities that do not comment on policy, viewpoint, or political actions by an entity but still may give a value judgment towards the entity. These can be applied to describe attributes of individual entities and the culture/image surrounding groups/organizations.
  - **Physical Attributes:** This category is invoked when a post provides commentary on the physical appearance or image of the highlighted entity. This includes description of the physical traits of the entity themselves, including their appearance and voice, or how they dress or present themselves.
  - **Character Traits:** Other non-physical traits associated with the entity that are not explicitly tied with policy but attribute some sort of value judgment towards the highlighted entity. This includes direct comparisons with other entities that indicate a

value judgment but with non-specific connotations that do not fall under any other category.

  - A statement such as "X is just like Hillary" would fit under this category, as it makes a direct comparison to another entity with a direct value judgment, but the specific connotations of the comparison (e.g. Competence? Trustworthiness? Policy?) are unclear and thus, none of the other categories can be checked off.

When to count assertions: We count assertions if the assertion is invoked by, in comparison to, or in direct interaction with the highlighted entity.

- **Example to count:** Well yeah, Obamas not white and not a **conservative** so obviously he's a socialist communist gay Muslim who's actually from Kenya.
  - In relation to Obama, conservative is contrasted with "socialist", "communist", "gay", "Muslim", and Kenya, which are used to intentionally associate Obama with taboo/undesirable groups (socialist, communist) and to question Obamas patriotism to the U.S. (Kenyan), both the **Sanctity/Degradation** and **Loyalty/Betrayal** categories are invoked and should be marked as 1, even though the author only provides a weak metalinguistic judgment on "conservative".
- **Example to not count:** I even see Bush being pushed as someone not-as-bad-as **Trump**. "Look at the poor fool, he was just used by Cheney". Sorry, don't buy it.
  - While the author makes an assertion that the highlighted entity Trump has caused harm, invoking the **Care/Harm** assertion category, the comments about being controlled/used by Cheney are directed at Bush and not connected with Trump. Thus, **Authority/Subversion** should not be marked 1.

Under our framework, a single statement may invoke multiple assertion categories. For example, the term "lynching invokes both the Care/Harm category, as a form of murder, and the Fairness/Inequity category, as lynchings are extrajudicial and have strong racial connotations in the U.S. In these cases, all the relevant categories should be marked as 1.

Bibliography

[1] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Understanding the Effect of Deplatforming on Social Networks. In *13th ACM Web Science Conference 2021*, pages 187–195, 2021. 7.2.2

[2] Yair Amichai-Hamburger. The contact hypothesis reconsidered: Interacting via Internet: Theoretical and practical aspects. *Psychological Aspects of Cyberspace: Theory, Research, Applications*, 2008. 7.2.1

[3] Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. *arXiv preprint arXiv:1806.05521*, 2018. 6.5.3

[4] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer, 2018. 2.1

[5] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 45–54, 2019. 2.2.1

[6] Kent Bach and RM Harnish. *Communication and Speech Acts*. Harvard UP, 1979. 3.4.4

[7] David Bamman and Noah A Smith. Open extraction of fine-grained political statements. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 76–85, 2015. 6.2.2

[8] Michele Banko, Brendon MacKeen, and Laurie Ray. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, 2020. 2.1

[9] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics, 2019. 2.2.2

[10] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit Dataset. *arXiv preprint arXiv:2001.08435*, 2020. 4.3, 5.3, 6.2.1

[11] Max H Bazerman, George Loewenstein, and Don A Moore. Why good accountants do bad audits. *Harvard Business Review*, 80(11):96–103, 2002. 3.2, 3.6.1, 3.6.2

[12] James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International journal of epidemiology*, 46(1):348–355, 2017. 5.4.1

[13] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009. 3.4.4

[14] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and its consequences for online harassment: Design insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–19, 2017. 1.1, 2.3.1

[15] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. When online harassment is perceived as justified. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018. 7.2.1

[16] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 2003. 4.4.1

[17] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008. 4.5.2

[18] Pierre Bourdieu. *Distinction: A social critique of the judgement of taste*. Harvard University Press, 1984. 7.3.2

[19] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, 2019. 2.1, 7.3.1

[20] Marilynn B Brewer and Sonia Roccas. Individual values, social identity, and optimal distinctiveness. *Individual self, relational self, collective self*, 2001. 7.3.2

[21] Thomas Lee Budesheim and Stephen J DePaola. Beauty or the beast? The effects of appearance, personality, and issue information on evaluations of political candidates. *Personality and Social Psychology Bulletin*, 20(4):339–348, 1994. 6.4

[22] Dallas Card, Justin Gross, Amber Boydstun, and Noah A Smith. Analyzing framing through the casts of characters in the news. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 4.5

[23] Mark Carman, Mark Koerber, Jiuyong Li, Kim-Kwang Raymond Choo, and Helen Ashman. 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*, 2018. 4.4.2

[24] Jordan Carpenter, Daniel Preotiuc-Pietro, Lucie Flekova, Salvatore Giorgi, Courtney Hagan, Margaret L Kern, Anneke EK Buffone, Lyle Ungar, and Martin EP Seligman. Real men dont say cute using automatic language analysis to isolate inaccurate aspects of

stereotypes. *Social Psychological and Personality Science*, 8(3):310–322, 2017. 6.2

[25] Jordan Carpenter, Daniel Preotiuc-Pietro, Jenna Clark, Lucie Flekova, Laura Smith, Margaret L Kern, Anneke Buffone, Lyle Ungar, and Martin Seligman. The impact of actively open-minded thinking on social media communication. *Judgment and Decision Making*, 13(6):562, 2018. 6.2

[26] Stephen L Carter. *Civility: Manners, morals, and the etiquette of democracy*. Basic Books (AZ), 1998. 3.1, 3.2

[27] Samuel Carton, Qiaozhu Mei, and Paul Resnick. Extractive Adversarial Networks: High-Recall Explanations for Identifying Personal Attacks in Social Media Posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. 2.2.2

[28] Michael Castelle. The linguistic ideologies of deep abusive language classification. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 160–170, 2018. 2.2.1, 7.3.2

[29] A Chadwick. *Internet Politics: States, Citizens, and New Communication Technologies*. Oxford UP, 2006. 3.1

[30] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1201–1213, 2016. 2.3.2

[31] Eshwar Chandrasekharan and Eric Gilbert. Hybrid approaches to detect comments violating macro norms on Reddit. *arXiv preprint arXiv:1904.03596*, 2019. 2.2.1

[32] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *CSCW 2017*, 2017. 1.2, 2.3.2, 3.4.4, 4.2, 4.5, 5.1, 5.2, 5.3.1, 5.3.2, 5.6.1

[33] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro Scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, 2018. 1.1, 7.3.1

[34] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–30, 2019. 2.3.3

[35] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *arXiv preprint arXiv:2009.11483*, 2020. 5.1, 5.4.2, 5.6.1, 5.8.1

[36] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems (NIPS)*, 2009. 4.5.3

[37] Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. Trajectories of Blocked

Community Members: Redemption, Recidivism and Departure. *arXiv preprint arXiv:1902.08628*, 2019. 4.2

[38] Anna Chung. How Automated Tools Discriminate Against Black Language . *One Zero. https://onezero.medium.com/how-automated-tools-discriminate-against-black-language-2ac8eab8d6db. Accessed*, 2019. 5.6.1

[39] Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 1990. 5.6.2, 6.4.2

[40] Robert B Cialdini, Raymond R Reno, and Carl A Kallgren. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015, 1990. 6.1

[41] Luc S Cousineau. Displaced Discussion: The Implications of Reddit Quarantine and the Movement of TheRedPill to Self-Hosting. *AoIR Selected Papers of Internet Research*, 2020. 5.1, 5.6.1, 7.2.2

[42] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International Conference on Machine Learning*, pages 894–903. PMLR, 2017. 5.7.1

[43] Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152, 2020. 7.3.2

[44] Srayan Datta and Eytan Adar. Extracting inter-community conflicts in reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 146–157, 2019. 5.3.3, 6.2.2

[45] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2017. 2.1, 2.2.1, 2.2.2

[46] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, 2019. 2.2.3

[47] Richard Davis. *The web of politics: The Internet's impact on the American political system*. Oxford UP, 1999. 3.1

[48] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018. 2.2.1

[49] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 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)*, 2019. 4.4

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019. 6.2.2

[51] Bryan Dosono and Bryan Semaan. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019. 2.3.1

[52] Arthur R Edwards. The moderator as an emerging democratic intermediary: The role of the moderator in Internet discussions about public issues. *Information Polity*, 7(1):3–20, 2002. 3.2

[53] Emory James Edwards and Tom Boellstorff. Migration, non-use, and the Tumblrpocalypse: Towards a unified theory of digital exodus. *Media, Culture & Society*, 43(3):582–592, 2021. 1.2

[54] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1041–1048, 2011. 5.5.2

[55] Robert M Entman. Framing bias: Media in the distribution of power. *Journal of communication*, 2007. 4.5

[56] Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, pages 87–96, 2019. 5.1

[57] Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 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)*, 2018. 4.4

[58] Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. Contextual affective analysis: A case study of people portrayals in online# metoo stories. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(1):158–169, 2019. 6.5.3

[59] Casey Fiesler and Nicholas Proferes. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 2018. 4.7.2, 6.6.2

[60] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*, 2018. 2.3.1, 7.3.1

[61] Frank Fischer, Ingo Kollar, Karsten Stegmann, and Christof Wecker. Toward a script theory of guidance in computer-supported collaborative learning. *Educational psychologist*, 48(1):56–66, 2013. 7.2.1

[62] Claudia I. Flores-Saviaga, Brian C. Keegan, and Saiph Savage. 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)*, 2018. 4.4.2, 5.2.1

[63] Jonathan Friedman. Chasm in the Classroom: Campus Free Speech in a Divided America. Technical report, PEN America, 2019. 4.7

[64] AA Frost. The necessity of political vulgarity. *Current Affairs*, 17, 2016. 5.2.2

[65] John D Gallacher, Marc W Heerdink, and Miles Hewstone. Online Engagement Between

Opposing Political Protest Groups via Social Media is Linked to Physical Violence of Offline Encounters. *Social Media+ Society*, 7(1):2056305120984445, 2021. 7.2.1

[66] Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*, 2017. 2.2.2

[67] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. 6.5.3

[68] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. Impact and dynamics of hate and counter speech online. *arXiv preprint arXiv:2009.08392*, 2020. 2.3.2

[69] Mor Geva, Yoav Goldberg, and Jonathan Berant. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, 2019. 6.2

[70] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018. 4.2

[71] Tarleton Gillespie. Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2):2053951720943234, 2020. 1.1

[72] Debbie Ging. Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and Masculinities*, 22(4):638–657, 2019. 5.1

[73] Erving Goffman et al. *The Presentation of Self in Everyday Life*. Harmondsworth London, 1978. 7.3.3

[74] Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945, 2020. 1.1

[75] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5), 2009. 4.7, 5.1, 5.6.2, 5.6.2, 6.4

[76] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47, 2013. 5.1, 5.6.2, 6.4.2

[77] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 1969. 5.6.4

[78] Ryan Greer. Weighing the Value and Risks of Deplatforming. *Global Network on Extremism and Technology*, May 2020. URL `https://gnet-research.org/2020/05/11/weighing-the-value-and-risks-of-deplatforming/`. 7.2.2

[79] Anatoliy Gruzd, Barry Wellman, and Yuri Takhteyev. Imagining Twitter as an imagined community. *American Behavioral Scientist*, 55(10):1294–1318, 2011. 1.2

[80] Anna Guimaraes, Oana Balalau, Erisa Terolli, and Gerhard Weikum. Analyzing the traits and anomalies of political discussions on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 205–213, 2019. 5.3.3

[81] Gahgene Gweon, Carolyn Penstein Rosé, Joerg Wittwer, and Matthias Nueckles. Supporting Efficient and Reliable Content Analysis Using Automatic Text Processing Technology. In *Human-Computer Interaction–INTERACT 2005*, 2005. 3.6.3

[82] Hussam Habib, Maaz Bin Musa, Fareed Zaffar, and Rishab Nithyanand. To act or react: Investigating proactive strategies for online community moderation. *arXiv preprint arXiv:1906.11932*, 2019. 7.2.1

[83] Jack Hessel and Lillian Lee. Somethings Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1648–1659, 2019. 2.2.2

[84] Ryan James Heuser. Word Vectors in the Eighteenth Century. In *DH*, 2017. 6.5.3

[85] Serena Hillman, Jason Procyk, and Carman Neustaedter. 'alksjdf; Lksfd' tumblr and the fandom user experience. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, pages 775–784, 2014. 1.2

[86] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM 2017*, 2017. 3.4.4

[87] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2017. 4.4.1

[88] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving Google's Perspective API built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017. 5.6.1

[89] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933. 6.5.3

[90] Aemilian Hron and Helmut F Friedrich. A review of web-based collaborative learning: factors beyond technology. *Journal of Computer Assisted Learning*, 19(1):70–79, 2003. 3.2

[91] Jina Huh. Clinical questions in online health communities: the case of see your doctor threads. In *CSCW 2015*, 2015. 3.2

[92] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2014. 6.4.2

[93] ITU. For the first time, more than half of the world's population is using the Internet. *ITU releases 2018 global and regional ICT estimates*, Dec 2018. URL `https://www.itu.int/en/mediacentre/Pages/2018-PR40.aspx`. 1

[94] Ravi Iyer, Spassena Koleva, Jesse Graham, Peter Ditto, and Jonathan Haidt. Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLOS ONE*, 7(8):e42366, 2012. 5.6.2

[95] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, 2014. 6.2.2, 7.3.2

[96] Edwin Jain, Stephan Brown, Jeffery Chen, Erin Neaton, Mohammad Baidas, Ziqian Dong, Huanying Gu, and Nabi Sertac Artan. Adversarial Text Generation for Google's Perspective API. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1136–1141. IEEE, 2018. 5.6.1

[97] Sue Curry Jansen and Brian Martin. The Streisand Effect and Censorship Backfire. *International Journal of Communication*, 9:656–671, 2015. 5.2

[98] Shagun Jhaver, Larry Chan, and Amy Bruckman. The view from the other side: The border between controversial speech and harassment on Kotaku in Action. *arXiv preprint arXiv:1712.05851*, 2017. 4.2

[99] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2018. 4.2

[100] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35, 2019. 2.3.1

[101] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27, 2019. 2.3.2, 7.2.1

[102] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. In *CSCW 2021*, 2021. 7.2.2

[103] Shan Jiang, Ronald E Robertson, and Christo Wilson. Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2019. 4.2, 5.1, 7.2.1

[104] Shan Jiang, Ronald E Robertson, and Christo Wilson. Reasoning about political bias in content moderation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 5.1, 7.2.1

[105] Yohan Jo, Michael Miller Yoder, Hyeju Jang, and Carolyn P Rosé. Modeling Dialogue Acts with Content Word Filtering and Speaker Preferences. In *EMNLP 2017*, 2017. 3.4.4

[106] Kristen Johnson and Dan Goldwasser. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, 2018. 5.6.2

[107] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen,

Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Googles multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. 6.5.3

[108] Kenneth Joseph, Lisa Friedland, William Hobbs, David Lazer, and Oren Tsur. ConStance: Modeling Annotation Contexts to Improve Stance Classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1115–1124, 2017. 6.2

[109] Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22, 2017. 6.4.2

[110] Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, 2017. 2.3.3

[111] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP):1–35, 2020. 2.3.2

[112] David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. A just and comprehensive strategy for using NlP to address online abuse. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 3658–3666. Association for Computational Linguistics (ACL), 2020. 1.1, 7.3.1

[113] Anna Kaatz, Belinda Gutierrez, and Molly Carnes. Threats to objectivity in peer review: the case of gender. *Trends in pharmacological sciences*, 35(8):371–373, 2014. 3.2, 3.6.1, 3.6.2, 3.6.3

[114] B Kamisar. Conservatives cry foul over controversial groups role in Youtube moderation. *The Hill*, 2018. 7.2.1

[115] Cecilia Kang. 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*, 2016. 4.4.2

[116] Ayse Karaevli. Performance consequences of new CEO Outsiderness: Moderating effects of pre-and post-succession contexts. *Strategic management journal*, 28(7):681–706, 2007. 7.2.2

[117] Mladen Karan and Jan Šnajder. Cross-domain detection of abusive language online. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 132–137, 2018. 7.3.1

[118] John Kelly, Danyel Fisher, and Marc Smith. Debate, division, and diversity: Political discourse networks in USENET newsgroups. In *Online Deliberation Conference 2005*, 2005. 3.4.3, 3.6.2, 5.4.2

[119] Lina Khatib, William Dutton, and Michael Thelwall. Public diplomacy 2.0: A case study of the US digital outreach team. *The Middle East Journal*, 66(3):453–472, 2012. 3.2

[120] Ashique R KhudaBukhsh, Rupak Sarkar, Mark S Kamlet, and Tom Mitchell. We Don't

Speak the Same Language: Interpreting Polarization through Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 7.3.3

[121] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design*, 2012. 1.2, 4.2

[122] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921*, 2020. 2.2.3

[123] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6.5.3

[124] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 6.5.3

[125] Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *arXiv preprint arXiv:2012.12305*, 2020. 1.1

[126] Aniket Kittur, Bryan Pendleton, and Robert E Kraut. Herding the cats: the influence of groups in coordinating peer production. In *WikiSym 2009*, 2009. 3.2

[127] Jason Koebler and Joseph Cox. Twitter Has Started Researching Whether White Supremacists Belong on Twitter. *Vice. https://www.vice.com/en_us/article/ywy5nx/twitter-researching-white-supremacism-nationalism-ban-deplatform. Accessed*, 2019. 1

[128] Yubo Kou and Bonnie Nardi. Complex mediation in the formation of political opinions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2018. 7.2.1

[129] Yubo Kou and Bonnie A Nardi. Governance in League of Legends: A hybrid system. In *FDG*, 2014. 7.2.1

[130] Yubo Kou, Xinning Gui, Shaozeng Zhang, and Bonnie Nardi. Managing disruptive behavior through non-hierarchical governance: Crowdsourcing in League of Legends and Weibo. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–17, 2017. 7.2.1

[131] Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433, 2004. 6.2.5

[132] Ivar Krumpal. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, 47(4):2025–2047, 2013. 6.2.5

[133] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, 2018. 2.2.2

[134] Renaud Lambiotte, J-C Delvenne, and Mauricio Barahona. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*, 2008. 4.5.2

[135] Cliff Lampe and Erik Johnston. Follow the (slash) dot: effects of feedback on new mem-

bers in an online community. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, pages 11–20, 2005. 6.1

[136] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly*, 31(2):317–326, 2014. 3.6.3

[137] Younghun Lee, Seunghyun Yoon, and Kyomin Jung. Comparative Studies of Detecting Abusive Language on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 101–106, 2018. 2.2.2

[138] Paige Leskin. Tumblr is banning all NSFW content and people are worrying it's the beginning of the end for the Verizon-owned website. *Business Insider. https://www.businessinsider.com/tumblr-bans-nfsw-content-and-users-say-the-platform-will-suffer-2018-12. Accessed*, 2018. 1.2

[139] Joseph CR Licklider and Robert W Taylor. The computer as a communication device. *Science and technology*, 76(2):1–3, 1968. 1

[140] Bill Yuchen Lin, Frank F Xu, Kenny Zhu, and Seung-won Hwang. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, 2018. 6.5.3

[141] Sally Lindsay, Simon Smith, Paul Bellaby, and Rose Baker. The health impact of an online heart disease support group: a comparison of moderated versus unmoderated support. *Health education research*, 24(4):646–654, 2009. 3.2

[142] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5.6.2, 6.4.2

[143] Alexander Maedche, Stefan Morana, Silvia Schacht, Dirk Werth, and Julian Krumeich. Advanced user assistance systems. *Business & Information Systems Engineering*, 58(5): 367–370, 2016. 7.2.1

[144] Kaitlin Mahar, Amy X Zhang, and David Karger. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018. 2.3.3, 7.2.1

[145] Diane Maloney-Krichmar and Jenny Preece. A multilevel analysis of sociability, usability, and community dynamics in an online health community. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):201–232, 2005. 3.2

[146] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014. 6.2.1, 6.2.2

[147] James R Martin and Peter R White. *The Language of Evaluation*, volume 2. Springer, 2003. 6.4, 6.4.2, C.2

[148] Ignacio Martinez. Chapo Trap House subreddit quarantined for allegedly encouraging violence. *The Daily Dot. https://www.dailydot.com/layer8/chapo-trap-house-subreddit-*

quarantine/. Accessed, 2019. 5.2.2

[149] Joan Massachs, Corrado Monti, Gianmarco De Francisci Morales, and Francesco Bonchi. Roots of Trumpism: Homophily and Social Feedbackin Donald Trump Support on Reddit. In *12th ACM Conference on Web Science*, pages 49–58, 2020. 5.2.1

[150] Adrienne Massanari. Reddits Alt-Right: Toxic Masculinity, Free Speech, and/r/The_Donald. *Fake News: Understanding Media and Misinformation in the Digital Age*, 2020. 5.1

[151] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 369–380, 2019. 7.2.1

[152] Louise Matsakis. Twitter releases new policy on "dehumanizing speech". *Wired. https://www.wired.com/story/twitter-dehumanizing-speech-policy/. Accessed*, 2018. 1

[153] Maxwell McCombs. The agenda-setting role of the mass media in the shaping of public opinion. In *Proceedings of the 2002 Conference of Mass Media Economics*, 2002. 4.4

[154] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011. 4.4.1

[155] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 2008. 4.6.3, 6.2, 6.2.5, 7.3.2

[156] Gianmarco De Francisci Morales, Corrado Monti, and Michele Starnini. No echo in the chambers of political interactions on Reddit. *Scientific Reports*, 11(1), 2021. 5.4.2

[157] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network? The structure of the Twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 493–498, 2014. 1.2

[158] Sarah Myers West. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383, 2018. 2.3.1, 4.2

[159] David Neiwert. 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*, 2017. 4.6.3

[160] Edward Newell, David Jurgens, Haji Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. User migration in online social networks: A case study on reddit during a period of community unrest. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, 2016. 5.5

[161] Casey Newton. 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*, 2019. 4.1

[162] Casey Newton. Reddit bans r/The_Donald and r/ChapoTrapHouse as part of a major expansion of its rules. *The Verge. https://www.theverge.com/2020/6/29/21304947/reddit-ban-subreddits-the-donald-chapo-trap-house-new-content-policy-rules. Accessed*, 2020. 5.2.1

[163] Viet-An Nguyen, Jordan L Ying, and Philip Resnik. Lexical and hierarchical topic regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1106–1114, 2013. 4.5

[164] Randal S Olson and Zachary P Neal. Navigating the massive world of Reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science*, 2015. 4.5, 4.5.1, 5.3.3

[165] Jahna Otterbacher, Libby Hemphill, and Erica Dekker. Helpful to you is useful to me: The use and interpretation of social voting. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10, 2011. 7.2.2

[166] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, 2018. 2.2.3

[167] Namsu Park, Kerk F Kee, and Sebastián Valenzuela. Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes. *Cyberpsychology & behavior*, 12(6):729–733, 2009. 1.2

[168] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. Toxicity Detection: Does Context Really Matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, 2020. 2.2.2

[169] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, Republicans and Starbucks Afficionados: user classification in Twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 430–438, 2011. 7.3.2

[170] Bryan Pfaffenberger. A Standing Wave in the Web of Our Communications: Usenet and the Socio-Technical Construction of Cyberspace Values. In *From Usenet to Cowebs*, pages 20–43. Springer, 2003. 3.1

[171] Elena Pilipets and Susanna Paasonen. Nipples, memes, and algorithmic failure: NSFW critique of Tumblr censorship. *New Media & Society*, page 1461444820979280, 2020. 1.2

[172] Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740, 2017. 7.3.2

[173] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 557–568, 2020. 6.1

[174] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Robert West. Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *arXiv preprint arXiv:2010.10397*, 2020. 5.1, 5.5, 5.6.1, 5.8.1, 7.2.2

[175] Annika Richterich. Karma, Precious Karma!Karmawhoring on Reddit and the Front Pages Econometrisation. *Journal of Peer Production*, 4(1):1–12, 2014. 7.2.2

[176] Stephen P Robbins and Timothy A Judge. Organizational behavior, 2013. 7.2.2

[177] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. The structural topic model and applied social science. In *Advances in Neural Information Processing Systems (NIPS) Workshop on Topic Models: Computation, Application, and Evaluation*, 2013. 4.4.1

[178] Sarah T Roberts. *Behind the screen: Content moderation in the shadows of social media*. Yale University Press, 2019. 1.1

[179] Adi Robertson. Was Reddit always about free speech? Yes, and no. *The Verge. https://www.theverge.com/2015/7/15/8964995/reddit-free-speech-history. Accessed*, 2015. 4.1, 4.2, 5.2, 5.7.2, 7.2.2

[180] Adi Robertson. Reddit quarantines Trump subreddit r/The_Donald for violent comments . *The Verge. https://www.theverge.com/2019/6/26/18759967/reddit-quarantines-the-donald-trump-subreddit-misbehavior-violence-police-oregon. Accessed*, 2019. 5.2.1

[181] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2015. 4.4.1

[182] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. 6.2.2

[183] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *arXiv preprint arXiv:1701.08118*, 2017. 2.2.1, 6.2

[184] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint*, 2019. 5.6.2, 6.4.2

[185] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, 2019. 2.2.3, 5.6.1

[186] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Association for Computational Linguistics*, 2020. 2.2.2

[187] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, 2017. 2.1, 2.2.1, 5.6.1

[188] Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A Smith. The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story

Cloze Task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, 2017. 6.2

[189] John R Searle. *Speech acts: An essay in the philosophy of language*. Cambridge UP, 1969. 3.4.4

[190] Joseph Seering. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, 2020. 1, 2.3

[191] Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong'Cherie' Chen, Likang Sun, and Geoff Kaufman. Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019. 7.2.1

[192] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, 2019. 2.3.1

[193] Qinlan Shen and Carolyn Rose. What Sounds Right to Me? Experiential Factors in the Perception of Political Ideology. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1762–1771, 2021. 6.2.5

[194] Qinlan Shen, Michael Miller Yoder, Yohan Jo, and Carolyn P Rosé. Perceptions of Censorship and Moderation Bias in Political Debate Forums. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2018. 4.2, 5.1

[195] Xing Shi, Kevin Knight, and Deniz Yuret. Why neural translations are the right length. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2278–2282, 2016. 6.5.3

[196] Yanchuan Sim, Brice DL Acree, Justin H Gross, and Noah A Smith. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101, 2013. 7.3.2

[197] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. A characterization of political communities on Reddit. In *30th ACM Conference on Hypertext and Social Media*, pages 259–263, 2019. 5.2.1, 6.2

[198] Tim Squirrell. 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*, 2017. 4.4.2

[199] Liam Stack. 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*, 2018. 1

[200] Nick Statt. 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*, 2018. 4.1

[201] Emily Stewart. Reddit restricts its biggest pro-Trump board over violent threats. *Vox. https://www.vox.com/recode/2019/6/26/18760288/reddit-the-donald-trump-message-board-quarantine-ban. Accessed*, 2019. 5.2.1

[202] John Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004. 1

[203] Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 940–950, 2019. 7.3.1

[204] Sali A Tagliamonte. *Variationist sociolinguistics: Change, observation, interpretation*, volume 39. John Wiley & Sons, 2011. 6.2

[205] Deborah Tannen. Language and Culture. In *An Introduction to Language and Linguistics*, pages 343–372, 2006. 6.2

[206] Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. Generating Counter Narratives against Online Hate Speech: Data and Strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, 2020. 7.2.1

[207] Oren Tsur, Dan Calacci, and David Lazer. 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)*, 2015. 4.4, 6.2, 7.3.2

[208] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. Detection and fine-grained classification of cyberbullying events. In *International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 672–680, 2015. 2.1

[209] Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one*, 15(12):e0243300, 2020. 1.1, 2.2.1

[210] Zeerak Waseem. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, 2016. 2.2.1, 6.2

[211] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016. 2.2.1, 2.2.2

[212] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, 2017. 1.1, 2.1, 7.3.1

[213] Tim Weninger. An exploration of submissions and discussions in social news: Mining collective intelligence of Reddit. *Social Network Analysis and Mining*, 2014. 4

[214] Mark H White II and Christian S Crandall. Freedom of racist speech: Ego and expressive threats. *Journal of Personality and Social Psychology*, 113(3):413, 2017. 4.7

[215] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long and Short Papers), pages 602–608, 2019. 2.2.1

[216] Anthony G Wilhelm. *Democracy in the digital age: Challenges to political life in cyberspace*. Psychology Press, 2000. 3.2

[217] Scott Wright. Government-run online discussion fora: Moderation, censorship and the shadow of control. *The British Journal of Politics and International Relations*, 8(4):550–568, 2006. 3.1, 3.2

[218] Scott Wright and John Street. Democracy, deliberation and design: the case of online discussion forums. *New media & society*, 9(5):849–869, 2007. 3.2

[219] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. Demoting Racial Bias in Hate Speech Detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, 2020. 2.2.3

[220] Michael Miller Yoder. *Computational Models of Identity Presentation in Language*. Carnegie Mellon University, 2021. 6.5.3, 7.3.2

[221] Omar Zaidan and Chris Callison-Burch. Crowdsourcing Translation: Professional Quality from Non-Professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, 2011. 6.2

[222] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, 2019. 2.2.2

[223] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, 2018. 2.2.2

[224] Justine Zhang, Cristian Danescu-Niculescu-Mizil, Christina Sauper, and Sean J Taylor. Characterizing online public discussions through patterns of participant interactions. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–27, 2018. 2.2.2, 7.2.1