COMBATTING ABUSIVE BEHAVIOR IN ONLINE COMMUNITIES USING CROSS-COMMUNITY LEARNING

A Dissertation
Presented to
The Academic Faculty

By

Eshwar Chandrasekharan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Computer Science in the
School of Interactive Computing

Georgia Institute of Technology

May 2020

COMBATTING ABUSIVE BEHAVIOR IN ONLINE COMMUNITIES USING CROSS-COMMUNITY LEARNING

Approved by:

Dr. Eric Gilbert
School of Interactive Computing
Georgia Institute of Technology

Dr. Amy Bruckman
School of Interactive Computing
Georgia Institute of Technology

Dr. Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology

Dr. Jacob Eisenstein
School of Interactive Computing
Georgia Institute of Technology

Dr. Cliff Lampe
School of Information
University of Michigan

Date Approved: March 3, 2020

“The past does not repeat itself, but it rhymes.”

Mark Twain

To Mom, Dad, Sri, & my grandparents.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mother, Usha Chandrasekharan, my father, Chandrasekharan Ananthanarayanan, and my sister, Srividhya Chandrasekharan, for their continued support and encouragement throughout this marathon of a PhD. Daily check-ins through video chat (more than once on some days), talking about the different milestones and deadlines on my path, making sure that I am eating, sleeping and feeling well. All of these mattered, helping me stay focused and healthy, making this endeavor an overall positive experience—one that I will cherish for years to come. I also want to thank my grandparents for their blessings and love. Without the love and support of my family, this thesis would not be possible. I would like to dedicate this thesis to you. We did this together, and from now on we shall be referred to as **Dr. Chandrasekharan**'s!

For the first time in my life, I moved away from the same city as my family—Chennai, one of the best cities in the world—for pursuing my PhD in 2015. I left home and moved to unknown territory, in a new country, USA. Thinking back, moving was overwhelming especially to an environment that was completely new to me. But with the help of family and friends, I was able to focus and cherish the opportunities that allowed me to grow as a scholar and as an individual. I want to thank my sister, Dr. Pavithra Chandramowlishwaran, for her continued support and making the transition to Atlanta fun and easy. Shout-out to my cuz's crew of lovely puppies—Minion, Boris, Dooley, and Bobby (honorary puppy, contemporary kitty). Special thanks to my music on Spotify that helped me focus, relax and get hyped—playlists like Verkaut, Hip Hop Controller, Work Jazz, Buddha Bar, and artists like Kanye, Logic, Eminem, Drake, Kendrick, J.Cole, Gorillaz, Maroon 5, Linkin Park, and Parcels, to name some of my favorites.

Without my compassionate and quirky friends, the past 5 years would have been tougher than what it has been. Thanks Dr. Akash Sridhar and Dr. Karthik Abhinav, both of whom I've known for over a decade now, for the conversations about our PhD journeys. Thanks

to my GT hoops crew and IITM basketball teammates—Amritraj Anand and Sagar Joshi—and my FIFA and NBA Live crew at Savannah Midtown—Sir Ravikiran Ramaswamy, Sumedh Parab and Pradyumna Tambwekar. I want to thank the social computing seminar group at GT's School of Interactive Computing, in particular Dr. Umashanthi Pavalanathan, Sandeep Soni, Koustuv Saha, Dr. Stevie Chancellor and Dr. Michaelanne Dye. Special shout-out to my "Team Awesome" partners, Ari Schlesinger and Ian Stewart. I also thank folks at the University of Michigan's School of Information who made my move to Ann Arbor, halfway through my PhD, an amazing experience. Thanks to members of the Social Media Research Lab (SMRL) for making me feel as part of the group. Special thanks to Katie Cunningham for all of the support and experiencing Ann Arbor together, and shout-out to Ashwin Rajadesingan and Srihari Sundar for the *ice cold kaapis*.

I have learnt how important lab members are to the PhD journey. To my amazing labmates—Dr. Catherine Grevet, Dr. Tanushree Mitra, Dr. Chaya Hiruncharoenvate, Dr. Shagun Jhaver, Dr. Mattia Samory, and Jane Im—I appreciate all your help and support during my time in the *comp.social* lab. Also, I would like to acknowledge students who have contributed to the work in this thesis: Anirudh Srinivasan, Hunter Charvat, Chaitrali Gandhi and Katherine Mustelier. It was a pleasure working with all you, and I hope our paths will cross again in the future. I would also like to thank my thesis committee, Dr. Amy Bruckman, Dr. Munmun De Choudhury, Dr. Jacob Eisenstein and Dr. Cliff Lampe. All of your guidance throughout my PhD has been instrumental in shaping my dissertation, and helped me craft a narrative for my research.

Last but not least, thanks to my amazing advisor, Dr. Eric Gilbert for everything! I am forever grateful for your guidance and advice throughout my PhD, and I will strive to emulate your brilliance and positive attitude towards research and mentorship. I will cherish our coffee runs, 1-on-1 shootarounds, and conversations—from talking about AI (*Allen Iverson*) during our recruitment call to talking about using AI (*Artificial Intelligence*) for creating better moderation tools. Thank you so much Eric!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This dissertation aims to develop a deep understanding of abusive online behavior via statistical machine learning techniques to build tools that help counter it. I have developed computational approaches to model abusive online behavior, aiming to address two of the major gaps in this line of research—the scarcity of labeled ground truth required to train effective ML models, and the contextual nature of online moderation by accounting for community-specific norms. First, I introduced a new class of machine learning tools that are based on *cross-community linguistic similarity*. Next, I discovered the existence of widely overlapping norms, across distinct online communities, suggesting that new automated tools for moderation could find traction in borrowing data from communities which share similar values. The abuse models that I build will enable a brand new class of interactive machine learning systems that can sidestep the need for site-specific classifiers. My thesis brings these pieces together in the form of open source software to detect abusive behavior online through *cross-community learning*, and thereby socio-algorithmically govern speech on large-scale Internet platforms like Reddit.

CHAPTER 1
INTRODUCTION

A key challenge for online communities is moderation. For example, the founders of the social media startup Yik Yak spent months of their early time removing hate speech [1]. Twitter has stated publicly that dealing with abusive behavior remains its most pressing challenge [2]. Powered by the disinhibition effect [3, 4, 5], rage-filled edit wars and the conversations that accompany them often break out on Wikipedia [6, 7]. Newspaper and blog comment sections often devolve into epithets [8, 9]. Behavior like this corrodes online communities. It turns away new Wikipedia editors [10, 11], threatening the long-term viability of the important Internet encyclopedia [12]. Recently, Anderson et al. showed in a controlled lab experiment that simply varying the tone of a blog post's comments (i.e., from civil to rude) made participants deeply distrust the post itself [13]. In response, many popular sites (like New York Times, NPR and Popular Science) have disabled the ability to comment at all because of problems moderating those spaces [14, 15], and empirical work has shown that people leave platforms after being the victims of online abuse [16]. Moreover, recent Pew surveys indicate that abuse happens much more frequently than often suspected: approximately 40% of Internet users report being the subject of online abuse at some point, with underrepresented users most often targeted [17, 18, 19].

1.1 Online Content Moderation

Recently, the moderators (or "mods") of a large gaming community on Reddit, r/Games, released screenshots[1] of bigoted, transphobic, racist, misogynistic, pedophilic, and otherwise hateful comments that they had moderated [20]. In their statement, the mods wrote:

[1]Content warning (see above): `https://imgur.com/a/umrdBYF`

> Unfortunately, this inflammatory content is not infrequent ... These are some of the more awful comments we see regarding transphobia, homophobia, islamophobia, racism, misogyny, pro-pedophilia/pro-rape, and vitriolic personal attacks against other users. These kinds of comments occur on a daily basis. We've compiled an entire album of examples of the horrible things people say on this subreddit. From bigotry to vitriol, this album merely scratches the surface of the magnitude of the problem.[2]

While most mainstream platforms prohibit obviously racist, homophobic, and hateful content, platforms still intake vast amounts of it [2, 21]. As the r/Games mods tried to make visible, moderators are on the front lines of the battle to keep such content out of their online communities and off platforms [22]. Platforms and their moderators use a variety of different approaches to regulate behavior in online communities, and subsequently limit the damage that bad actors cause [23]. On most sites, those techniques take two primary forms: human moderation, and human moderation augmented by automated techniques.

1.1.1 Regulating platforms through human moderation

Most social platforms employ the services of moderators (either paid or unpaid) who regulate content generated within the platform. Human moderation typically takes two forms: centralized [23] and distributed [24, 25] approaches. In the centralized approach, teams of human moderators such as externally contracted workers, and/or a small number of power users—manually go through posts, and scrub the site of content with racist, homophobic, or misogynist language or imagery [26]. In the distributed approach, a social platform's users triage inappropriate submissions via voting or reporting mechanisms—after which the site can take action (often, moderators take these actions).

[2] `https://www.reddit.com/r/Games/comments/b7ubwm/rgames_is_closed_for_april_fools_find_out_why_in/`

1.1.2 Challenges faced by human moderation

Human moderation approaches require moderators to perform a great deal of manual labor, and these suffer from drawbacks when deployed within large-scale platforms [27, 28]. In the centralized approach, groups of paid or volunteer moderators constantly regulate all of the content generated within platforms. This constant exposure to disturbing content negatively and substantially affects the mental health of moderators [29, 30, 31]. In the distributed approach, platforms require users to report inappropriate content before taking action—the exact type of content platforms wish their users did not have to encounter in the first place [32]. In addition, human moderation struggles to keep up with the immense volume of content generated within large-scale platforms—plenty of content that violates site guidelines remains online for years [22].

1.1.3 Using AI to triage content for human moderators

To keep up with the volume of content created by users, social platforms—like Facebook [33], YouTube [34], and Twitter [35]—are known to train machine learning (ML) algorithms by compiling large datasets of past moderation decisions on the platform. Deploying these algorithms without any human oversight can be detrimental; for example, Tumblr "caused chaos" recently when it launched a new, unsupervised anti-porn algorithm on the site [36]. Nonetheless, ML approaches can be especially helpful for algorithmically *triaging* comments for human moderators to review. However ML-based approaches face drawbacks that prevent them from being easily deployed—scarcity of labeled ground truth data, and the contextual nature of moderation.

Scarcity of labeled ground truth data

First, ML-based approaches require vast amounts of labeled ground truth data for training effective models. These data are difficult to obtain since platforms do not share moderation decisions publicly due to privacy and public relations concerns. Brand new or small

online communities, by definition, have little to no training data at all. Moreover, new and emerging communities lack the resources to develop effective moderation tools from scratch [1]. In this thesis, I introduce a technique called *cross-community learning* which can side-step the need for site-specific data and classifiers for moderation [37]. In cross-community learning, data obtained from one or more source communities is used to detect violations within a completely different target community.

Contextual nature of moderation

Second, moderation is a highly contextual task. Moderation decisions about what is considered acceptable or undesirable are guided by an online community's norms [38, 39]. But norms vary widely across communities; even behavior considered undesirable in one community may be valuable in another [40, 41]. A common failure mode for moderation algorithms is failing to understand the community norms where they are being deployed.

An online community's norms play an important role in guiding acceptable behaviors, and therefore in its governance [38]. Moderators (mods) help maintain normative behaviors within their respective communities in a variety of ways. One common approach taken by the mods to enforce community norms is by removing content that violates normative guidelines. Online community moderators have to sanction pedestrian normative violations like posting spoilers about a TV show, as well as more serious infractions like online abuse [2], harassment [17, 18, 39], and fake news and misinformation [42]. Yet, norms for what is appropriate can vary widely from one community to another. Even behavior considered harmful in one community might be celebrated in another (e.g., 4chan's /b/ [40], Something Awful Forums [41]).

When designing automated tools for moderating online communities, it is important to take the community's norms into account. Given that different communities care about different sets of norm violations, the severity of infractions can vary given the context or the nature of the topic of discussion (e.g., sensitive topics around politics or mental health

would require the moderators to be less tolerant of trolling or vitriol). In other words, the communities (members and moderators) themselves are the best judges of the types of speech that are either valued, or considered to be norm violations. These nuances are important to take into account, as platforms and researchers are doubling down on ML approaches towards moderation. As a step towards this goal, I work closely with Reddit moderators, and develop a new *mixed-initiative* [43] system for regulating content. This human-AI collaboration will allow moderators of target communities to use ML classifiers trained on data obtained from other source communities, that share similar norms.

My dissertation aims to bridge the gaps discussed above by introducing a new class of ML systems, that are based on *cross-community* learning, in order to combat abusive behavior in online communities. Next, I will briefly introduce the different components of my thesis, highlighting key methodologies and findings from each chapter.

1.2 The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data

In Chapter 3, I introduce a new analytic concept for studying and building online communities: the *Bag of Communities* (BoC) approach, which aims to sidestep site-specific classifiers (and their data) by computing similarity scores between one community's data and preexisting data from other online communities. I go on to use it in an "existence proof:" identifying abusive posts from a major online community, demonstrating that an algorithm with access to only *out-of-domain* data can predict abusive posts in another community—without access to data from that community. In essence, BoC could allow communities (especially new ones with limited resources) to spend time on what differentiates them from other places on the Internet, and less time on common problems shared across sites.

1.2.1 Summary of methods and findings.

Relying on 10M posts collected from 9 different online communities from 4chan, Reddit, Voat and MetaFilter, I show that a BoC-based linguistic classifier outperforms an in-domain classifier with access to over 4 years of site-specific data. I demonstrate that a BoC classifier can be used on a target community "off the shelf" with roughly 75% accuracy—no training examples are needed from the target community. That is, an algorithm with access to only *out-of-domain* data can predict abusive posts in another community—without access to data from that community. In addition to this static model, I also explore a dynamic BoC model mimicking scenarios where newly moderated data arrives in batches (similar to online learning [44]). It outperforms a solely in-domain model at every batch size, achieving 91.18% accuracy (95% precision) after seeing 100,000 human-moderated posts. This is notable since it implies that while the BoC approach will help communities without moderators to generate training data (static model), the BoC will continue to boost systems that predict abusive behavior after years of professional moderation (dynamic model).

1.3 An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales

Next, I examine community norms on Reddit in a large scale, empirical manner, in Chapter 4. I identify three types of norms within Reddit: *macro norms* which are universal to most parts of Reddit; *meso norms* which are shared across certain groups of subreddits; and *micro norms* which are highly specific to individual subreddits.

1.3.1 Summary of methods.

I first train linguistic classifiers for 100 top subreddits, using moderator-removed comments from each subreddit; those classifiers only "see" their own subreddit's data and predict moderator removals in that subreddit. Next, I ask the classifiers to estimate a counterfactual:

For every comment in my dataset, *what would this subreddit have done if this comment had been posted there?* Using this, I cluster subreddits that often agree to remove the same comments (based on their classifiers' predictions). Finally, I compile all comments that subreddits within each cluster agree to remove, and employ open coding to identify three different types of norms on Reddit: *macro*, *meso*, and *micro* norms.

1.3.2 Findings.

Macro norm violations include employing personal attacks, misogyny, and hate speech in the form of racism and homophobia. In addition, *controversial views around Donald Trump*[3] , and *criticizing moderators* are norm violations on most parts of Reddit. Meso norms, by contrast, are not universal, and are only enforced by subgroups of subreddits. As expected, not sharing personal anecdotes, and not posting links to promotional spam are meso norms. Perhaps surprisingly, comments only *expressing thanks*, or acknowledging a good point, are meso norm violations. Furthermore, I observe that "mansplaining," mocking religion and nationality, and hostility toward immigrants exist only at meso scales—they are not considered norm violations on all of Reddit. Finally, I find highly specific micro norms that apply to individual, relatively unique subreddits. These are not widely enforced on most other parts of Reddit; one example is using high school-level science to explain new scientific discoveries (e.g., on r/AskScience).

This discovery of widely overlapping norms, across distinct online communities, suggests that new automated tools for moderation could find traction in borrowing data from communities which share similar values.

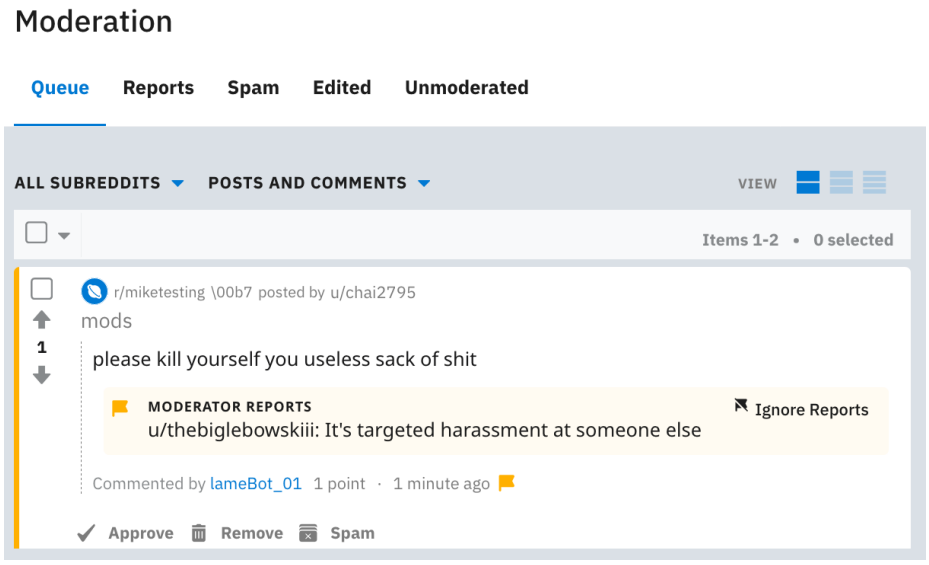

Figure 1.1: This illustration describes how the system I build will be integrated into Reddit's interface to support different moderation actions. When the moderator (mod) tool is configured to triage norm violations, comments flagged by the mod tool will be *reported*, and sent for further review by mods. In this illustration, the first comment in the moderation queue is reported by the new system (denoted by the username */u/thebiglebowskiii*) for *targeted harassment*. All comments in the Moderation *Queue* can be reviewed by mods, and the mod can perform three different actions based on whether they agree/disagree with the reports/flags raised by the mod tool. If they disagree with the report, and feel that the comment does not violate community norms, they can "approve" the comment. Instead, if they agree that the comment does indeed violate community norms, then they can either "remove" the comment, or mark it as "spam".

1.4 *Crossmod*: A Cross-Community Learning-based System to Assist Reddit Moderators

Building on the previous chapters, I build a new, open-source, AI-based moderation system for assisting moderators of communities on Reddit. I call this system the CrossModerator or *Crossmod*. Crossmod aims to overcome the problems above by embracing a sociotechnical partnership with mods, who understand their community's norms. Specifically, I adopt a *mixed-initiative* [43] approach in Crossmod, allowing moderators of subreddits to

[3]I have seen instances where some subreddits disallow posting about Donald Trump so as not to attract the attention of Trump-supporters elsewhere on Reddit.

augment the automatic predictions obtained from cross-community learning with human decisions and oversight. An illustration of how I do this using the new moderation system is shown in Figure 5.1. The Bag of Communities (BoC) approach, and the discovery of widely overlapping norms among disparate online communities lay the foundation for this new class of auto-moderation tools.

1.4.1 Formative interviews.

In Chapter 5, I conduct a formative interview study with 11 mods from 10 different subreddits to understand the current state of automated moderation tools on Reddit, as well as opportunities for extending those tools. I also work closely and iteratively with these moderators through all stages of building Crossmod.

1.4.2 System development.

Using the insights gained from these interviews, I design a new, large-scale, automated moderation system for Reddit in Chapter 6. Through the socio-technical system [45] I build, I explore how Reddit moderators may regulate their communities by augmenting automatic predictions from my cross-community classifiers (*technical*) with human judgement (*social*). Developed with iterative, participatory methods, Crossmod is a AI-based moderation system that is freely available and open-source. The machine learning backend for Crossmod leverages *cross-community learning* [37, 21]; specifically, it uses classifiers trained on the moderation decisions from 100 other communities over roughly a year. For example, Crossmod's ML-backend provides counterfactual estimates about what 100 communities would do with new content, as well as whether that content resembles racism, homophobia, or other types of abuse. Driven by my formative interviews, Crossmod wraps this backend in a sociotechnical architecture that fits into existing moderator workflows and practices.

1.4.3 Summative evaluation.

Finally, I deploy Crossmod in a controlled environment, simulating real-time conversations from two large subreddits with over 10M subscribers each—r/science and r/Futurology. Two moderators from each subreddit evaluated Crossmod's moderation recommendations by manually reviewing comments scored by Crossmod that are drawn randomly from existing threads in their own subreddit. Moderators reported that they would have removed 648 (95.3%) of the 680 comments surfaced by Crossmod; however, 637 (98.3%) of these comments were still online at the time of this writing. In other words, moderators reported that those comments should have been removed, but that the current sociotechnical moderation architecture failed to help them do so.

1.5 Contributions of this thesis

Here I discuss the design and theoretical implications of my work for online moderation, and online communities more broadly.

First, I present a novel approach based on cross-community learning to side-step the need for site-specific data and classifiers for moderation in Chapter 3. One of the major gaps that my thesis aims to address is the scarcity of labeled ground truth required to train effective ML models. This is also known as the *cold-start* problem. Through my *Bag of Communities* approach, I demonstrate that online communities can leverage training data obtained from other preexisting online communities to address the challenge of data scarcity. Limited resources prevent new and emerging online communities from investing too much time and effort into moderation strategies (automated or otherwise).

For CMC and HCI theory, BoC provides a new analytic lens through which existing online phenomena may be examined. For example, a researcher might use BoC to empirically derive a taxonomy of online communities based on their similarity to one another. From a systems perspective, BoC may allow sites to address a variety of common problems. In

addition to identifying abusive behavior (the focus of the present work), sites need to sort content based on its likelihood to be interacted with, identify spam, and decide whether a post requires intervention by professionals (e.g., suicidal ideation). In the latter case, for example, one could imagine deploying BoC using r/suicidewatch, a Reddit suicide support forum, as a companion data source. In essence, BoC could allow communities (especially new ones with limited resources) to spend their time on what *differentiates* them from other places on the Internet, and less time on common problems shared across sites.

Given the size of Reddit's user base—and the wide range of topics covered by different subreddits—I believe this work is the first large-scale study of norms across disparate online communities. In other words, the findings from Chapter 4 shed light on *what Reddit communities values*, and how widely-held those values are. For the design of online communities, it may be possible to use the frame of macro, meso, and micro norms to derive normative guidelines for new communities. That is, the norms identified in this work may serve as sensible defaults for a new online community. Moreover, the discovery of widely overlapping norms suggests that new automated tools for moderation could find traction in borrowing data from communities which share similar values—a direction I explore in 6.

The contributions of the system I develop in Chapter 6 are two-fold. Firstly, I make a systems contribution where I develop a novel sociotechnical system for moderation on Reddit. To the best of my knowledge, Crossmod is the first open source, AI-backed sociotechnical moderation system to be developed and released publicly.[4] Prior work on online governance tend to focus on the algorithmic *detection-side* of moderation. Crossmod extends this line of research by exploring ways to go beyond detection towards enforcement. Second, I develop Crossmod by upholding the principles of participatory design—workers who are involved in the work system should be given a voice in the design process to determine how the new system could improve the quality of their work [46, 47]. In this thesis, I present a grounded approach to include the voice of the moderators (i.e., the workers) in

[4]The code for Crossmod is publicly available at `https://github.com/ceshwar/crossmod`.

the design process to determine how Crossmod (i.e., the work system) could improve the quality of moderation (i.e., their work). Through this work, I hope to inform the design of similar AI-backed sociotechnical systems in the future, and ameliorate the *social-technical gap* in similar CSCW applications [48].

CHAPTER 2
BACKGROUND

In this Chapter, I discuss related work on abusive online behavior, online moderation, role of social norms in online governance, and discuss commonly deployed approaches to moderation on Internet platforms. I conclude by laying out the challenges faced by current methods, and discuss how my work helps address these problems.

2.1 Abusive behavior online

Abusive behavior and norm violations can be difficult to define precisely, but as Papacharissi puts it, "we know it when we see it" [49]. For the purpose of this work, abusive behavior is an omnibus term including harassment, threats, racial slurs, sexism, unexpected pornographic content, and insults—all of which can be directed at other users or at whole communities. Abusive behavior has been part of the fabric of online communities since the Internet's earliest days [3, 4, 5, 23, 50, 51, 52]. Moreover, while online communities have changed significantly since the 1980s and 1990s, abusive behavior and harassment have migrated with those changes [53]. Precisely measuring the extent of abuse in modern social media and online communities is difficult—as sites that gather such data rarely share or report on it due to public relations concerns—however, very recent survey studies from Pew indicate that abuse happens much more frequently than many people suspect (approximately 40 % of Internet users report being the subject of online abuse at some point), particularly for underrepresented users of the Internet [17, 19]. Recent work by communication scholars has shown that abusive comments have structure—notably in terms of topicality and evidentiary support, both examined in the present work—uncovered through labor-intensive manual coding [8].

My thesis looks to make headway on this problem by building foundational science

around the structure of abusive behavior online via data-intensive methods—science that should help new and emerging online communities going forward.

2.2 Online moderation

Lawrence Lessig argued that in social interactions mediated by computers, there are four factors that can be used to shape behavior: markets, architecture, policy and norms [38]. First, I survey related research in two of these areas: *online moderation*, and *social norms*, both offline and online.

There are a variety of different approaches for regulating behavior in online communities. In a comprehensive meta-analysis, Keisler et al. present ways to limit the damage that bad behavior causes when it occurs, and to limit the amount of bad behavior that a bad actor can perform [23]. Current online platforms tend to rely on a combination of policy and design for regulating behavior. Policies are posted to make clear what is allowed and what is not [54] and then technical tools are used and human workers employed to enforce those rules. Technical tools depend on the ability to either edit or delete content (including users) or to append new information to content to inform future users. Sites like Reddit, Stack Overflow, and Yik Yak use *distributed social moderation* [24, 25]. On these sites, the content is moderated through a voting mechanism where registered users up-vote or down-vote each submission or comment. Such voting determines how prominently any content is displayed on the site. This model allows the community to collectively decide its threshold for what content is acceptable and which issues need to be articulated and discussed.

Online communities like Facebook groups and subreddits also use *centralized moderation* [23]. In this model, a small number of users called *moderators*, who are usually regulars from within the community, manually remove posts and comments that violate community norms. Such communities usually specify the rules for posting content on their forums, and these rules guide the moderation. This model often employs automated tools to flag posts (for example, posts containing any of a list of pre-specified offensive words

or violating formatting requirements) for review by the moderators. In some communities, the moderators also review posts that are flagged by regular users on the community. After review, the moderators either remove the content from the site if they find it inappropriate for their subreddit, or allow it appear on the subreddit otherwise. Research on automatic approaches to moderating online antisocial behavior has shown that textual cyberbullying [55, 56] and undesirable posting [57, 58, 59, 37] can be identified based on topic models, presence of insults and user behavior.

Although the approaches mentioned above are widely used, they suffer from shortcomings. The first two approaches require a great deal of human labor. Particularly, in the centralized moderation approach, a few moderators have to spend countless hours in order to maintain the community [32, 31]. While some kinds of distributed moderation can be effective [39, 60], it can also make things worse and serve as a potential trigger for deviant behaviors [61, 62]. The literature around automated moderation approaches lack empirical studies about the effectiveness of various abusive content moderation strategies. This is largely due to the fact that when a site employs a moderation approach that removes content from the internet, it is therefore no longer visible.

Despite there being several studies about online moderation and building computational tools to assist moderators, regulating bad behavior still remains a pressing challenge for online communities [63, 64]. Therefore, an understanding of what norms are actually being enforced by moderators is important. I will build on this line of research by deriving norms from removals by human moderators.

2.3 Social norms online and offline

Social norms are rules and standards that are understood by members of a group, and that guide and/or constrain social behavior without the force of laws [65]. These norms emerge out of interaction with others; they may or may not be stated explicitly, and any sanctions for deviating from them come from fellow members of the social group, not the legal

system. Norms vary to the extent to which they are *injunctive*, prescribing the valued social behavior, versus *descriptive*, informing us about how others act in similar situations [66, 67]. In addition to commonly accepted rules of desirable behavior, norms include rules forbidding unacceptable social behaviors, such as taboos against incest or infanticide, and laws or standards for conduct established by a government or elected body [68]. Norms shape our behavior related to more quotidian activities as well, from how loudly one should speak on a cell phone in a public space, to what the appropriate dress is in different social situations.

Regulation through policies, rules, and guidelines is not always visible, with governance occurring at the level of informal norms instead. Prior work on governance in online communities suggests the importance of social norms in regulating behavior, yet we also know that the difficulty for newcomers learning norms can lead to high drop-out rates [69]. Norms on Reddit are nested. Some norms are adopted from the general social context, for example that pejorative adjectives indicate rudeness. Some norms are shared across the internet, like all caps being the equivalent to shouting. Some norms are Reddit-wide, while others exist in some subreddits, but not others.

2.3.1 Rules vs norms

Rules and norms are interrelated, differing in their degree of explicitness [65]. Certainly in the context of Reddit, rules and norms are loosely coupled, with some mods in some subreddits turning norms that are enforced behind-the-scenes into explicit rules that face the community. Recent work has surveyed outward-facing subreddit rules [54], finding the frequencies of different rule types across Reddit (though approximately half of all subreddits have no explicit rules at all). It may be fair to think of Reddit rules as the front-stage to the norm's back-stage; that is, a rule is a formalized norm, and a norm is an informal rule, with a fluid boundary between the two. For the purposes of my work, I treat norms as the emergent themes in the record of mod removals, some of which may overlap with explicitly

formalized subreddit rules (however, a far greater proportion do not; see Tables 4.3–4.5.)

In this thesis, I will leverage the language used in comments removed by moderators to identify and understand community norms. By exploring where these norms overlap across communities, the present work is one of the first large-scale studies of norms across disparate online communities.

2.4 Complexities around content moderation decisions

A key point of contention in content moderation decisions is how standards and guidelines for what is acceptable differ widely across nations and platforms. Across nation-states, there are no laws that apply universally for governing online discourse. For example, the USA and European countries have adopted different strategies to regulate the use of racist hate speech over the past 50 years or so [70, 71]. A major factor behind this difference is due to the rulings by the US Supreme Court and the European Court of Human Rights—"while the Supreme Court has elected to uphold the freedom to express inflammatory racism in public, the European Court has almost always sided with its 47 member states when they have enforced laws curbing racist hate speech" [71]. On the other hand, countries like China block access to platforms like Twitter and Facebook, and impose strict censorship policies around what can be posted and shared on domestic social media platforms [72, 73]. Such differences across countries lead to debates around content moderation policies, especially when dealing with content that has cultural and political significance, and the institutions for governing online speech are under flux.

At the platform-level, there are disputes around moderation policies of internet companies. For instance, Facebook's decision to moderate (or remove) the "Napalm Girl" photograph—one of the most powerful images that represented the conditions during the Vietnam War—in September 2016, has sparked a debate around censorship power and policies of large-scale, privately owned platforms like Facebook [74, 75, 76, 77]. As more and more of our social interactions and public discourse move online, it is crucial that we

examine the choices made by moderators on internet platforms.

Despite these differences in standards, scholars have observed that all societies have a set of core values to maintain their existence [78, 79]. Quinn (2017) argued that the existence of common values among all societies is a powerful response to the contention that different social contexts demand different moral guidelines, which is at the heart of the argument in favor of cultural relativism [80]. In the context of moderation decisions, identifying and understanding such shared values can serve as a starting point in the creation of a new class of moderation approaches for governing online communities. Irrespective of the type of online platform, there should be certain values that need to be upheld in the design of moderation systems. What are these values supposed to be? Who gets to make these decisions? These are complex questions to consider during the design of sociotechnical systems. Value-sensitive design mechanisms can help navigate these nuances, by taking into account the perspectives of the multiple stakeholders affected by these decisions—e.g., moderators, platform administrators, policy makers, end-users, designers and so on. In this thesis, I focus primarily on community moderators on Reddit (who moderate subreddits on a voluntary-basis), interviewing them to understand their design needs and build a new moderation system to assist them.

2.4.1 Rethinking institutions for governing online discourse

There is a pressing need to rebuild these systems and to rethink the institutions that we are unable to count on anymore. There have been debates around coming up with "absolute" laws for what are acceptable types of content online, and measures are being taken towards this goal, both at platform and legal-levels. At the platform-level, a recent example is Facebook's Safety Advisory Board ("Oversight Board"), which is made up of independent online safety organizations and experts [81]. Members provide expertise, perspective and insights that inform Facebook's approach to safety. The plans for the new Oversight Board came amid intense scrutiny of Facebook and other social media companies over their con-

tent moderation policies [82]. In response to debates around how the unprecedented size and influence of social media has given private companies too much power to unilaterally police speech online, and Facebook proposed an Oversight Board—a quasi-judicial body that will oversee its moderation apparatus, hear appeals, and make rulings that will govern the company's approach on the issue [81, 82].

At the legal-level, laws to govern online speech are under flux as well. Recent examples are the *Stop Enabling Sex Traffickers Act* (SESTA) and *Allow States and Victims to Fight Online Sex Trafficking Act* (FOSTA)—the U.S. Senate and House bills that as the FOSTA-SESTA package became law on April 11, 2018 [83, 84]. These laws "clarify the country's sex trafficking law to make it illegal to knowingly assist, facilitate, or support sex trafficking, and amend the *Section 230* safe harbors of the Communications Decency Act (which make online services immune from civil liability for the actions of their users) to exclude enforcement of federal or state sex trafficking laws from its immunity" [83]. But FOSTA-SESTA creates an exception to Section 230 [85, 86] meaning that website publishers would be responsible if third parties are found to be posting ads for prostitution—including consensual sex work—on their platforms. The goal of this was supposed to be that policing online prostitution rings gets easier [83, 84]. But recent debates argue that FOSTA-SESTA has instead created confusion and immediate repercussions among a range of internet sites as they grapple with the ruling's sweeping language [87].

2.4.2 Facilitating changes from the *bottom-up*: Community-level moderation tools

Given the legal debates around content moderation policies, and how things are in flux at the nation-level and platform-level, I take a bottom-up approach in this thesis, helping facilitate changes at the *community-level*. I work closely with groups of individuals (i.e., community moderators on Reddit) who have inadequate moderation tools to support their communities, and aim to build new tools to support them. Recent reports have found that Reddit is systematically failing to limit the damage caused by bad actors, and moderators

are struggling due to the large volumes of abuse constantly directed at them—resulting in moderator burnout, and even mental health risks [88]. Moderators play a key role in governing communities on Reddit. Moderators enforce rules that are community-specific [89], in addition to site-wide (content[1] and anti-harassment[2]) policies. My thesis will allow community moderators define and regulate behavior within their groups through a new class of interactive ML systems, by introducing a novel approach called *cross-community learning*. This research presents a step towards automated tools that account for community norms by examining the types of norms (or values) around acceptable online behavior that are shared across disparate communities on Reddit. I examine the moderation decisions made by moderators across different online communities, shedding light on actual moderation practices on a large Internet platform like Reddit, and how widely spread these values are.

The key idea in this thesis revolves around the use of *cross-community* data from the past to train ML systems that can inform decision-making by moderators. The feasibility of using prior assumptions or past decisions made by moderators to inform future moderation decisions was an empirical question that needed to be answered. This hypothesis was validated by data, demonstrating the power of *cross-community similarity* in improving the current state of systems for online moderation. My empirical approach can also be used to examine past moderator enforcement of community norms, and allows established communities to reflect on and refine their current moderation practices, answering questions like: "Are prior moderator decisions good?", "Are there norms that are problematic?". In Chapter 4, I find that racist, misogynistic and homophobic speech is generally considered unacceptable on most Reddit communities (i.e., macro-level norms). For the design of online communities, it may be possible to use the frame of these macro norms to derive normative guidelines for new and emerging online communities. That is, the norms identified in this work may serve as sensible defaults for a new online community. On the other hand, some norms can be problematic (e.g., *do not criticize mods*, *do not express thanks*)

[1] `https://www.redditinc.com/policies/content-policy`
[2] `https://redditblog.com/2015/05/14/promote-ideas-protect-people/`

and suggest challenges for designing large-scale discussion systems. *How do you support dyadic relational maintenance without interfering in the larger discussion? How do you provide a place for discussion and arbitration of mod actions?*

2.5 Commonly Deployed Approaches to Moderation

Most online platforms today curate content generated by their users in different ways—presenting content in an ordered form to increase user engagement (e.g., Facebook News-Feed [90]), promoting sponsored content (e.g., ads [91]), and moderating (or screening) undesirable content [32, 64, 92]. Different social platforms adopt different approaches to moderating content. But these approaches can be broadly categorized as two types—human moderation (centralized and distributed), and automated moderation. Next I describe commonly deployed approaches from each category in detail, highlighting how they are currently used in practice.

2.5.1 Human approaches to content moderation

Human moderation take two primary forms—centralized and distributed approaches. In the central moderation approach, most sites employ the services of commercial content moderation workers or moderators who regulate content generated within the platform (e.g., Facebook, Twitter, YouTube, Reddit) [31, 76, 93]. Moderators either screen all of the content before it gets posted (*proactive*), or deal with it after the content is either reported by other users or triaged by an AI-agent (*reactive*). Content found to violate site guidelines, community norms, or even the law are typically removed off-site, with users being sanctioned and even banned from the platform in severe cases [29]. In the distributed approach, users up-vote desirable content making it more visible, and down-vote undesirable content, sometimes even reporting content [32, 94, 95, 96].

Moderators are critical to platforms that rely heavily on human moderation, acting as digital gatekeepers who decide what content gets posted on the platform and what con-

tent gets taken down [76]. By curating content generated by users, moderators help guard against serious infractions that might do harm to a social platform's digital presence [22]. More importantly, moderators keep users from (unintentionally) viewing racist, homophobic, violent, misogynist, etc., content. [21, 26].

Challenges faced by human moderation approaches

Human moderation approaches suffer from drawbacks when deployed at scale, especially the need for a great deal of human labor [27, 28]. In the centralized approach, the labor falls on a small number of paid workers or volunteers who must work tirelessly to maintain the community [29]. Despite their significant role in shaping online discourse and improving user experience on social platforms, moderation workers typically receive low wages and work long hours. Moderators on most platforms are dispersed globally, typically hired from firms located overseas [30]. Prior work on commercial content moderation observed that:

> The work is almost always done in secret for low wages by relatively low-status workers, who review content day in and day out, digital content that may be pornographic, violent, disturbing, or disgusting. [26]

Recent reports have found that continued exposure to such violent and disturbing content can lead to detrimental effects on the mental well-being of moderators [30, 97]. The tasks performed by moderation workers range from repetitive and pedestrian normative violations like posting spoilers about a TV show, to exposure to images and material that can be violent, disturbing and, at worst, psychologically damaging [21, 98]. The workers are often further isolated because the work they do is carried out in secret, because their employers consider their work to be a threat to the brand [29].

In addition, the rapid pace and scale at which content is constantly generated within large-scale platforms restricts the effectiveness of *proactive* review processes that rely on human moderation alone—where a moderator examines every new piece of content before

it appears on the site. As a result, most platforms employ human moderation in a *reactive* manner. In other words, even horrific content may exist on the site for some period of time where other users view and experience it, until the content gets reported, and subsequently removed off-site by a human moderator [26]. Due to the sheer challenge of regulating content within such large-scale platforms, plenty of content that violates site guidelines remain online for days, and sometimes even years [22]. Moderation is hard, and not all types of content are simple or easy for moderation workers to spot or to adjudicate, either. Some of these tasks involve complicated matters of judgment that involve familiarity with community norms and context information. Moreover, when the work of moderating platforms is outsourced to other parts of the world (e.g., Philippines [99]), it creates an additional hurdle to be familiar with social norms: workers must become steeped in the racist, homophobic, and misogynist tropes and language of another culture for which the content is destined [76].

2.5.2 Automated approaches to content moderation

In an effort to keep up with ever-growing content, technical approaches are reported to exist within social platforms communities [100, 64, 101, 102]. These range from simple word and source-ban lists, to more sophisticated AI-based techniques that can flag inappropriate content. Word and source-ban lists take automated actions by filtering content based on either the use of black-listed words [1], or posting from blacklisted IP addresses [103]. Social platforms are also known to train machine learning algorithms by compiling large datasets of example posts that have been moderated off-site [33, 34, 35]. Machine learning-based approaches can be especially helpful for algorithmically *triaging* comments for a much smaller number of (perhaps paid) human moderators.

Challenges faced by automated approaches to content moderation

Automated approaches to triage undesirable content have the potential to allow human moderators to review content more effectively and target their manual labor towards reviewing content that are likely to be violations. But current automated moderation approaches face some key drawbacks.

Static methods are crude by modern standards. Static word and source-ban lists have been observed to perform poorly [104]. They need to be constantly updated to keep up with the evolving nature of online behaviors to remain effective, and they are also prone to "false positives" as they typically do not account for context-specific information [105, 61].

Scarcity of labeled ground truth data. Machine learning-based approaches are generally more effective than static word-ban lists, but they require vast amounts of labeled ground truth data for developing reliable models. Moreover, new and emerging platforms suffer from the "cold start problem"—they lack enough data from their respective users. Such platforms lack the data and resources required to develop reliable automated moderation systems for triaging undesirable content [106]. In this thesis, I use *cross-community learning* to address the scarcity of labeled ground truth. The core idea behind cross-community learning is to learn from data obtained from different source communities to detect violations within a completely different target community [37]. Recent research has shown that *cross-community learning* can be useful to side-step the need for site-specific data and classifiers for moderation [37, 21]. This thesis extend this line of work.

Online moderation is contextual. Online moderation is a highly contextual task, and community norms guide moderation decisions by defining what is acceptable, or undesirable within an online community [38, 40, 41]. Community norms can vary widely across communities, and sometimes behavior considered undesirable by most may even be promoted in certain places (e.g., r/fatpeoplehate [39], Something Awful Forums [41], and r/RoastMe [107]). Such nuances are important to take into account, as platforms and

researchers are doubling down on automated approaches towards moderation.

2.5.3 Recent advances in online moderation

Prior work on ML-based approaches to detect online misbehavior (e.g., abuse [108, 37], toxicity [109, 110], hate speech [39], violations [111]) focus mainly on the detection-side of online moderation as if it is the ultimate step. Despite being a critical part of how norms are enforced and communities are regulated, the enforcement-side of online moderation remains relatively unexplored [112, 39]. As a first step towards this goal, I work closely with Reddit moderators to develop a new AI-based moderation system that can be easily customized to detect content that violate a target community's norms and enforce a range of moderation actions. Though moderation systems have been developed in the past, they are either completely proprietary, therefore unavailable for public use and study (e.g., internal tools at Facebook [33], The New York Times[3], YouTube [34]), or they are not supported by AI (e.g., Twitter Blocklists [100], The Coral Project[4], HeartMob [113], and SquadBox [114]). In Chapter 6, I introduce Crossmod, the first open source, AI-backed moderation system to be released publicly.

2.6 In-domain approaches to moderate antisocial behavior

Prior research has looked at technical approaches to moderating online antisocial behavior. All of them tend to focus on in-domain methods to study different kinds of antisocial behavior and develop strategies to counter them. Studies have shown that antisocial behavior like undesirable posting [59, 58, 57], and textual cyberbullying [55, 56] can be identified based on the presence of insults, user behavior and topic models. In recent years, politeness has been studied in online settings, to help keep online interactions more civil. Researchers have built a *politeness* classifier using a computational framework for identifying linguistic aspects of politeness [115]. While these methods are effective within-domain, learning

[3]https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html
[4]https://coralproject.net

across domains or *communities* remains an open question.

2.6.1 Challenges faced by current work

Current moderation techniques employed by researchers and community moderators face key challenges. Supervised detection techniques require labeled ground truth data for building and evaluating a model [116]. These data are difficult to obtain, and manual annotation is a common approach to address this challenge. But this task requires a large amount of manual labor to hand-annotate (or label) the data. This method is also inherently subject to biases in the annotator's judgment, which could affect the quality of the analysis results [117].

A constant struggle is to identify good data sets that researchers can study. Communities do not publicly share data containing moderated content due to privacy and public relations concerns. This restricts data access, and makes it difficult to model the types of abuse present in a community. In addition, new and emerging online communities lack enough data from their respective users. A new community has few contributions, by definition, and therefore even fewer labeled examples; this does not allow them to build robust automated detection systems for identifying abusive content. As a result, building cross-domain moderation systems remains a challenge. Yet, studies have shown that it is important to define community tolerance for abusive behavior as early as possible [106].

My thesis aims to address many, but not all, of these challenges. In particular, *cross-community learning* allows online communities to piggyback on the data of others, requiring far fewer (and perhaps no) labeled training examples. BoC may form the backbone of cross-domain classifiers built on the data of many Internet communities.

CHAPTER 3
THE BAG OF COMMUNITIES: IDENTIFYING ABUSIVE BEHAVIOR ONLINE WITH PREEXISTING INTERNET DATA

Machine learning-based approaches can help by algorithmically triaging comments for a much smaller number of (perhaps paid) human moderators; yet, they typically require vast amounts of labeled training data. Moreover new and emerging online communities lack enough data generated by their own users to develop effective machine learning-based solutions to moderate content, especially in their initial stages. This chapter bridges this data gap by introducing a new analytic concept for studying and building online communities: the *Bag of Communities* (BoC) approach. In brief, BoC aims to sidestep site-specific models (and their data) by computing similarity scores between one community's data and preexisting data from other online communities. Next, I introduce the concept of BoC, and use it in an "existence proof:" identifying abusive posts from a major online community.

3.1 Bag of Communities (BoC)

In this section, I define a new approach to identify certain kinds of online behavior by leveraging large-scale, preexisting data from other Internet communities. The intuition behind my approach is to use the similarity of a post to a known, existing community as a feature in later classification. For example, a post that seems at home within a corpus of 4chan posts may likely be inappropriate for npr.org.

First, I define a method to compute *cross-community similarity* (CCS), a building block of my approach. I then introduce a new model where a variety of CCS data points act in concert to aid predictions in a new community. Analogous to the well-known Bag of Words representation, I call this the *Bag of Communities* approach.

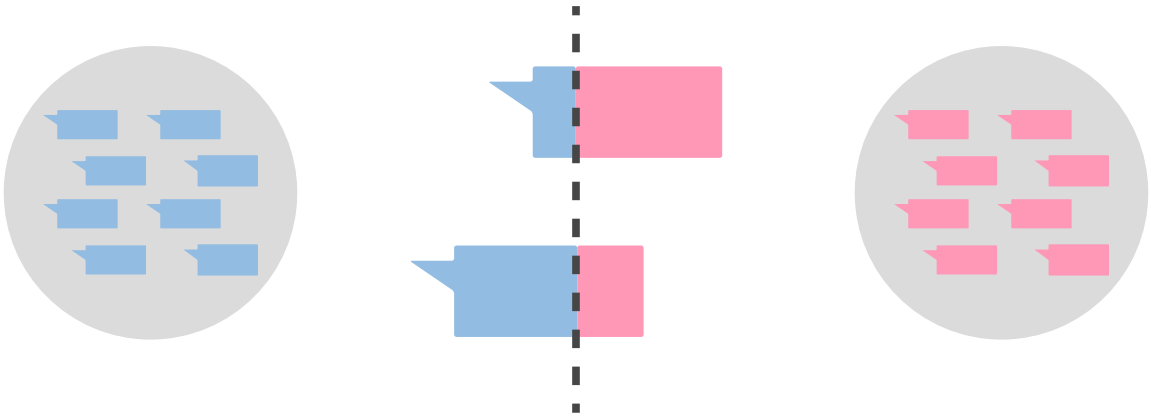

Figure 3.1: A conceptual illustration of Bag of Communities approach, here with two source communities employed. When new and unlabeled posts are generated in a community, similarity scores can be assigned by comparing them to preexisting posts from other communities (*blue* and *pink*, in this example). A downstream classifier uses similarity scores to make predictions, in my case about abusive behavior.

3.1.1 Cross-Community Similarity (CCS)

Let S be a source community, with whose data we will compare a community of interest—or target community T. While one could approach representing S and T in a variety of ways, it seems natural to model S and T via their posts: let $p \in R^n$ be a vector-space representation of a post in n dimensions. S and T then comprise all vectors corresponding to their constituent posts. One dimension might represent whether the post was created on a weekday, another might represent whether it contains the word "happy," another might represent whether the post contains an image, etc.

S and T could represent posts along a variety of (possibly infinite) dimensions: temporal characteristics (burstiness vs. spread-out), posting medium (textual vs. image-centric), network structure (connected vs. disconnected), identity (anonymous vs. identifiable), community norms (supportive vs. judgmental). In this chapter, I focus on a linguistic representation: the words and phrases used in S and T serve to define S and T. That is to say, a post is represented as a vector with 1's connoting a word or phrase's presence, and 0's otherwise.

There are as many ways to compute $CCS(S, T)$ as there are to compute similarity between vector spaces [118, 119, 120]. Its application may drive the particular method. For example, a straightforward approach might involve computing the centroids s and t of S and T, respectively, and next computing $cos(\theta)$ for the angle θ between them. However, I adopt an approach in this chapter inspired by Granger causality [121]. Let M_S be a statistical model that predicts (real-valued) membership in S. $CCS(S, T)$ is then the information provided by $M_S(p)$ in predicting membership in T, for some post p. In other words, I let a model predicting membership in S to predict membership in T. This is analogous to the Granger-causal idea of letting one time series at time t predict the value in another time series at time $t + k$. By "information provided by $M_S(p)$," I mean that M_S may not be used directly, but as the raw material of some encapsulating function. The range of $CCS(S, T)$ is $[0, 1]$, with $CCS(S, T) = 0$ for entirely dissimilar communities and $CCS(S, T) = 1$ for entirely similar communities.

3.1.2 Bag of Communities definition

In a Bag of Communities representation, a post $p \in T$ generates CCS scores $CCS(S_1, T)$, $CCS(S_2, T)$, $CCS(S_3, T)$, ... for a variety of source communities S_1, S_2, S_3, ... A Bag of Communities model develops a function $f(CCS(S_1, T), CCS(S_2, T), CCS(S_3, T), ..., T)$ that maps these CCS scores and local, site-specific information to a prediction in $[0, 1]$.

In other words, as illustrated in Figure 3.1, a post might be compared against 4chan, MetaFilter, hateful subreddits, etc. These scores are then fed to another, higher-level classifier that also takes site-specific information into account. In other words, an ensemble classifier might use site-specific information (e.g., the words and phrases used in *that* community) along with CCS scores representing similarity to 4chan, MetaFilter, etc. to make a prediction. Figure 3.2 illustrates the overall process in function notation, mapping the domain of source communities to the range of the target community.

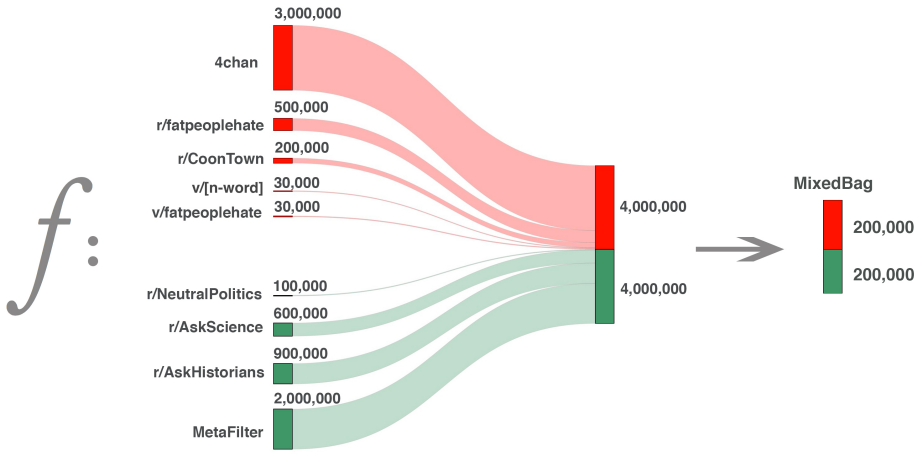

Figure 3.2: An illustration of the overarching Bag of Communities concept, along with the approximate number of posts collected from each source community in my empirical work. Cross-community similarity values are obtained by comparing target community posts to preexisting posts from source communities ($CCS(S_i, T)$). Communities in *red* were selected because I hypothesized they contained abusive content and those in *green* because they are well-moderated. The goal is to learn a function that maps the source communities to the target community.

3.2 Source and Target Communities

Next, I explore using the BoC approach in predicting abusive behavior in a new online community. I draw data from 9 communities from 4chan, Reddit, MetaFilter and Voat—with MixedBag[1] as my *target* community. My BoC models will aim to learn from content on source communities, and make predictions about a post's likelihood of being labeled as abusive on MixedBag.

Next, I describe my source and target communities in more detail, and explain the motivation behind using each community to build my BoC models.

[1]Pseudonym for the actual target website/community, as per research agreement.

3.2.1 Source: 4chan's /b/ and /pol/

4chan is made up of themed online discussion boards, where users generally post anonymously. 4chan is popularly known as the "Internet hate machine" [122], and "the rude, raunchy, underbelly of the Internet" [123]. The use of racist, sexist and homophobic language is common on 4chan. Groups are often referred to using a "fag" suffix (e.g., new members are "newfags", British users are "britfags"), and a common response to any self-shot picture by a woman is "tits or GTFO" [40].

/pol/ is 4chan's *politically incorrect* board. As per 4chan's rules page, /pol/ is a board where debate and discussion related to politics and current events is welcome. /b/ is 4chan's "random" board, and is 4chan's first and most active board, representing 30% of all 4chan traffic. In the words of its creator, /b/ is the "life force of the website," and a place for "rowdiness and lawlessness" [124]. These boards are infamous for exhibiting a range of explicit content. Despite being a funny, open and creative board that is credited for the creation and promotion of numerous memes, the content on /b/ is frequently intentionally offensive, with little held sacred.

3.2.2 Source: Reddit's r/fatpeoplehate and r/CoonTown

In the wake of Reddit's new anti-harassment policy, the website banned several hate communities that it found in violation of the site's rules [125, 126]. According to Reddit's announcement, "We will ban subreddits that allow their communities to use the subreddit as a platform to harass individuals when moderators don't take action." [127]

I collected posts from two of Reddit's most controversial communities which routinely engaged in hate speech, namely *r/fatpeoplehate* and *r/CoonTown*. r/fatpeoplehate is a fat shaming community devoted to posting (among other things) pictures of overweight people for ridicule [126]. It was one of the most prominent removals from Reddit, and had 151,404 subscribers at the time of its banning, as reported by Reddit Metrics.[2]

[2]http://redditmetrics.com/r/fatpeoplehate

r/CoonTown is a racist subreddit dedicated to violent hate speech against black people. It contained "a buffet of crude jokes and racial slurs, complaints about the liberal media, links to news stories that highlight black-on-white crime or Confederate pride, and discussions of black people appropriating white culture." [128] It had 21,168 subscribers at the time of banning, as reported by Reddit Metrics.[3]

3.2.3 Source: Voat's v/fatpeoplehate and v/[n-word]

Voat is a media aggregator website which claims to emphasize free speech above all other values. Following Reddit's banning of subreddits for violating its harassment policy, users from those banned communities migrated to Voat, creating hate subverses to take the place of their banned subreddit counterparts [129, 130]. In particular, I collected posts from v/[n-word] and v/fatpeoplehate, which are the Voat equivalents of r/CoonTown and r/fatpeoplehate on Reddit.

3.2.4 Source: MetaFilter

In addition to sites like 4chan and Voat, I also try to use well-moderated sites, like MetaFilter, as distractors or counterexamples of abusive content. MetaFilter requires $5 to establish an account, and is one of the most strictly moderated communities on the Internet. Moderators hide inappropriate material quickly, and reinforce positive norms by making good behavior far more visible than bad [131]. Whenever needed, moderators step in and temporarily suspend an offending user's account.

3.2.5 Source: r/AskHistorians, r/AskScience & r/NeutralPolitics

r/AskHistorians and r/AskScience are communities that are actively moderated, and have well-defined rules regarding user behavior and interactions on the subreddit. These rules are regularly enforced by moderators and exist to ensure that debates on the subreddit do

[3] http://redditmetrics.com/r/CoonTown

not devolve into personal insults or ad hominem attacks.

r/AskScience urges its users to "Be civil: Remember the human and follow Reddiquette", in its guidelines [132, 133]. r/AskHistorians has a strict "Civility" rule which says, "All users are expected to behave with courtesy and politeness at all times. We will not tolerate racism, sexism, or any other forms of bigotry. This includes Holocaust denialism. Nor will we accept personal insults of any kind." [134]

r/NeutralPolitics is a well-moderated community "dedicated to evenhanded, empirical discussion of political issues." The community urges its users to be courteous in its comment rules,[4] which states that "Name calling, sarcasm, demeaning language, or otherwise being rude or hostile to another user will get your comment removed."

3.2.6 Target: MixedBag

My collaborators and I have a research partnership with a large online community who provided data moderated off-site for violating abuse policies. Getting data such as these is typically a major hurdle, as companies fear the blowback that may occur after its release. As per our partnership agreement, I will refer to this target community using a pseudonym: *MixedBag*. The community has on the order of 100M users, and is typical of user-generated content sites: the site has profiles, posts, comments, friends, etc. I obtained comments that were deleted by the site's moderators as abusive, and flagged by users, as part of this partnership.

A notable challenge is that *a priori*, the target and source sites share little in common. For example, MixedBag is a pseudonymous community where conversation is structured into threads of comments, in response to a piece of shared content; 4chan is an image board where anonymous people often post short and unrelated phrases in response.

[4] https://www.reddit.com/r/NeutralPolitics/wiki/guidelines

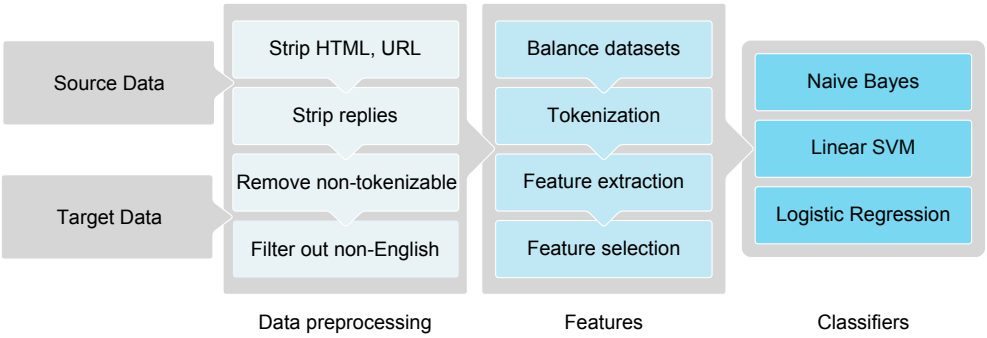

Figure 3.3: Flowchart depicting the overall CCS model-building pipeline. After collecting Bag of Communities and MixedBag data, text undergoes a number of preprocessing steps before acting as input for three different classifiers. Each CCS classifier tries to distinguish a source community's posts from a random background cohort of distrators.

3.3 Data

I collected data from each of my source communities, as well as data from MixedBag as target data. In my static model, I use the source and target datasets as classic train and test datasets. In the dynamic model, I iteratively allow a model trained on source data to update itself as it sees new batches of target data.

3.3.1 Source data

I collected varying amounts of data from each source community, as it was available:

- **3M posts** from 4chan /b/ and /pol/ boards, spanning 14 months in 2015 and 2016
- **700K posts** from r/fatpeoplehate and r/CoonTown, spanning January to July 2015
- **70K posts** from v/fatpeoplehate and v/[n-word], spanning August 2015 to February 2016
- **2M posts** from MetaFilter, which contains all posts archived on the site, spanning July 1999 to July 2015

- **1.5M posts** from r/AskScience and r/AskHistorians, spanning 2007 to 2015
- **130K posts** from r/NeutralPolitics, spanning 2007 to 2015

I also obtained **3.5M** random comments from MixedBag, which were publicly available at the time of data collection. The comments serve as *distractors* for building a BoC model: they represent a random sample of the site's publicly visible comments. In total, I collected over **10M** posts to serve as training data using a variety of archives and crawlers. *Note:* my training phase does not give static models access to comments moderated from MixedBag.

3.3.2 Target data

To evaluate my model, I obtained the text in **200,000** *moderated comments* from MixedBag. The dataset contains over 4 years of human-curated data—comments moderated off-site by moderators and users for violating abuse policies. As mentioned before, these data were given to me by MixedBag as part of a research partnership. I also obtained the textual content of **200,000** random MixedBag comments, which were still present online during the time of data collection, using the same procedure as in the section above. Note that there is no overlap between (on-site) MixedBag comments used for training and testing.

To provide readers a sense of the types of comments moderated off-site, the following randomly-sampled ones represent typical instances. Readers are forewarned that most are offensive and "not safe for work" (NSFW):

> SO I WILL LOOK YOU'RE FAMILY UP ON WHITEPAGES AND MURDER YOU!!
>
> Lol, go kill yourself
>
> go fuk yourself ugly
>
> YOUR GRANDFATHER IS BURNING IN HELL KIKE!!
>
> hehehe! we're gonna have a lot of fun with this! now, lie on your back.
>
> This is full of fail and AIDS!
>
> awwwww what a cute [n-word]

3.4 Applying BoC to Abusive Behavior Online

I used the data collected from my *Bag of Communities* and MixedBag to build and evaluate multiple machine-learning models. In this section, I will discuss the components of my BoC model and steps in the model's pipeline. For reference and overview, Figure 3.3 visualizes the pipeline for training every internal CCS estimator.

3.4.1 Data preprocessing

I began preprocessing the data sets by stripping replies, HTML elements and URLs in the collected comments. Next, I discarded posts that were not tokenizable. These were comments that were either not in Unicode or did not contain any text/tokens. Finally, I performed language detection and discarded comments that were not in English. I used *langdetect* [135], an open-source Java library, for language identification.

3.4.2 Balancing datasets

After the preprocessing steps, I shuffle the datasets and balance them to ensure an equal number of posts from each class. Note that balancing the number of samples from each class likely does not mimic real-world situations. In general, abusive posts are relatively rare. However, balancing across all conditions ensures that I can easily interpret model fits relative to one another. In other words, since the in-domain model will also act on the balanced datasets, balancing will not privilege either approach.

3.4.3 Tokenization & feature extraction

I tokenize comments, and break the text contained in each comment into words. Using these words, I go on to build the vocabulary for all comments in the sample. Each comment is represented as a feature vector of all words and phrases present in the vocabulary (i.e., a *Bag of Words* (BoW) model). The feature values are either the binary-occurrence values

Table 3.1: Grid of parameter values used when running classification tests to find the best combination of parameter values for my model. The best values shown for all the parameters, found with a grid search, were used in all classifiers. *max features* refers to the upper limit placed upon the hashing vectorizer.

Parameter	**Values**	**Best Value**
n-gram range	[(1,1), (1,2), (1,3)]	(1,3)
binary	[on, off]	off
lowercase	[on, off]	on
max features	$[2^{22}, 2^{26}]$	2^{26}
tf-idf	[on, off]	on
alpha	[0.1, 0.01]	0.01
feature selection	[on, off]	on
top k	$[100, 10^3, 10^4, all]$	10^4
classifier	NB, LinearSVC, Logit	NB

(present or not) or the frequency of occurrence. I extract n-grams ($n \in [1, 2, 3]$) from the text and perform vectorization using a Hashing Vectorizer. Hashing Vectorizers create a mapping between tokens and their corresponding feature value (TF-IDF) [136].

3.4.4 Feature selection

I compute the ANOVA F-values for the provided sample, and select the most distinguishing features using the F-value between features and labels. I perform feature selection on features in the *top k* in $[100, 10^3, 10^4, all]$. For example, when *top k* $= 10^3$, only the top $1,000$ BoW features are selected based on their ANOVA F-values.

3.4.5 Classifiers

To build internal CCS estimators, I ran classifications tests using three different classifiers: *Multinomial Naive Bayes* (NB), *Linear Support Vector Classification* (LinearSVC), and *Logistic Regression* (Logit). My BoC models use the output likelihoods from these classifiers as internal estimators.

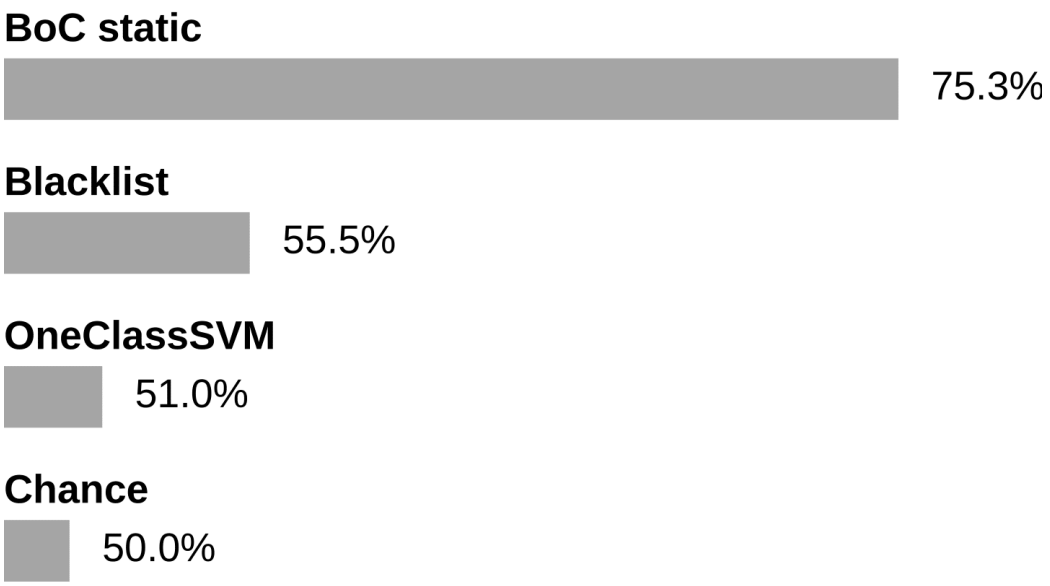

Figure 3.4: Accuracy values for baselines and the BoC static model. *Chance* refers to a random classifier.

3.4.6 Parameter search

I ran classification tests on the datasets using different settings to identify the best configuration for my BoC model. I performed a grid search on a held out 5% sample of my data, provided by *GridSearchCV* [136]; it exhaustively generates candidates from a grid of parameter values specified in the grid shown in Table 3.1. For example, *n-gram range* refers to the range of n-grams extracted, and with *n-gram range* $= (1, 3)$, I extracted uni-grams, bi-grams and tri-grams. These parameter values were used to find the best configuration for my models, and are used in all subsequent phases.

3.4.7 BoC static model

I explore two models in this work. The first I call the "BoC static model," a model that sees no target data from MixedBag. This BoC static model trains the underlying CCS estimators, but gets no access to test data; therefore, it resembles a pre-trained model that could be deployed "off the shelf," similar to how blacklists are often used in practice today.

Table 3.2: Precision, recall and accuracy for different models. The dynamic (online learning) models were trained on 100,000 test samples.

Model	Precision	Recall	Accuracy
BoC static	77.49%	71.24%	75.27%
In-domain	88.20%	91.66%	89.77%
Only abuse BoC dynamic	95.04%	85.85%	**91.18**%
All BoC dynamic	91.09%	87.93%	90.20%

I built two different baselines to compare with my BoC static model, and arrive at performance (lower) bounds.

3.4.8 BoC static baseline: Blacklist

I first trained a model to classify an input comment as abusive or not based on the presence of blacklisted words. I obtained a list of profane terms used in previous work [137]. Such list-based detection mechanisms are commonly deployed in the wild. This model essentially checks for the presence of at least $threshold$ number of blacklisted term(s) in the comment. I tested the model for all values of $threshold \in [1, 2, 3, \ldots]$.

3.4.9 BoC static baseline: OneClassSVM

In the absence of labeled, rare and supposedly different data points, one known approach is treating such points as outliers of a known distribution. In my case, I trained a OneClassSVM [136] to learn the distribution of n-grams for naturally-occurring MixedBag posts, in the hope that abusive posts will deviate from this distribution. The OneClassSVM was trained on just 3.5 million random MixedBag posts, and tested on the target data from MixedBag. The parameter configurations used are shown in Table 3.1.

3.4.10 BoC dynamic model

In addition to the static model, I also explore a "dynamic" (or online learning) model that iteratively sees more and more target community data to aid prediction. This mimics what

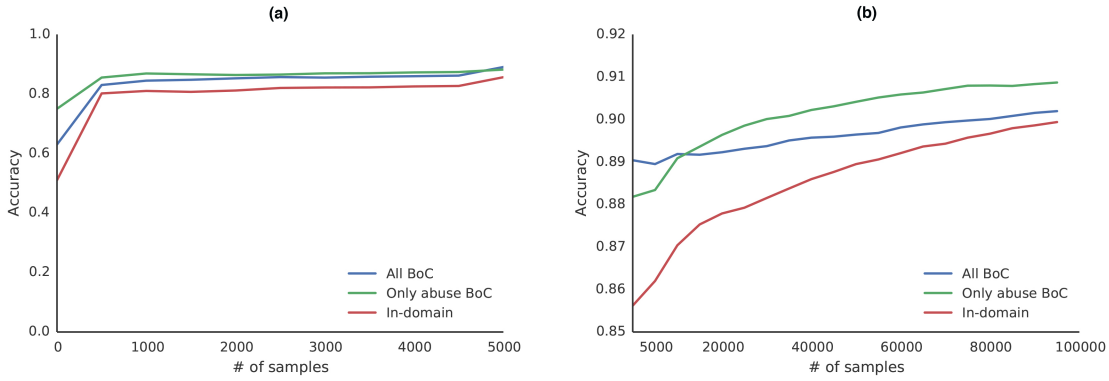

Figure 3.5: Dynamic model performance when trained only on target community (MB) data and including CCS, BoC features. *In-domain* denotes the plain partial fit model that uses only MB data, *Only abuse BoC* denotes the dynamic model only using communities that are hypothesized abusive, and *All BoC* denotes the dynamic model using all communities in my dataset. Performance of the models when iteratively trained on up to 5,000 target community samples are shown in (a), and the remaining batch sizes in (b). The plots are separated for better resolution, and (b) is scaled up for clarity.

an upstart community might face when building its own abuse detection models as new moderator labels come in. The BoC dynamic model uses these data in conjunction with internal CCS estimators to make final predictions. That is, it has access to the cross-community similarity scores, $CCS(S_i, T)$, described earlier, which gives the likelihood of a (target) post belonging to a (source) community S_i.

In particular, the dynamic BoC model is provided with similarities to each source community, $CCS(S_i, T)$, which is the $predict\_proba()$ returned by an internal estimator (NB) trained on source community data. These probability scores are used as features, in addition to textual features learned from the target community data by the final estimator (also NB), which predicts whether a given post is abusive or not. I compare it against a purely in-domain linguistic model with the same parameter setting. Both models—the online, in-domain model and the dynamic BoC model—are trained the same way, on a fractional batch of the target data, and then evaluated on the remaining (unseen) target data.

More formally, at a given iteration where the model sees a fractional batch of size f of the target data, the models are constructed as follows:

$$M_{in} : BoW(f \cdot MB)$$

$$M_{dyn} : BoW(f \cdot MB) + CCS(S_i, T), \forall S_i \in BoC(all)$$

$$M_{dynabuse} : BoW(f \cdot MB) + CCS(S_i, T), \forall S_i \in BoC(abuse)$$

All models build BoW linguistic models of the data to which they have access so far. The *in-domain* model (M_{in} above) is trained only on posts from the target community, and does not see any of the BoC data. The *all BoC dynamic* model (M_{dyn} above), is trained on posts from the target community, in addition to CCS (internal) estimations from all 9 source communities. Whereas the *only abuse BoC dynamic* model ($M_{dynabuse}$ above) uses CCS estimations from only the abusive communities (i.e., 4chan, r/fatpeoplehate, r/CoonTown, v/[n-word], v/fatpeoplehate).

I aimed to observe the growth in accuracy of predictions over time (as more and more moderated posts from the community are available for training the model) and understand when the performance values saturate.

3.5 Results

While I ran trials with three different classifiers (see above), Multinomial Naive Bayes (NB) performed best in all conditions. The simplest model, its performance may reflect its limited ability to overfit the training data. Hereafter, I report results for the NB model across conditions. The parameter values used for the best model are available in Table 3.1.

3.5.1 BoC static model performance

I compared the performance of my best BoC static model with two different baselines. Figure 3.4 displays the accuracies across models. I observed that the Blacklist gave a best

performance of 55% (with threshold 1), while the OneClassSVM achieved an accuracy of 51%. My BoC static model performed at 75.3% accuracy.

3.5.2 BoC dynamic model performance

The *BoC dynamic* online learning models performed uniformly better than a purely *in-domain* model built only using moderated posts from the target community. The differences in performance of the *in-domain*, *all BoC dynamic*, and *only abuse BoC dynamic* models at various stages of data access are shown in Figure 3.5. At 0 test samples seen, the in-domain model performed at 51% accuracy (it is equivalent to a single-class classifier used to detect outliers, without any access to moderated posts). The BoC dynamic models outperformed the purely in-domain model even after 100,000 (moderated) test samples were seen. The best performing BoC dynamic model achieved 91.18% accuracy, after seeing 100,000 (moderated) test samples. At all batch sizes measured, the differences are statistically significant.

3.6 Discussion

I find that the Granger-causal, CCS-based, *Bag of Communities* models perform well in both static and dynamic settings. The static model likely performs well enough right now that it could be deployed as is with human oversight on a new community; the dynamic model uniformly outperforms purely in-domain classifiers with access to years of curated data. This means that models operating entirely on out-of-domain (4chan, Reddit, Voat and MetaFilter) data can learn significant cross-domain knowledge applicable to a community the model has never seen before. Given that I performed no domain adaptation [138, 139, 140], this result signals deep overlap between, for instance, large-scale preexisting Internet data and comments on another site.

I do not intend to intimate with these results that sites should substitute a BoC model for their existing moderation systems. Rather, this chapter presents a promising empirical

result about the utility of using preexisting community data to inform abuse detection. It suggests that gathering data from other communities could be extremely useful. Next, I reflect on my models, discuss some of their error patterns, strategize about selecting source communities, and conclude with reflection on how designers and researchers could use BoC models.

3.6.1 Reflection on models

In post-hoc inspection I observed that my BoC-based model identified a significantly larger variety of *abusive* content than the other models. This is in accordance with the high precision values achieved by the BoC classifiers when classifying abusive content (see Table 3.2). This derives from the source communities. For instance, the BoC data provides background information not available to the in-domain model, ranging from popular Internet phrases (e.g., "full of fail", "FOR THE LULZ") and terms (e.g., "desu", "nips") to variants of commonly used terms (e.g., "fuk"). Most of these comments were not identified by the in-domain model, as it sees only a handful of such terms in MixedBag posts.

As seen in Table 3.2, the BoC exhibit better precision-to-recall trade-offs than purely in-domain models. That is, they naturally seem to trade recall for precision more than the in-domain models. In discussions with site operators, this seems to be the way they would prefer the model's error patterns to behave. As many social media companies are owned and operated in the United States, concerns about censorship understandably pervade discussions around moderation [141]. High precision models (i.e., if the model declares it "abusive," then it very likely is, even at the cost of missing more abusive posts on average) would fit well in this context.

3.6.2 Error analysis

Both the in-domain and the BoC models missed a significant fraction of abusive posts. In an error analysis, many of them used character-level substitutions to evade automatic filters

(e.g., “f**king”, “f**k”), but were identified by human moderators on MixedBag. You could imagine normalization filters that help to uncover substitutions like these [105], a fruitful area for future work improving these models and data pipelines. Run of the mill spam also seemed to evade all models, suggesting that a future enhancement would be to add existing spam filters to data pipeline in Figure 3.3.

Some moderated posts were sarcastic in nature, and automatic detection of sarcasm is an open research problem [142, 143]. While neither the in-domain models nor the BoC could catch these instances, they were identified by human moderators on MixedBag:

> if i had a dollar for every pixel in this picture, i’d have 50

> Oh mai gowd I have never been so enlightened in mai hole laif.

> aww your going blind ???

Reflecting the noise of the real world, I also observed the presence of non-abusive posts in my test samples, which were (perhaps wrongly) removed by site moderators. Sometimes, moderators delete entire threads of comments, posted in response to inflammatory or offensive (parent) posts. Examples of likely mislabeled data:

> This is really cool! Superb job < 3

> Wonderful job

> Its WOW!

> Aww! You should nominate him for [award]! See [link] for details, okay?

3.6.3 Best performing model: Only abuse BoC

r/NeutralPolitics, r/AskScience, r/AskHistorians and MetaFilter are all well-moderated communities. I observed that training the *All BoC* model including data from these communities, in addition to the hypothesized abusive communities, increased the number of false

positives (i.e., non-abusive posts being misclassified as being abusive). This can be attributed to the fact that typical (onsite) content found in MixedBag is more similar to 4chan, v/[n-word], v/fatpeoplehate, r/fatpeoplehate and r/CoonTown, than the former. In other words, the content found on the former communities are too polite (or well-moderated), and observed to not be representative of the normative behavior in the target community. As a result, the *Only abuse BoC* model achieved the best accuracy in my tests.

3.6.4 Choosing source communities

The choice of 4chan boards, hate-filled subreddits and subverses as source communities required some community-level insight. The intuition that many of these communities perform "bad behavior" motivated my data collection. I have also looked at a variety of well-moderated communities like MetaFilter, r/AskHistorians and r/AskScience, and r/NeutralPolitics as counterexamples.

How do you choose the community data required for BoC? At present, there is some "black magic" involved in collecting the right communities so as to be useful for a given context—not unlike the infrequently discussed black magic surrounding feature engineering in many applied machine learning contexts. For the moment, I believe this will be driven by the problem at hand. Intuition and domain knowledge will likely drive BoC data collection, and more work should be done to explore how to reduce search and collection costs. While community data such as this only needs to be collected once, it does require some investment of time and energy to write crawlers, debug them, etc.

However, it is possible to envision scenarios where many of the Internet's most important and popular communities have been crawled, stored, and used in training CCS classifiers. For example, given these encouraging results, I will explore simply building a CCS classifier for every subreddit, for all contributions ever posted to that subreddit, in the next chapter.

3.6.5 Design Implications

I have released my models under an LGPL license. I host the pre-trained internal CCS estimators used for both the static and dynamic models I presented here.[5] Even at this early stage in the research, I believe the present BoC models can be used to address the challenge of moderating new, emerging and established communities. As demonstrated in this conceptual and empirical work, the BoC approach may allow communities to deal with a range of common problems, like abusive behavior, faster and with fewer engineering resources.

Communities themselves may choose to operationalize the models I provide in a number of ways. A site might choose to wrap them in a human-in-the-loop system, where moderators review comments triaged by BoC score. Another community with fewer resources may automatically hold comments that score above a certain threshold, requiring users to petition to have them looked at by a human moderator. I hope that the open source models provide the kind of flexibility necessary for site operators to build these and other moderation approaches.

3.6.6 Theoretical Implications

As alluded to in the *Introduction*, the BoC approach may enable new kinds of theoretical advances. Returning to my example of a maximal BoC classifier that returns thousands of similarity features for all of Reddit, these scores place communities in a metric space: they are embedded in a high-dimensional space either closer or farther away from other communities. In this way, this approach might allow a more theoretically-inclined researcher to extract latent clusters of online communities into a taxonomy. An advantage of using semi-automated clustering techniques such as this one is that it would update as the Internet changes, without requiring another full round of researcher effort. I hope to see theoretical work exploring approaches along these lines.

[5] https://bitbucket.org/ceshwar/bag-of-communities.git

3.7 Limitations & Future Work

While I find these results encouraging, they raise a number of questions, challenges and issues. Here, I reflect on some of the limitations present in my work, with an eye toward how future research might build upon it.

Focus on linguistic information. The empirical part of this chapter only examines the similarity between words and phrases in two communities. I have left out many pieces of data including temporality, other media such as images, etc. Work exploring and including these data would paint a richer picture of communities, and very possibly aid in prediction tasks. For example, it seems very reasonable to assume that reply chain dynamics (i.e., fast-arriving replies vs. slow-arriving replies) might interact with a post's likelihood of exhibiting abusive properties.

Focus on one target community. Here, I only look at the transference of information between 4chan, Reddit, Voat, MetaFilter, and one community, MixedBag. While I find the results encouraging and surprising, whether BoC-based techniques will work with other communities remains open. While I see it as a positive to have obtained MixedBag data at all, given the challenges facing companies who would release it, I encourage other researchers to explore extending the BoC approach to other target communities. For example, I will extend this work to moderated comments from various subreddits (observable through the open Reddit API), in the next chapter.

Human-in-the-loop systems. In real situations, human moderators will still need to supervise any automated triage system, including ones built on BoC. It remains open exactly how well BoC would fare in situations like these, and how best to design it to do so. I see this as very profitable future line of work for human-centered computing researchers, as it will certainly involve talking with and understanding the work of professional and amateur community moderators. As a first step in this direction, I build a *mixed-initiative* moderation system to assist (volunteer) moderators on Reddit, by working closely with a

group of partner subreddits, which I will detail in Chapter 6.

3.8 Conclusion

I presented a new method for transferring one community's data to another—the *Bag of Communities* approach—and apply it to a key challenge faced by communities and moderators. To the best of my knowledge, the BoC representation is novel within social computing and applied machine learning more generally. Further, I presented an existence proof of Bag of Communities: building a classifier for identifying abusive content using data from 4chan, Reddit, Voat and MetaFilter. I hope that other designers find BoC useful in predicting a variety of important online phenomena; I hope researchers can make use of it to examine other communities with the Bag of Communities representation.

CHAPTER 4

THE INTERNET'S HIDDEN RULES: AN EMPIRICAL STUDY OF REDDIT NORM VIOLATIONS AT MICRO, MESO, AND MACRO SCALES

4.1 Introduction

An online community's norms play an important role in guiding acceptable behaviors, and therefore in its governance [38]. Online community moderators have to sanction pedestrian normative violations like posting spoilers about a TV show, as well as more serious infractions like online abuse [2], harassment [39, 17, 18], and fake news and misinformation [42]. Yet, norms for what is appropriate can vary widely from one community to another. Even behavior considered harmful in one community might be celebrated in another (e.g., 4chan's /b/ [40], Something Awful Forums [41]).

4.1.1 Regulating behavior on Reddit

In this chapter, I study norms across a wide variety of communities on Reddit. Reddit is an assemblage of over one million online communities[1] known as *subreddits*. Subreddits can be created by anyone, and they are moderated by members of the community. They exist for almost any topic, including specific sports (e.g., r/nba), science (e.g., r/science), TV fan theories (e.g., r/gameofthrones), and standing cats (e.g., r/standingcats).

Reddit has a multi-layered architecture for regulating behavior on the platform. It has site-wide content[2] and anti-harassment[3] policies that all subreddits are expected to follow. In cases where there are violations of some of these policies, Reddit is known to ban subreddits and user accounts [39]. In addition to Reddit's content policies, each subreddit has

[1] `http://redditmetrics.com/history`
[2] `https://www.redditinc.com/policies/content-policy`
[3] `https://redditblog.com/2015/05/14/promote-ideas-protect-people/`

its own set of subreddit-specific rules and guidelines regarding submissions, comments, and user behaviors [54]. Moderators (or "mods") enforce the rules and guidelines.

4.1.2 Community norms on Reddit

Rules and norms are loosely coupled on Reddit, with subreddit moderators sometimes turning (often implicit) norms that are enforced behind-the-scenes into rules that face the community. While rules tend to be explicit, norms are emergent, arise from interaction over time, and respond to current demands on a community [144]. Community norms play an important role in online moderation, and moderating online communities is strongly contextual because norms can vary widely between communities. An understanding of community norms is generally gained through experience [145]: observing posts and comments posted on the subreddit, peer feedback in the form of votes or replies to comments, and interactions with mods. This work of enforcing norms is important to both communities and platforms: people may leave sites and communities after being the victims of norm violations [16]. Importantly for the present work, norm enforcement by mods also creates a record of norm violations across disparate communities.

4.1.3 Summary of methods and findings

In this chapter, I study community norms on Reddit with a large-scale, empirical approach. By working from over 2.8M comments removed by moderators of 100 subreddits over 10 months, I use both computational and qualitative methods to identify three types of norms within Reddit: *macro* norms that are universal to most parts of Reddit; *meso* norms that are shared across certain large groups of subreddits; and *micro* norms that are specific to individual, relatively unique subreddits.

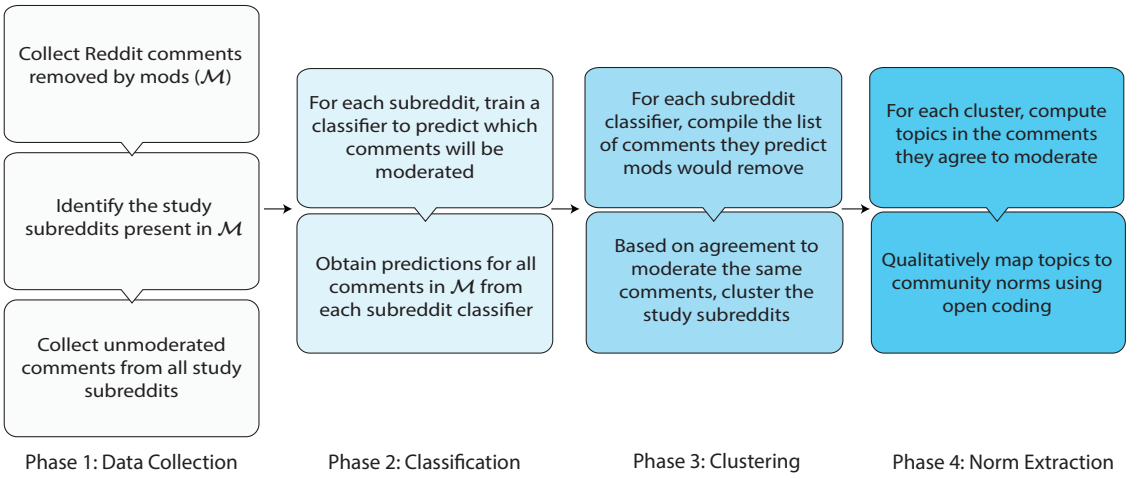

Figure 4.1: Flowchart depicting the different phases of my research pipeline. $\mathcal{M}$ denotes all the moderated Reddit comments I collect in *Phase 1*, and *mods* denote the subreddit moderators on Reddit. The final output derived from *Phase 4* gives me the different community norms on Reddit.

4.2 Data

Next, I transition to my dataset construction, and describe the procedure I use to collect Reddit comments removed by moderators. An illustration of this approach is shown in Figure 4.2.

4.2.1 Moderated comments from Reddit ($\mathcal{M}$)

I construct a dataset that includes all Reddit comments that were moderated off-site[4] during a 10-month period, from May 2016 to March 2017, in a three-stage process.

Stage 1: Stream Reddit comments into master log file continuously.

I use the Reddit streaming API [5] to crawl all comments as they are posted on Reddit on a continuous basis. These are all comments posted to r/all, which can be from any subreddit that is not "private," and chooses to post its content to r/all. As I keep streaming comments

[4]Therefore, these comments are no longer publicly visible on the internet.
[5]`https://praw.readthedocs.io/en/latest`

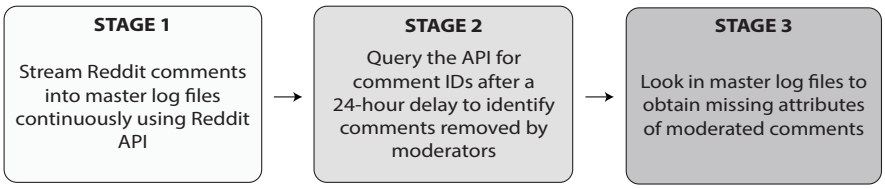

Figure 4.2: Flowchart depicting the different stages involved in my collection of moderated (and unmoderated) comments from Reddit.

continuously, I store all of the data in a master log file.

Stage 2: Query the API for comments after a 24-hour delay.

After a 24-hour delay, I query the Reddit API for each comment in my master log file that was collected in the past day, using a comment's unique *comment_ID*. If a comment is removed by a moderator, then the text that was previously present in the comment (represented by the "body" field) is replaced with **["removed"]**.

Via conversations with Reddit moderators and an inspection of Reddit's source code [21] , I know that only when moderators or admins remove a comment, its text is replaced by ["removed"], and most comments violating norms are moderated within the first 24 hours of posting on the subreddit. Note that a comment removed by a moderator (either auto- or human moderators) is different from an author removal, and I can distinguish between the two by looking at the text of the comment. The text in comments deleted by the authors is replaced by ["deleted"], while only moderator removals are replaced by ["removed"]. Using this method, I compile the "*comment_IDs*" of all comments that were removed by the moderators in the previous day.

Stage 3: Look in master log file to obtain missing attributes.

For each removed comment I identify in the previous stage, I perform a look-up in the master log file (compiled in Stage 1), using the *comment_IDs* of the removed comment.

Through this look-up, I obtain all the fields that were previously contained in the removed comments (like 'body', 'subreddit', 'author', and so on) before it was removed by moderators.

Using this 3-stage process, I collect 4,605,947 removed comments from Reddit during a 10-month period, from May 2016 to March 2017. All the removed comments I identify constitute my Reddit moderated comments corpus (denoted by $\mathcal{M}$ in the remainder of the chapter).

4.2.2 Preprocessing moderated comments in $\mathcal{M}$

AutoModerator replies.

I observe the presence of comments authored by *AutoModerator* in my removed comments corpus ($\mathcal{M}$). These are comments posted as replies to comments removed by the *AutoModerator*s of subreddits. *AutoModerator* is a customizable moderation bot used by many subreddits to automatically moderate posts from specific users or websites, and flag content that is inappropriate based on a predefined word list [6]. Upon removing a comment or link, *AutoModerator* posts a reply to the moderated comment indicating why it was removed. An example *Automoderator* comment is shown below:

> This submission has been removed. Submissions must be direct links to images in the imgur, minus, or gfycat domains. When using Imgur, simply right-click the image, select "Open in a new tab", and submit that URL. * I am a bot, and this action was performed automatically. Please [contact the moderators of this subreddit] if you have any questions or concerns.*

Given that different subreddits may use *AutoModerator* differently, I take precaution and remove all comments authored by *AutoModerator*, even if they do not appear in the form of replies to moderated comments in my dataset. Since these comments authored

[6] https://www.reddit.com/r/AutoModerator/

by *AutoModerator* are just warnings issued to users following actual removals, I do not consider them in my analysis. As a result, I discard all 101,502 comments which were authored by *AutoModerator* from $\mathcal{M}$.

Discarding replies to moderated comments.

Next, I strip replies to moderated comments in $\mathcal{M}$. Through interactions with various subreddit moderators on a separate project, I learned that comments posted as replies to moderated comments are often also removed by moderators. These are replies that get removed due to their parent comment's removal, and they are sometimes referred to as the "children of the poisoned tree." Since these replies are not always removed intentionally, I decided to err on the side of caution, and discard such replies. I do this by identifying comments whose parents are themselves contained in $\mathcal{M}$. Through this procedure, I discard 1,051,623 moderated comments which I identify to be replies to comments that were removed, giving me the final dataset which I use for further analysis.

Study subreddits.

After preprocessing the data, there are over 3 million moderated comments contained in $\mathcal{M}$, and they are collected from 41,097 unique subreddits. But I was only able to collect very few moderated comments (i.e., less than 10) for most of these subreddits. My current goal is to build machine learning (ML) models that can predict moderator removals for the subreddits they are trained on. In order to build robust ML classifiers, I restrict my analysis only to the subreddits for which I was able to collect a reasonable amount of moderated comments. Therefore, I discard all subreddits that generate fewer than *5,000* moderated comments in $\mathcal{M}$.

Next, I discard all comments from any non-English subreddits present in my corpus. I use *langdetect* [7] and examine all comments from subreddits to decide whether the subreddit

[7] https://pypi.python.org/pypi/langdetect

interactions are predominantly in English or not. For each subreddit, I predict only the top language using *langdetect*, and count the fraction of comments from that subreddit with English as first language. Via this step, I identify and discard 18 non-English subreddits, each of which contains more than 50% of their overall comments in languages other than English (e.g., r/podemos, r/svenskpolitik, r/Suomi, r/argentina, r/brasil, r/italy, r/france, and so on).

Finally, 2,831,664 moderated comments remain in $\mathcal{M}$, all originating in the 100 subreddits generating the most removed comments in my corpus. I call these 100 subreddits my *study subreddits* for the rest of the chapter. At the time of my analysis, the study subreddits had an average 5.76 million subscribers, with r/funny having the highest subscriber count (19 million), and r/PurplePillDebate having the lowest subscriber count (16,000). On average, each subreddit contributes 20,070 moderated comments, with r/The_Donald contributing the most (184,168) and r/jailbreak contributing the least (5,616) number of moderated comments in $\mathcal{M}$.

4.2.3 Unmoderated comments from Reddit

In addition to collecting comments that were moderated from different subreddits, I also collect comments that were not removed by moderators (i.e., unmoderated comments). As shown in Stage 1 of Figure 4.2, I store all comments obtained from r/all through the PRAW API in master log files. These master logs include comments that are both moderated subsequently after posting, and comments that still remained online at the time of data collection. In order to build my corpus of unmoderated comments, I use all comments present in the daily logs, which are not in $\mathcal{M}$. Essentially, any comment that I crawl from Reddit, which is not removed by a moderator within 24 hours from the time of posting is added to my unmoderated comments corpus. These comments are collected similarly to the moderated comments, from the same set of subreddits, and through the same API. Using this data, I compile a dataset of unmoderated comments for all study subreddits.

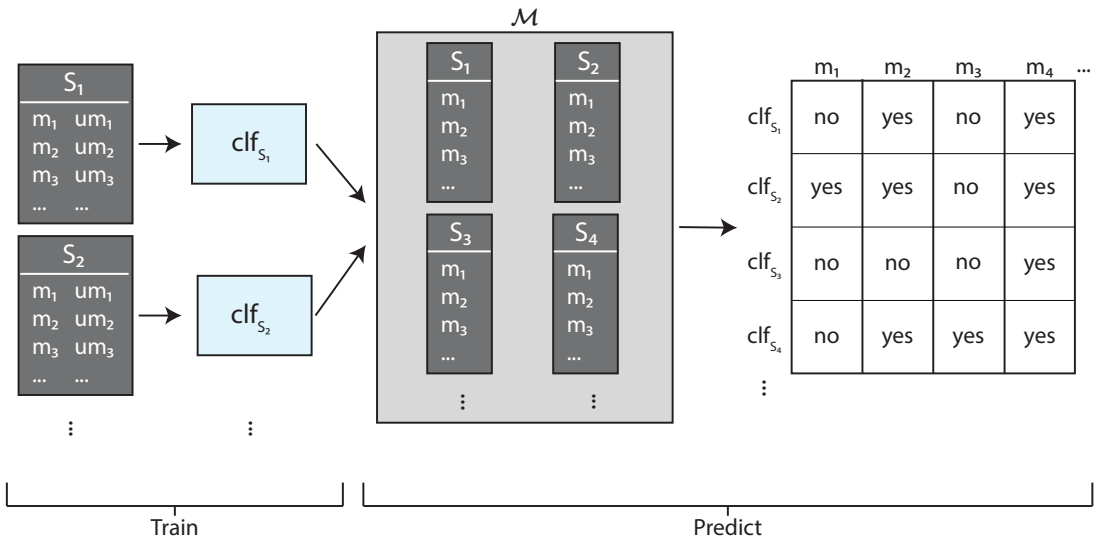

Figure 4.3: In the first step (*Train*), I train classifiers to predict whether a comment posted on a subreddit will get moderated or not. For each study subreddit S_k, I build a classifier clf_{S_k} using moderated (e.g., m_i) and unmoderated (e.g., um_i) comments obtained entirely from S_k. In the next step (*Predict*), I obtain predictions from each subreddit classifier (e.g., clf_{S_k}) for each comment present in $\mathcal{M}$, and generate a *prediction matrix*. Columns in this matrix are comments in $\mathcal{M}$, and rows are subreddit classifiers. Each cell [i,j] in the prediction matrix contains a *yes* or *no*, depending on what classifier clf_{S_i} predicted for comment m_j: *If it were hypothetically posted here, would it get moderated?*

4.3 Method: Classifiers for predicting comment removals

In this section, I detail the procedure used to train classifiers that can predict moderator removals within the study subreddits. Using the comments that were removed by moderators of each study subreddit, along with unmoderated comments collected from study subreddits, I train machine learning models to predict whether a comment posted on the subreddit will get moderated or not. An illustration of this approach is shown in Figure 4.3.

4.3.1 Building classifiers for study subreddits

I will refer to each in-domain classifier built entirely using moderated and unmoderated comments from a single subreddit as a "subreddit classifier". I go on to build 100 such classifiers, one for each of my study subreddits. Each subreddit's classifier is trained on

comments removed by moderators from the subreddit under consideration, along with an equal number of randomly sampled comments that were not moderated (at the time of my data collection).

Next, I describe the construction of my 100 subreddit classifiers, and evaluation of the in-domain classifiers through 10-fold cross-validation tests.

Balancing datasets for each subreddit.

I shuffle and balance each subreddit's dataset to ensure an equal number of comments from each class (moderated and unmoderated). Note that balancing the number of samples from each class likely does not mimic real-world situations. In general, moderated posts are less frequent than unmoderated posts. However, balancing across all conditions ensures that I can easily interpret model fits relative to one another.

`FastText` classifiers.

`FastText` is a state-of-the-art library for text classification [146, 147]. It represents each instance by the average of vector representations for words and n-grams, which are short units of adjacent words. These "representation vectors" enable generalization to words and n-grams that are not encountered in the training data. Supervised training is used to estimate another set of vectors, per label, which characterize the classification rule. If learning is successful, then subreddits with similar moderation patterns will have similar vectors.

Parameter tuning using gridsearch.

Using `FastText`, I build 100 in-domain subreddit classifiers, each trained on an equal number of moderated and unmoderated comments obtained from the study subreddit (i.e., binary classification with balanced classes). Like all classifiers, `FastText` has a number of parameters that must be tuned to achieve optimal performance. I tune these parameters by grid search, trying a large set of values, and selecting those which maximize the F1

Table 4.1: Grid of parameter values used when running classification tests to find the best combination of parameter values for my models. The best values shown for all the parameters, found with a grid search, were used in all classifiers.

Parameter	Description	Range	Best value
lr	Learning rate	[0.05, 0.5]	0.05
epoch	Number of epochs	[25,30,50]	25
dim	Size of word vectors	[100,200]	200
ngram range	Max length of word ngrams	[1,2,3]	3
lowercase	Converting text to lowercase	[on,off]	on
punctuation removal	Remove punctuation in text	[on,off]	on
number removal	Remove numbers in text	[on,off]	on

(f-measure) across 10-fold cross-validation. The parameter space, along with the best performing parameter values are shown in Table 4.1.

Evaluation of in-domain subreddit classifiers.

Using the best performing parameter values shown in Table 4.1, I train in-domain subreddit classifiers for each of the 100 study subreddits. In order to evaluate the subreddit classifiers, I perform 10-fold cross-validation tests using the balanced set of moderated and unmoderated comments collected from each study subreddit. The mean 10-fold cross-validation F1 score for the 100 study subreddits was 71.4%. This is comparable to the performance achieved in prior work on building purely in-domain classifiers to identify moderated comments within an online community [37, 59].

4.3.2 Compute agreement among subreddit classifiers' predictions

I obtain the predictions from each of the 100 subreddit classifiers for all moderated comments present in $\mathcal{M}$. The prediction from each subreddit classifier for a comment represents a probabilistic answer to the following question: *If this comment were posted on this subreddit, would the moderators remove it?*

Next, I compute the overall agreement among all subreddit classifiers' predictions for each comment present in $\mathcal{M}$. By overall agreement among subreddit classifiers for a comment, I refer to the number of classifiers that agree to remove the same comment. The

output of this step is a *prediction matrix*, with number of rows equal to the number of comments in $\mathcal{M}$ (2.8M), and the number of columns equal to the number of subreddits for which I have trained classifiers (100).

4.3.3 Methodological limitations

Access to only textual data.

My current method of data collection does not give me access to removed content in the form of pictures, GIFs or videos. As a result, I am not able to identify community expectations around multimedia content.

False negatives.

Anecdotal evidence shows that not all comments posted on a subreddit have been seen by moderators [148]. This could lead to some *false negatives* (i.e., comments that should have been removed) being present in my collection of unmoderated comments, which could affect the classifiers I build. Future work can investigate this issue, and examine the amount of such comments that the moderators typically fail to see on Reddit.

Passive norms.

Note that the classifiers learn about rules and norms that are actively enforced by moderators on a subreddit. It could be the case that there exist "passive norms" that have not needed to be actively enforced—no one has thought to violate those norms within the subreddit. For example, posting TV show spoilers may actually be considered a norm violation on many subreddits, but the classifiers may not identify that such a comment would be moderated on a specific subreddit if no one has posted such spoilers within that subreddit before. Such passive norms could serve as "blind spots" for the subreddit classifiers, and may be an intriguing avenue for future study.

Table 4.2: Clusters obtained from K-means clustering, based on agreement among classifier predictions to remove comments. The subreddits in each cluster are ordered by cosine distance from their respective cluster's center. *Size* denotes the number of subreddits present in the cluster, *type* denotes the cluster type or type of "norm" that is shared by subreddits present in the cluster, and *name* denotes the assigned cluster number by which I will reference each cluster in further sections.

Cluster subreddits	**Name**	**Size**	**Type**
conspiracy, Android, atheism, Incels, PurplePillDebate, IAmA, canada, tifu, india, SubredditDrama, dataisbeautiful, pics, LifeProTips, hiphopheads, fantasyfootball, explainlikeimfive, worldnews, SandersForPresident	C_0	18	Meso
CanadaPolitics, spacex, changemyview, NeutralPolitics, personalfinance, AskHistorians, history, whatisthisthing, science, Games, philosophy, space, Futurology, syriancivilwar, legaladvice, PoliticalDiscussion, AskTrumpSupporters, TheSilphRoad, Christianity, DIY, OutOfTheLoop, UpliftingNews	C_1	22	Meso
DestinyTheGame, hearthstone, Overwatch, jailbreak, 2007scape, wow	C_2	6	Meso
CFB, me_irl, books, movies, nba, nfl, asoiaf, pokemon, MMA, relationships, AskWomen, food, pcmasterrace, Showerthoughts, GlobalOffensiveTrade, pokemongo, leagueoflegends, depression, gonewild, hillaryclinton, SuicideWatch, The_Donald, gaming, GlobalOffensive, anime, politics, photoshopbattles, television, ShitRedditSays, GetMotivated, aww, EnoughTrumpSpam, sex, gameofthrones, TwoXChromosomes, funny, nottheonion, europe, LateStageCapitalism, news, technology, soccerstreams, socialism	C_3	43	Meso
churning, NSFW_GIF, pokemontrades, nosleep	C_4	4	Meso
videos, OldSchoolCool, gifs	C_5	3	Meso
AskReddit	C_6	1	Micro
BlackPeopleTwitter	C_7	1	Micro
askscience	C_8	1	Micro
creepyPMs	C_9	1	Micro

4.4 Method: Clustering subreddits and extracting norms

Next, I identify subreddits where moderators enforced similar community norms, by finding clusters of subreddits that would moderate the same comments. An illustration of this approach is shown in Figure 4.4. Using the predictions obtained from the 100 subreddit classifiers for all moderated comments in $\mathcal{M}$, described in the previous section, I cluster

the subreddits based on their agreement with respect to moderating comments. For each cluster of subreddits, I then extract the norms enforced by moderators of most of these subreddits. I employ open coding to qualitatively identify norm violations exhibited by comments predicted to be moderated by subreddit classifiers.

Note that an alternative scheme like matrix factorization (or topic modeling, or clustering) on the *comments themselves* would likely just group subreddits by content, rather than by moderation practices (i.e., the decision to moderate comments or not). I believe that this would be true even if I focused exclusively on moderated comments, since a norm-violating comment in, say r/nba, is still likely about basketball. There is one key aspect that the procedure described above would miss: the labeling associated with the moderated posts. The procedure I use in this work, on the other hand, focuses on moderated comments. I begin by identifying commonality in the language of moderated comments via the subreddit classifiers, and then use this commonality to cluster the subreddits, arriving at subreddits clustered by *moderation practices*. Finally, I return to the language of the moderated comments to analyze the topics that characterize the obtained clusters.

4.4.1 K-means clustering

I use the K-means clustering algorithm [149] to cluster subreddits based on their predictions on all removed comments present in $\mathcal{M}$. Here, I find coherent clusters of subreddits that would remove similar comments. Because the matrix of subreddit-comment predictions is large, I first reduce its dimensionality by performing Principal Component Analysis [150]. Intuitively, PCA reduces the size of the input matrix by iteratively computing a projection that explains the most variance in the input. I find that a projection on 81 dimensions is sufficient to explain 90% of the original variance. I then cluster the PCA-transformed predictions of the subreddit classifiers using K-means. I determine the number of clusters k by examining the mean silhouette coefficient [151]—the similarity of predictions within a cluster with respect to other clusters. I test the clustering algorithm with different initial-

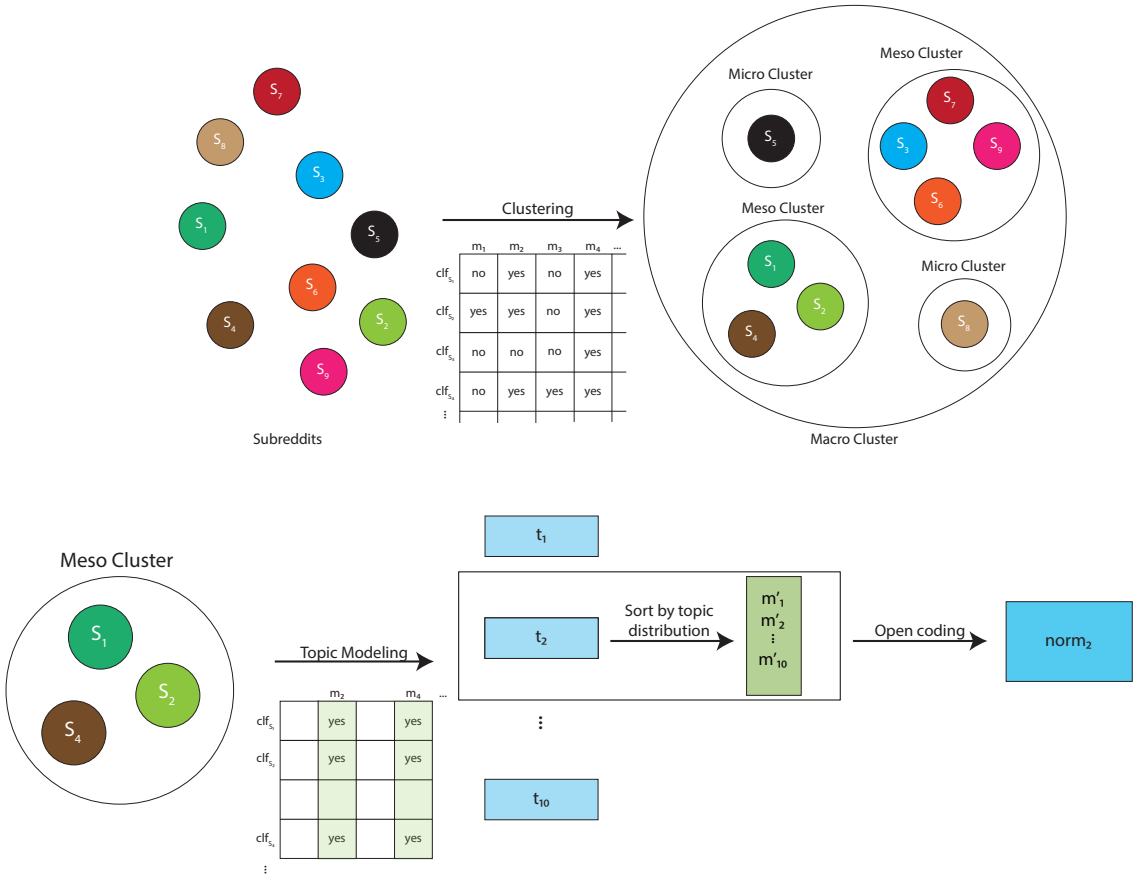

Figure 4.4: Based on agreement among subreddit classifiers (e.g., clf_{S_k}) to remove comments (e.g., m_j), I cluster subreddits (e.g., S_k) into three different types of clusters: *macro*, *meso*, and *micro* clusters. For each cluster of subreddits, I perform topic modeling only on comments in $\mathcal{M}$ that the subreddit classifiers agreed to moderate, using the prediction matrix shown in Figure 4.3. Finally, I employ open coding to extract the norms violated by 10 comments that rank highly in the topics I identify. By repeating this procedure for the macro cluster containing all subreddits, and each cluster shown in Table 4.2, I extract *macro*, *meso*, and *micro* norms.

izations of K (ranging from 1 to 20), in order to identify the most stable configuration.

4.4.2 Clustering results

By increasing K from 2 to 20, I find that after an initial local maximum the coefficient peaks around $K = 10$, before degrading for higher values. Therefore, I cluster the predictions in $K = 10$ groups. The resulting 10 clusters are shown in Table 4.2, and the 2-D t-SNE [152]

representation of the clusters is shown in Figure 4.5.

Based on the amount of agreement among subreddit classifiers, I identify three different levels of clusters among the study subreddits.

Macro cluster

First, I consider all 100 study subreddits to be part of *one large cluster*, so that I can identify comments that a large majority of subreddit classifiers belonging to this cluster agree to remove. These comments are highly likely to be removed by moderators of all study subreddits, when posted on their subreddit. Using the text contained in these comments, I will extract norms that extend across most study subreddits. I call these *macro norms*, as I observe them to be enforced by moderators of a large majority of my study subreddits.

Meso clusters

I identify six meso-clusters of subreddits (C_0 to C_5), obtained through K-means clustering, shown in Table 4.2. Moderators from all subreddits belonging to a cluster tend to agree on what comments to remove from their subreddits (based on the predictions obtained from the subreddit classifiers). For each *meso cluster*, I identify comments that a large majority of subreddit classifiers belonging to this cluster agreed to remove, while subreddits that do not belong to the cluster agreed to not remove. For each comment, I compute the following ratio: fraction of subreddits within the cluster that agree to remove the comment (based on classifier predictions), normalized by the fraction of subreddits outside the cluster that agreed to remove the comment. Then, I rank all comments based on this computed ratio, and then pick only the top 1% out of all comments. These comments are highly likely to be removed only by moderators of subreddits present in the same cluster. Using the text in these comments, I will go on to qualitatively extract cluster-specific norms that extend across most study subreddits in the same meso cluster. I will call these *meso norms*, as they are likely to be enforced by moderators of communities (subreddits) in the same meso

cluster, but not on other parts of Reddit.

Micro clusters

Finally, I have the four micro-clusters (C_6 to C_9) obtained in Table 4.2, each containing a single, isolate study subreddit, in order to identify comments that are only removed by moderators of these individual subreddits. I identify comments that only the individual subreddit belonging to each micro cluster agreed to remove, while all other subreddits agreed to not remove. For each comment, I compute a similar ratio: fraction of subreddits within the cluster that agree to remove the comment (either 0 or 1 since there exists only one subreddit in the cluster), normalized by the fraction of subreddits outside the cluster that agreed to remove the comment. I rank all comments based on this computed ratio, and then pick only the top 1% comments. These are comments that violate highly specific norms that are enforced by moderators of micro cluster subreddits, while the same comments are not removed when posted on most other study subreddits. Using the text in these comments, I will qualitatively extract norms that are highly specific to each individual study subreddit. I call these *micro norms*, as I observe them to be enforced exclusively by moderators of individual subreddits.

Note: Subreddits are clustered based on the comments that their classifiers agree to moderate, and these are not necessarily representative of typical comments found on these subreddits. As a result, some of the obtained clusters may not be intuitive, and subreddits present in the same cluster need not appear to be topically similar. Instead, what I observe in these obtained clusters are subreddits that share similar moderation policies and norms. As mentioned in 5.1, the obtained clusters were determined to be the most stable configuration by examining the mean silhouette coefficient [151].

4.4.3 Norm extraction through topic modeling and open coding

As explained in the previous subsection, I identify clusters of subreddits that share norms among themselves at three different levels (macro, meso and micro).

Topic modeling

I next adopt a computational approach to reduce the dimensionality of my textual data. I employ topic modeling on the comments agreed to be moderated by subreddit classifiers belonging to each cluster to identify the underlying topics contained in these comments. I frame the task as follows:

Applying Latent Dirichlet Allocation (LDA) [153], I estimate topic distributions on the comments that have high agreement among classifiers belonging to the same cluster. I use LDA to estimate the topic distributions among 10 topics for each cluster. Every comment belonging to each cluster I analyzed is considered to be a document for this analysis. In further analysis, I tested by increasing the number of topics from 10 to 20 for LDA, but observe that no new types of norms emerged. As a result, I estimate topic distributions among 10 topics for each subreddit cluster.

Open coding for mapping topics to norm violations

Finally, I introduce a qualitative step, where I use open coding to manually code each topic by the norm violation it represents (in the form of a 1-2 line explanation behind the comment's removal). Using the topic distribution computed for all comments agreed to be removed by subreddits within each cluster, I identify 10 comments that ranked highly in each of the 10 topics obtained for the cluster. Then, three annotators independently code each topic by the norm violation it represents, using the 10 comments ranking highly in that topic as context. This way, I manually map all 10 topics (using 10 randomly sampled comments for context) to their respective norms for each cluster. Then, the three annotators come together to compare the norms they coded independently, and resolve any disagree-

ments. By repeating this process for all clusters at the three different levels, I extract the macro, meso and micro norms contained in $\mathcal{M}$.

Through open coding, a total of 100 different topics were coded manually, and I observed the presence of 32 topics for which the annotators could not identify the exact norms being violated. This could arise from a number of different factors: computational noise introduced by the classifiers in the data; lack of background knowledge about the actual subreddits as outsiders; and, missing the context information for comments that were moderated. For example, some of these comments could have been removed by moderators due to reasons that are very highly context-specific to the type of discussions they were a part of. I discarded such topics for which I could not identify the exact norms being violated, and only present the norms that were identified and agreed upon by all three annotators.

4.4.4 Methodological limitations

My findings hinge on the algorithms I use in my methods—classifiers I train and the clustering algorithm I choose to employ can play a role in the types of norms I uncover. On the other hand, using these algorithms give me the ability to study site-wide norms holistically in a large-scale empirical manner, which is not possible to do by manual inspection alone.

Lack of context for removed comments.

In my current analysis, I do not have access to the conversations surrounding the comments that were removed by the moderators of different subreddits. This lack of context for some of the removed comments could make interpreting the reasons behind moderator actions a hard task. Future work examining removed comments within the context of the larger discussions they are a part of could help understand moderator actions at a discourse-level. While it is true that I do not have context information for all moderated comments, the three independent annotators were able to agree on the norm violations represented by 68 out of the 100 topics that were coded.

Confounding factors.

Note that I do not know the exact reason behind each moderator removal, and I do not account for differing levels of moderator activity within different subreddits. Currently, I do account for one common type of mass-removal: "children of the poisoned tree". Moderators sometimes remove all the children that were posted in response to comment that needs to be removed for violating community norms. The rationale behind this being, given that the parent needs to be removed, it is "safer" to remove its children, since there is high possibility of users responding in undesirable ways to an undesirable comment.

Treating auto-moderated and human-moderated comments equally.

My analysis treats auto-moderated and human-moderated comments equally when constructing norms for communities. I are currently unable to systematically determine whether comments were removed by AutoModerator or human moderators. In fact, the ways in which AutoModerator is used for moderation varies from subreddit to subreddit. Anecdotally, I know that some subreddits remove all comments that trigger any of the hard-coded rules defined for the AutoModerator, while other subreddits use AutoModerator for triaging comments, which are subsequently reviewed by human moderators. Future work can examine the biases introduced by automated tools, like AutoModerator, that moderate based on hard-coded rules, when constructing community norms.

Temporal aspects of community norms.

It is important to note that norms can change within and across communities over time, and tools that moderate automatically based on the "right" set of norms for a community must be flexible to change over time. my current analysis presents a static snapshot of norms identified through moderator actions, and does not examine the temporal aspect of community norms. Future work may find traction exploring the temporal nature of community norms, examining how norms evolve within communities over time.

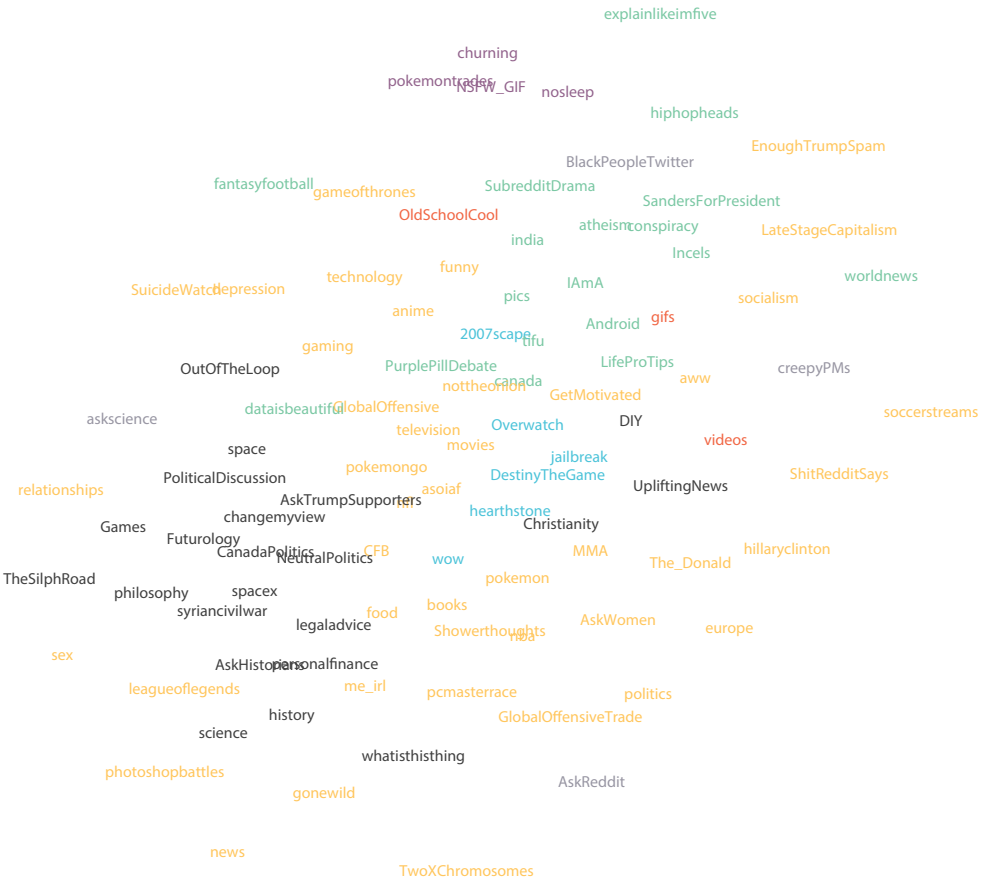

Figure 4.5: 2-D t-SNE representation of the clusters, obtained from the high-dimensional space of subreddit classifier predictions. Intuitively, subreddits that are spatially nearby have similar moderation practices, according to the classifiers. Clusters are indicated by color, with all singleton subreddits shown in gray.

4.5 Results

I identified 8 macro, 21 meso, and 15 micro norms by employing open-coding on comments agreed to be removed by subreddits from different clusters of subreddits. They are shown in Table 4.3, Table 4.4, and Table 4.5 respectively.

Table 4.3: Macro-norms extracted by analyzing comments that at least 96 out of 100 subreddit classifiers predicted to moderate from their respective subreddits. For each norm, I include an example comment found to be violating it.

Norm violations	**Example comments**
Using misogynistic slurs	*what a dumb cunt lol what a pussy*
Opposing political views around Donald Trump (depends on originating subreddit)	*stay classy trump supporters you bunch of worthless fucking pricks*
Hate speech that is racist or homophobic	*you're allow to swear on the internet you fucking [n-word]*
Verbal attacks on Reddit or specific subreddits	*drain the swamp, u/spez is a kek kekadooddledoo fuck reddit this site sucks*
Posting pornographic links	*you dont like senpai [URL]*
Personal attacks	*please kill yourself you useless sack of shit*
Abusing and criticizing moderators	*lets see if this gets deleted. fuck you r/news mods*
Claiming the other person is too sensitive	*fucking cry about it you fucking baby*

Table 4.4: Meso-norms extracted for Clusters C_0 to C_5 by analyzing comments agreed to be moderated by most subreddits within each cluster. For each norm, I include example comments and also the names of the clusters that enforce it.

Norm violations	**Example comments**	**Clusters**
Meme responses	*mitochondria is the powerhouse of the cell*	C_0
Comments that only express thanks	*thank you so much for sharing this*	C_0, C_5

Table 4.4 continued

Ad hominem attacks that demean and undermine users, based on flairs or usernames	*just looking at your rank flair I wouldn't really criticize*	C_0
Mocking the concept of safe space	*poor snowflake do you need a safe space*	C_0
Attempts to be funny, sarcastic, or make jokes	*its obvious god is really keen on what one eats and dinner etiquette*	C_1
Personal reactions, opinions	*and this is why i love science, always on the pursuit of knowledge*	C_1, C_4
Phatic talk	*they're making a new austin powers movie*	C_1
Outbound links to illegal live streams	*free live streaming chicago bulls los angeles lakers basketball*	C_2
Personal anecdotes (details about one's family, past events)	*According to my parents, my dad wanted to name me 'Taylor made' and my mom was like there is no way in hell you are doing that to my child. So Taylor with a normal first name was the compromise*	C_2, C_4
References to trading items	*i have a competitive shiny gengar which i could trade for that lugia if you're interested*	C_2
Expressing disagreement and criticisizing opinions shared by others	*i totally see where your coming from. i don't think however that I should approach this with a gradient...*	C_2, C_3
Talking about romantic relationships and sex	*just ask him out and see how it goes*	C_2, C_3
Mansplaining	*I'm not saying it was her goal. I'm saying her actions were akin to someone who had that goal*	C_2

Table 4.4 continued

Talking about guns	*laws vary by jurisdiction in a lot of places pointing the gun is automatically a threat but not pointing the gun is not automatically not a threat*	C_3
Excessive hedging	*maybe that is how you interpreted it but that is not necessarily what they meant*	C_3
Using Wikipedia articles and other web links to support arguments	*it was an acquired accent taught in schools source [Link to Wikipedia]*	C_3, C_4
Generalized complaining (e.g., electoral system, censorship, airport rates, etc.)	*yeesh airport rates are always silly but it s still disheartening to see a rate like that in any context*	C_3, C_4
Acknowledging a good point	*that's a valid point. honestly i had not thought about it that way before*	C_4
Links to promotional spam	*would you rate kenya coffee [Link to blog] ... as the best in the world or at least amongst the best*	C_5
Mocking religion and nationality	*but but but but but but but but but islam is a religion of peace*	C_5
Hostility towards Muslims, and immigrants	*the country has to import rubbish from other countries. is that what I are calling it now*	C_5

Table 4.5: Micro-norms extracted for Clusters C_6 to C_9 by analyzing comments predicted to be moderated only by the individual subreddit within each cluster. For each norm, I include example comments and also the names of the clusters that enforce it.

Norm violations	**Example comments**	**Clusters**
Comments that only express thanks	(see above)	C_6, C_8
References to movies and TV shows	*on the other hand what other episode could it really be*	C_6
Offering commerce tips	*i could offer 25k so i think around somewhere there*	C_6, C_8
References to history	*perhaps the initials are emperor wilhelm, as in wilhelm ii who reinstitute it in 1914. that would also explain the crown*	C_6
Using Wikipedia links as source	(see above)	C_6
Personal reactions	(see above)	C_7
Guessing at other people's motives	*even by the fn their private views are likely different from their public views*	C_7
Talking about past regrets and lost opportunities	*wow i smoked pot for the first time at 13 and also dropped out of high school*	C_7
Merely indicating agreement conversation	*definitely i agree*	C_7
Personal anecdotes	(see above)	C_8

Table 4.5 continued

Diet advice, and pro-anorexic content	*i'm 153 lbs 5 9 and if i don't eat much a couple days in a row i can lose up to 5 or 6 lbs. once i've gone more than 3 or 4 days without much food. i completely lose my appetite and have to force feed. 1lb a day doesn't seem like much, i've lost up to 20 lbs in 3 weeks and that was when i was just eating when ever i was hungry*	C_8
High-school science theories	*when i was in highschool, i misunderstood the myth even more thinking that overnight a car battery would turn into some sort of acidic goo pile. i left a car battery on the front walk of my high school principal's house one night as a prank, again thinking he would come out the next morning to a pile of acidic slime. i wonder how confused he was to find a perfectly normal car battery in his yard the next morning*	C_8
Undermining and arguing against author opinions	*how is this disrespectful or hateful? would you remove my comment if it was criticizing the prevalence of homophobia related to christianity i don t think that's fair in the slightest*	C_9
Calling out previous authors for flaws	*i call bullshit on it*	C_9
Showing lack of confidence in one's own position	*i didn't say they were the same, i said a certain unnamed insult fits both*	C_9

4.5.1 Macro norms on Reddit

Working with moderated comments from 100 different communities on Reddit, I identified 8 macro norms that are enforced by the moderators on most subreddits.

Hate speech in the form of homophobic and racist slurs are considered as norm violations on most parts of Reddit. In addition, name-calling, use of misogynistic slurs, graphic verbal attacks, and distributing pornographic material are not condoned. Comments presenting opposing political views around Trump, either for or against depending on originating subreddit, are also removed by moderators. Such content could potentially lead to highly polarized comment threads, thereby hijacking ongoing discourse towards unrelated topics. This indicates that such comments are considered to be norm violations on Reddit because they hurt the process of discussion, and not necessarily because they are universally abhorrent. Another common norm violation is criticizing and abusing subreddit moderators, and most of the time, these are members of the community expressing their discontent with moderator actions (e.g., removing or promoting certain posts, lack of a formal escalation system, and the need for transparency in moderation). Sometimes, this discontent goes beyond certain specific subreddits, and users verbally attack Reddit (and its admins) due to a variety of reasons (e.g., policy change, banning communities, public statements, and so on). In such cases, moderators of most subreddits intervene and remove such comments.

4.5.2 Meso norms on Reddit

Cluster C_0

There are 18 subreddits present in this cluster, and they are on a range of topics (news, countries, politics, lifestyle, and so on). These communities have norms against ad hominem attacks, especially demeaning and undermining user opinion based on flairs or usernames. Moderators of these subreddits also remove comments mocking the concept of a safe space,

and purely *meme* responses [154] (e.g., *"mitochondria is the powerhouse of the cell"*). Interestingly, comments that only express *thanks* are often observed to be removed by moderators. Though these comments serve a purpose for the two individuals that are part of the conversation, they do not necessarily add value, to other participants, within the context of the overall discussion. This type of removal could also be for archival reasons, where you want to minimize the amount of noise in current snapshots of the subreddit being archived for future references.

Cluster C_1

There are 22 subreddits present in this cluster, and most of them are subreddits that are known to be heavily moderated (e.g. r/NeutralPolitics, r/science, r/AskHistorians). Personal reactions, opinions, and (failed) attempts to make jokes or be sarcastic are considered to be violations of community norms on these subreddits. Additionally, references to movies, phatic talk, and comments that generally do not add value to ongoing conversation are removed by moderators.

Cluster C_2

There are 6 subreddits present in this cluster, and most of them are gaming-related subreddits (e.g., r/DestinyTheGame, r/Overwatch, r/wow). Moderators remove outbound links to (illegal) live streams and references to trading items (especially Pokemon). Other common removals include comments sharing personal anecdotes, stories about romantic relationships and sex. Mansplaining and criticizing opinions shared by other users are not condoned by these subreddits.

Cluster C_3

There are 43 subreddits present in this cluster, including highly popular subreddits focused on topics like politics (e.g., r/The_Donald, r/hillaryclinton), sports (e.g., r/NBA, r/nfl), and

mental health (e.g., r/depression, r/SuicideWatch). Hedging language, criticizing other users' opinions, and the use of weblinks, including Wikipedia articles, to support arguments are not encouraged within these subreddits. Moderators also remove comments complaining about current state of things (e.g., electoral system, censorship, and so on).

Cluster C_4

There are 4 subreddits present in this cluster, and they are r/churning, r/NSFW_GIF, r/pokemontrades, and r/nosleep. Norm violations include complaining about the state of things, and using Wikipedia articles to make a point. Moderators also remove comments that are personal reactions, personal anecdotes, or just acknowledging a good point.

Cluster C_5

There are 3 subreddits present in this cluster, and they are r/videos, r/OldSchoolCool, and r/gifs. Hostility towards muslims and immigrants, and mocking religion and nationality violate the norms of these communities. Moderators also remove links containing promotional spam, and low value comments that only express *thanks*.

4.5.3 Micro norms on Reddit

Micro norms are context-dependent, and highly specific to individual subreddits, and are not found to be widely enforced on most parts of Reddit. For instance, moderators of r/AskReddit, a Q&A forum, consider low value comments that express gratitude, contain movie or TV show references, and offering commerce tips as norm violations. In addition, references to historical events, and comments using Wikipedia links to support their arguments are removed by moderators, despite there being no written rules against them. r/BlackPeopleTwitter is intended for hilarious and insightful social media posts by black people, with an emphasis on hilarity.[8] As a result, posting personal reactions to issues,

[8] `https://www.reddit.com/r/BlackPeopleTwitter/`

guessing at the motives of users, and talking about past regrets or missed opportunities are considered norm violations by the moderators. On a science Q&A forum promoting scientific literacy like r/askScience, posting personal anecdotes is against comment rules, as specified by subreddit moderators. I also observed that mods do not tolerate high-school science theories, and diet advice (especially pro-anorexic content) in discussions. On a support subreddit like r/CreepyPMs, undermining, arguing against, and calling out flaws present in comments/post by previous authors are considered norm violations. Sometimes, comments showing a lack of confidence in one's own position are also removed by mods.

4.6 Discussion

My findings describe the ecosystem of norms on Reddit. Some of the community norms I identified are mirrored in the written rules and guidelines provided by Reddit, or individual subreddits (an encouraging face validity sign); however, many are not. I also see many *unpublished* norms that are widely enforced by subreddit moderators.

4.6.1 Norms at different scales on Reddit

My findings document the existence of norm violations that are universally removed by moderators of most subreddits. These include comments that contain personal attacks, misogyny, and hate speech in the form of racism and homophobia. The presence of these macro norms are in many ways encouraging, as they indicate that engaging in such behavior is considered a norm violation site-wide. I would argue that knowing about the presence of such site-wide norms could also help moderators of new and emerging communities shape their regulation policies during the community's formative stages, and feel more confident doing so.

I also documented norms that are local to specific groups of subreddits—the meso norms. For instance, sharing personal anecdotes, and posting links containing promotional spam are considered norm violations in certain clusters of subreddits (C_2, C_4, and C_5),

while most other communities on Reddit do not consider such comments norm violations. I also found some meso norms that are seemingly counter-intuitive. For instance, comments expressing thanks, or acknowledging a good point are considered to be norm violations in clusters C_0, C_5, and C_4 respectively. Though these comments appear to be polite, and add value to one-to-one conversations between individual users, they may be perceived as *noise* or low-value comments by users trying to follow the larger discussion. On the other hand, I observe that only certain clusters considered mansplaining (C_2), mocking religion and nationality (C_5), and hostility towards Muslims and immigrants (C_5) as norm violations. Despite being important societal issues, they do not appear to be norm violations on most parts of Reddit.

Finally, I observed the presence of highly specific micro norms that apply to individual subreddits. These are distinctive to the particular subreddits they emerge from, and are not widely enforced on most other parts of the site. For example, using Wikipedia as a source and presenting high-school science theories are considered to be norm violations within r/AskReddit, and r/askscience respectively, while most other parts of Reddit would not remove such comments. These idiosyncratic micro norms are important for understanding the reasoning behind moderator removals within individual communities—as well as understanding the range of norms on an umbrella site like Reddit.

4.6.2 Ethical considerations

I recognize that the use of "deleted data" (here in the form of moderated comments) is controversial territory in social computing research. I debated and discussed these issues with my local colleagues, remote colleagues, and my IRB before performing this research. In the end, I arrived at the conclusion that examining moderated comments provides invaluable insights about the governance of online communities, and as long as any downside risks are mitigated, those benefits outweighed the risks. For example, as I discuss next, I believe these findings may enable new mixed-initiative governance tools for online communities.

I actively worked to minimize potential risks by not linking moderated comments back to their authors (who may not want to be immortalized in a research paper next to their norm violation). Moreover, I did not use *posts deleted by their authors* in this work, as those felt qualitatively different to everyone with whom I discussed this work. Finally, in an effort to protect Reddit itself from harm, I used only public data collected via Reddit's official API.

4.6.3 Theoretical implications

Norms play a key role in the governance of online communities [38]. Norms can be nested, in that they can be adopted from the general social context (e.g., use of pejorative adjectives are rude), or from Reddiquette[9], and more general internet comment etiquette (e.g., using all caps is equivalent to shouting at someone). Yet, norms for what is considered to be acceptable can vary significantly from one community to another, thereby making them challenging to study at scale. Through my work, I presented an empirical description of an ecosystem of community norms on Reddit, and my findings shed light on *what Reddit values*, and how widely-held these values are. I believe this is the *first large-scale study* of norms across disparate online communities.

Despite having established moderation strategies, including rules and guidelines, in place to regulate subreddits, bad behaviors continue to remain a challenge for online communities [2, 17, 42]. In the context of Reddit, rules and norms are interrelated. Moderators create formalized rules and guidelines for the front-stage of their subreddits, based on the norms they enforce in the back-stage. In my work, I identified norms as the emergent themes contained in the record of moderated comments. I observed that some of the norms I identified may overlap with outward-facing subreddit rules, but a far greater proportion of them do not. Future work could examine this apparent divide between the formal rules and informal norms enforced by moderators in online communities in greater detail. An understanding of the ecosystem of norms within online communities known to be success-

[9] `https://www.reddit.com/wiki/reddiquette`

ful in regulating behaviors could provide an empirical understanding of the driving factors behind effective online governance.

4.6.4 Implications for online communities

For established online communities, an understanding of the macro, meso and micro norms on Reddit could help moderators reflect on the norms they typically enforce within their subreddits. Moderators can adopt existing norms from other communities known to be successful in regulating behaviors (e.g. r/AskHistorians, r/askscience, and r/NeutralPolitics). This could also help train new moderators by surfacing the implicit norms in the community. For new communities, I believe these macro norms (and some meso norms, depending on the community) may serve as sensible defaults for regulating behavior.

4.6.5 Classifiers that learn from other communities' norms

Finally, the discovery of widely overlapping norms suggests that new automated tools for moderation could find traction in borrowing data from communities which share similar values—a direction I plan to explore in Chapter 6. I observed that the F1 scores obtained for relatively smaller subreddits (with less than 5000 removed comments) was approximately 68%, despite using state-of-the-art classifiers. This indicates the potential for using cross-community data to augment and improve completely in-domain classifiers. By understanding the types of norms that are valued by the target community, researchers could use classifiers trained on other source communities that share similar community norms.

One way to operationalize this idea is take into account the general agreement between the source and target community classifiers, measured by the number of comments both classifiers agree to moderate. In Chapter 6, I introduce a new cross-community classification framework for automated moderation that employs inter-subreddit agreement measures to scaffold other communities' norms.

4.7 Conclusion

I examined community norms on Reddit in a large scale, empirical manner. Using computational and qualitative methods, I examined over 2 million comments removed over 10 months by moderators of 100 top subreddits. I identified three types of norms within Reddit: *macro* norms which are universal to most parts of Reddit; *meso* norms which are shared across certain groups of subreddits; and *micro* norms which are highly specific to individual subreddits. I argued that my findings represent the first large-scale study of norms across disparate online communities, given the size and diversity of Reddit's user base. I concluded by reflecting on the apparent sharing of norms among distinct online communities, discussing implications for theory, and the design of internet communities more broadly.

CHAPTER 5

FORMATIVE INTERVIEW STUDY TO UNDERSTAND MODERATOR NEEDS

Like many social platforms, Reddit relies on human moderation to regulate content. Moderators on Reddit are groups of users with special privileges who regulate content generated within subreddits on a voluntary basis—they are not paid for their labor. Moderators enforce rules that are subreddit-specific [89], in addition to site-wide (content[1] and anti-harassment[2]) policies. Reddit has existing architecture in place to support human moderation on the platform, as well as automated moderation tools to assist moderators. However recent reports have found that Reddit is systematically failing to limit the damage caused by bad actors, and moderators are struggling due to the large volumes of abuse constantly directed at them—resulting in moderator burnout, and even mental health risks [88]. In this chapter, I begin with a formative interview study to learn about the current state of automated moderation tools on Reddit, and explore opportunities for extending these tools[3]. The goal of this study is to understand how moderators use automated tools on different parts of Reddit (either to *proactively* remove content, or to triage content for human review). Through these interviews, I examine challenges faced by existing automated tools and identify unmet moderator needs.

5.1 Methodology

5.1.1 Recruitment

I conducted in-depth, formative interviews with 11 individuals who moderate Reddit communities on a regular basis. Interviewees were recruited by sending private messages

[1] `https://www.redditinc.com/policies/content-policy`
[2] `https://redditblog.com/2015/05/14/promote-ideas-protect-people/`
[3] This study was approved by the IRB at the authors' institution.

Table 5.1: The list of moderators I interviewed. P0 preferred to be identified by his real name in this study.

Subreddit Name	Participant Name	Age	Country
r/AutoModerator	Chad Birch (P0)	35	Canada
r/photoshopbattles, r/blackpeopletwitter	P1	18	USA
r/science	P2	44	USA
r/news, r/funny, r/todayilearned	P3	60	UK
r/science	P4	30	USA
r/relationships	P5	31	USA
r/Sakartvelo	P6	39	UK
r/femalefashionadvice	P7	29	Australia
r/homeimprovement, r/homeautomation	P8	36	USA
r/itsaunixsystem, r/jailbreak	P9	23	Canada
r/computers	P10	39	UK

through Reddit.[4] In order to maintain diversity in the topics and sizes of the communities represented in my interviews, I targeted five mainstream subreddits with over 1M subscribers each, five mid-level subreddits with 100k to 1M subscribers each, and five niche subreddits with less than 10k subscribers. Through this procedure, I reached out to moderator groups from 15 subreddits, and eventually recruited 10 active moderators who regulate content across 11 unique subreddits. In addition, I also interviewed Chad Birch, a Reddit moderator who created AutoModerator (or Automod)[5]—an automated moderation tool that is widely used on Reddit. Overall, I interviewed 11 participants based on a variety of English-speaking countries. Further details about my interview participants are shown in Table 5.1. All the interviewees were compensated with a $20 Amazon gift card for their time; further participation in the study was completely voluntary.

5.1.2 Interview goals

I began the interviews by asking moderators about their general experiences regulating content on Reddit, and then dove into specific moderation practices currently employed

[4]Reddit allowed me to contact each subreddits' moderators through a group *private message*, following which interested moderators responded to my recruitment message individually.

[5]`https://www.reddit.com/wiki/automoderator/full-documentation`

within their subreddits. Next, I explored the breadth of existing automated moderation tools for Reddit, learning about commonalities and differences between how subreddits employ these tools. Finally, the moderators explained the advantages and drawbacks of existing automated tools. Through this process, I was able to understand the needs of moderators and the challenges faced by existing tools. In Chad's interview, I also asked about his experiences creating Automod for a handful of subreddits, and it subsequently getting adopted by Reddit as an internal moderation tool. All 11 interviews were conducted remotely over Skype or Hangouts, and lasted between 40-60 minutes. Interviews were recorded and then transcribed using a paid transcription service.[6] Finally, my co-author and I read through the transcripts, coded them using thematic analysis.

5.2 Current state of automated moderation tools on Reddit

Reddit has existing infrastructure built for supporting moderators in the process of manually curating content within subreddits. Different subreddits use Reddit's infrastructure differently, but the underlying moderation interface and internal tools remain the same across subreddits.

5.2.1 Existing moderation interface

Reddit's current moderation interface is shown in Figure 5.1. Each subreddit has a dedicated moderation queue or "mod queue," which is a central listing of all the content generated within the community that needs to be reviewed by moderators. This includes all of the posts and comments reported by users, and content marked as spam by Reddit's site-wide spam filter. In addition to the mod queue, moderators also use two other tabs to review specific types of content. First is the "Reports" tab, which only shows content flagged through user-reporting, and the second is the "Spam" tab, which only lists removed content (mods said that these might be used by Reddit for refining the spam filter). In addition

[6] www.rev.com

to Reddit's moderation interface, 9 out of the 11 moderators also use third-party browser extensions like Toolbox and Reddit Enhancement Suite (RES). P4 said, "*RES and toolbox are the standard ones (i.e., third party tools) that make the site usable and (provide) extra analytics.*" These browser extensions offer better user-interface and additional features that help mods sift through content manually.

Given the sheer amount of content generated due to Reddit's popularity, reviewing all content manually is infeasible. P2 said, "*Ultimately, for smaller websites you can do human based moderation. But right now, r/science has like 20 million subscribers.*" In order to keep up with the large volumes of user-generated content within subreddits, 10 out of the 11 moderators I interviewed rely heavily on automated tools to triage comments for manual review.

5.2.2 Moderation bots

Moderation bots are a popular class of automated tools that currently support Reddit moderators [155]. Bot development is facilitated through the openly-available Reddit API, and the associated Python Reddit API Wrapper (PRAW), both of which offer a range of scripted functionality [156]. A well-known bot that performs moderation tasks is */u/Botwatchman*, which detects and removes other *blacklisted* bots (based on a predefined list of bots that are not allowed on the subreddit). Additionally, moderators also employ Reddit bots specifically written to perform certain tasks, based on a set of predefined conditions. P1 said, "*r/photoshopbattles only allows submission of images with reasonable quality, and employs a bot to automate checking image quality.*"

5.2.3 AutoModerator

From my interviews, I found that AutoModerator or *Automod* is the most commonly used automated tool on Reddit, and some subreddits rely entirely on Automod to regulate content. P9 said, "*AutoModerator does bulk of the moderation for our subreddit.*" Automod

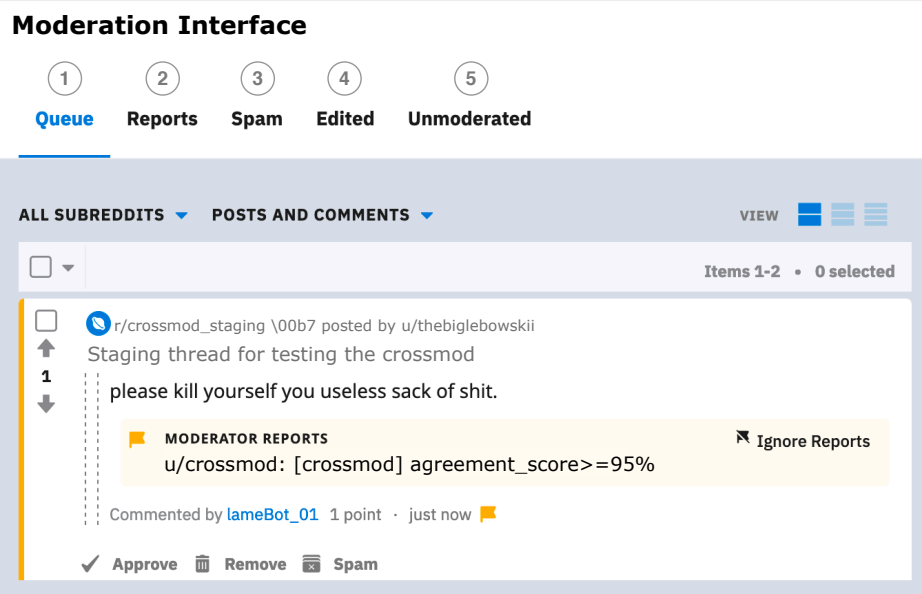

Figure 5.1: Reddit's moderation interface for moderators to curate content manually. Five different tabs exist in this interface: *Queue* (or mod queue), *Reports*, *Spam*, *Edited*, and *Unmoderated*. When automated moderation tools are configured to "triage" content violating community norms, posts and comments are sent to the Mod queue (labeled as tab #1 in the Moderation Interface) for manual review by moderators. All posts and comments in the Moderation *Queue* are reviewed by moderators, after which they make moderation decisions. If a comment does not violate community norms, they can "approve" allowing it to remain on the subreddit. If they feel that a comment violates community norms, then they can either "remove" the comment, or mark it as "spam", taking the content down (i.e., off-site).

is a customizable moderation tool that was created by Chad Birch (P0) for a handful of subreddits in its initial stages. Automod was subsequently adopted by Reddit, and released as an internal tool to assist moderators on the platform. All of the moderators I interviewed had used Automod for moderating their subreddits at some point, and 9 moderators continue to actively use Automod. As a result, I focused my interviews on understanding how Automod is employed within subreddits, and its strengths and weaknesses in the opinions of moderators.

5.3 Current uses of Automod

Automod employs a regular expression (*regex*)-based filtering approach to scan, and proactively remove or triage[7] content within a subreddit. Automod rules are written and maintained by moderators (within a subreddit-specific config file). P9 mentioned that the documentation for Automod is well-maintained and easy to follow. A full description[8] of all the capabilities of Automod can be found on the subreddit called *r/AutoModerator*. 9 out of 11 moderators used Automod predominantly as a *triaging tool*—sending content violating hard-coded rules to the moderation queue for human review. P9 said, "*Automod can* filter *comments: remove from public thread, and push to moderation queue for human review. Apparently, this functionality isn't available to any-one/thing other than Automod.*"

5.3.1 Automod can detect violations based on simple, hard-coded rules

Automod works based on simple, hard-coded rules, and is effective at taking care of repetitious and mundane tasks. P0 said, "*What Automod did was get rid of all the really tedious, easy moderation work and allow the moderators to pay attention to more subjective things.*" For example, P1 uses Automod to enforce formatting guidelines within their subreddit:

> r/photoshopbattles has a submission rule that states, "All titles must begin with *PsBattle*:". This rule helps distinguish r/photoshopbattles images from other pictures when they appear on the Reddit front page along with images from other subreddits. To ensure that users comply with this rule, we have hard-coded an Automod rule that detects and removes all submissions that do not begin with "PsBattle:"

In addition to enforcing formatting restrictions, Automod is also used to ban problematic users, prohibit URLs linking to objectionable websites (P6, P9, P10), and detect comments

[7] Triaging comments and posts for further review by moderators.

[8] `https://www.reddit.com/wiki/automoderator/full-documentation`

using phrases that are known to commonly occur in undesirable content within their subreddit (P1, P2, P4).

Chad, creator of Automod, had created a library of common rules available to everyone for reference. Shown below is an example rule that is used to configure AutoModerator to remove comments that consist of only capital letters:[9]

> **type:** comment
>
> **body (case-sensitive, regex, full-text):** "([A-Z0-9]—""W)+"
>
> **action:** remove
>
> **action_reason:** CAPS ONLY

Many subreddits also share AutoModerator config files among themselves, allowing them to re-use commonly employed rules. P3 said, "*I just copy paste them from a preexisting AutoModerator in another subreddit? You know, the knowledge, you don't need to know how to create it from scratch.*"

5.4 Challenges in using Automod

Moderators noted that hard-coded rules and regexes are prone to mistakes

Automod's filtering mechanism, based on hard-coded rules and regexes, is effective at tedious and straightforward tasks. But Automod is prone to make mistakes as tasks get harder and need to take contextual information into account. P1 and P5 said that Automod consistently misses content that are violations (*false negatives*). Additionally, Automod filters "innocuous" content by mistake (*false positives*), and moderators stopped using Automod due to this reason. P6 said that, "*A lot of good comments get flagged by the AutoModerator and there was a lot of backslash from the community which is why I stopped using Automoderator.*" This aligns with findings from prior work examining AutoModerator's

[9] `https://www.reddit.com/r/AutoModerator/comments/2l4e2j/can_automod_remove_or_report_postscomments_that/`

challenges [157]. Moderation is a highly contextual task, that needs to take the community's norms into consideration before taking down content. Encoding all of the norms of a community in the form of simple regular expressions is not feasible.

5.4.1 Moderators find it hard to configure Automod

Given that regexes are written by the moderators themselves, configuring Automod is non-trivial. Despite the presence of thorough documentation, a lot of moderators have a hard time configuring Automod by themselves, including long-time moderators like P3 with over 8 years of experience (and currently moderates 10 different subreddits). P3 said, "*I personally have problems with configuring Automod because I don't understand the process. It gets quite complex because it uses filters, mark down rules and so on which are beyond me. They seem to require a level of coding knowledge.*" Moreover, the current design of Automod's configuration file makes it restrictive for a lot of moderators to use. Along these lines, P2 said, "*You have to be very comfortable with what feels a lot like a command line interface in a lot of ways. And that restricts the number of people who are able to interact with AutoModerator, to the ones who have the skillset and the inclination, which is probably the biggest restriction.*" Additionally, Automod configuration files for most subreddits quickly grow into massive lines of hard-coded rules in order to keep up different types of misbehavior. P5 said that this makes it hard to maintain: "*When configuration files for AutoModerator become huge, it can cause errors on Reddit due to large amounts of processing time required by the rules.*"

5.4.2 Moderators need to manually come up with new rules and constantly update Automod.

In order to keep up with the dynamic nature of abuse on the platform, 5 out of 11 moderators said that they have to constantly update Automod rules (e.g., come up with new lists of undesirable phrases or domain names for URLs). P2 said,

> "The AutoModerator is what it is, right? It is text string recognition. It doesn't

> learn, it doesn't adapt. If you're not constantly on top of it, it will quickly lose its relevancy. So what I had to do was constantly read bad comment strings and constantly update it with the latest dumb jokes, and that's an entirely manual process."

This static nature of Automod rules is a major drawback, despite its effectiveness in enforcing simpler formatting rules.

5.5 Summary of findings from formative interviews

Via formative interviews with 11 active Reddit moderators, I learned about currently available automated moderation tools on Reddit. Current automated tools like Automod fit into a deeply sociotechnical system of human- and machine-moderation. I identified three major areas of improvement for extending the capabilities of these tools.

- Current automated tools like Automod, though effective at performing simpler tasks, but are prone to making mistakes when tasks get harder.
- Moderators find that configuring Automod is hard and unintuitive, and the lack of an easy way to configure the tool is a major drawback.
- In order to keep up with the evolving nature of misbehavior on the platform, moderators maintain long lists of rules that need to be manually updated regularly.

CHAPTER 6

CROSSMOD: A CROSS-COMMUNITY LEARNING-BASED SYSTEM TO ASSIST REDDIT MODERATORS

Through my interviews, I found that mods need tools that adapt and learn. As P4 said:

> "I just need a smarter Automod. Automod is great because it can act on regular expressions. It can ban (spam) bots and report problems. It's a very strong tool, but it's a very simple tool. A machine learning model that can learn from past mod actions and remove content would be powerful, especially if it can do what a properly socialized and culturalized moderator can."

Next, I introduce *Crossmod*, a sociotechnical moderation system built to extend the capabilities of moderators while also fitting into their existing workflows reported in the previous sections. Figure 6.1 demonstrates this through an example walkthrough. Scaling the idea of machine learning-based sociotechnical interventions up to groups interacting via an online community presents challenges, because of the social norms that emerge within a particular group [158]. In order to address this challenge, I construct Crossmod by working with the selected partner subreddits in Table 5.1, incorporating key principles of mixed-initiative user interfaces [43]. The moderators (or "mods") of those groups know those norms best, and expend considerable effort enforcing them. At each stage of my system design process, I worked with mods in a participatory framework to inform key features: as a consequence, Crossmod permits great deal of moderator control over internal machine-generated predictions.

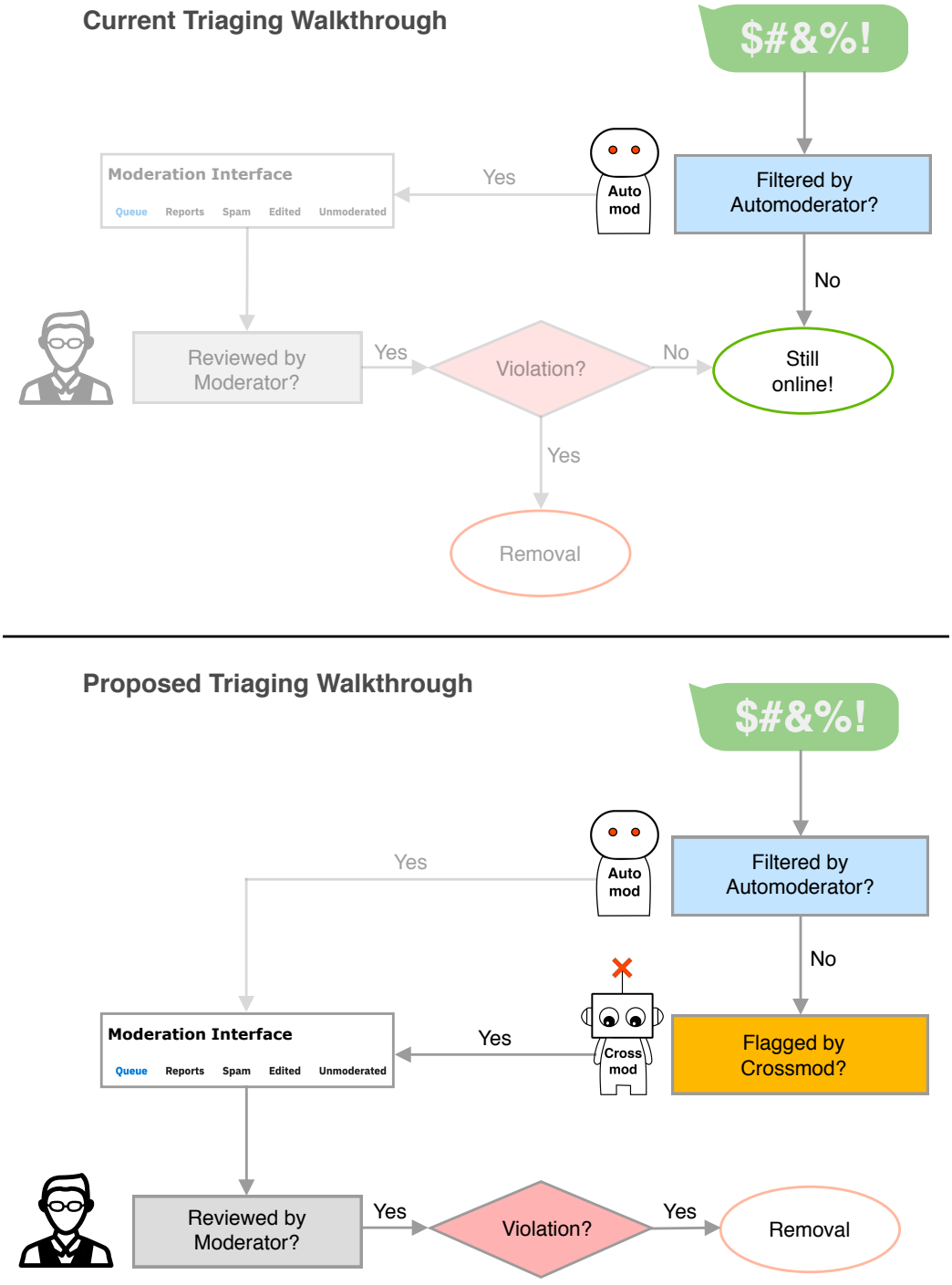

Figure 6.1: Comparison between the current triaging workflow a comment undergoes when posted to a subreddit, and how it changes with Crossmod. Note that this illustration represents subreddits where moderators only review content flagged by automated tools (i.e., the complete workflow is more complex than depicted here).

6.1 Crossmod: System design

Crossmod provides moderation recommendations for incoming comments in an automated manner, and makes its decisions based on *cross-community learning* (described in more detail below)—an approach that leverages a large corpus of previous moderation decisions via an ensemble of classifiers. Crossmod is designed so that it can be easily integrated into each subreddit's dedicated moderation interface.

6.1.1 How Crossmod is integrated into Reddit's moderation interface

Figure 5.1 illustrates how Crossmod is integrated into Reddit's interface to support different moderation actions. For example, if Crossmod is configured to report comments that violate community norms, the comments flagged by the system will be added to the *Reports* and *Mod Queue* tab in the Moderation interface, and this can be further reviewed by human moderators (since Crossmod has been granted *moderator-permissions* on the subreddit shown in Figure 5.1, all comments reported by Crossmod are also added to the central Mod Queue). In this illustration, both of the comments in the Mod Queue are automatically reported by Crossmod. All comments in the Mod Queue remain online until they are reviewed by human moderators. A moderator can perform one of three types of mod actions based on whether they agree with Crossmod's moderation recommendations or not. If they agree with Crossmod's moderation recommendation, then they just click "remove" on the interface, or mark the comment as "spam". If instead they disagree with Crossmod's report, and feel that the comment does not actually violate community norms, they can just "approve" the comment, and take it out of the Mod Queue.

In addition to reporting comments, Crossmod is designed to support a range of other functionalities: to send alerts to moderators in case of particularly sensitive topics (in the form of *modmails*), or to proactively remove comments that garner very high *abuse* scores computed by the pre-trained machine learning models described in Section 6.3. I detail

Figure 6.2: An illustration of the core idea behind *cross-community learning*. Using an ensemble of classifiers, we provide counterfactual estimates about what a set of source communities would do with new content from a completely different target community. In other words, "What would r/science do if this comment was posted there?" In my work, Crossmod's ML-backend provides counterfactual estimates about what 100 subreddits would do with new content, as well as whether that content resembles racism, homophobia, and so on.

the different moderation actions that can be supported by the system in further sections (an overview is shown in Table 6.2). The system pipeline is shown in Figure 6.3, and I describe each component in detail next.

6.2 System pipeline for Crossmod

Crossmod is a Reddit bot written in Python, and works as a subreddit-level moderator tool that can automatically detect comments violating community norms. Crossmod performs three main tasks:

Task 1: Constantly listening to the stream of incoming comments posted on a target subreddit. Every new comment that is posted on the subreddit will be ingested by Crossmod, and sent to the appropriate tab in the moderation interface, if necessary.

Task 2: Querying the ML back-end to obtain scores (or predictions) for each ingested comment. Crossmod's ML back-end is developed using *cross-community learning*,

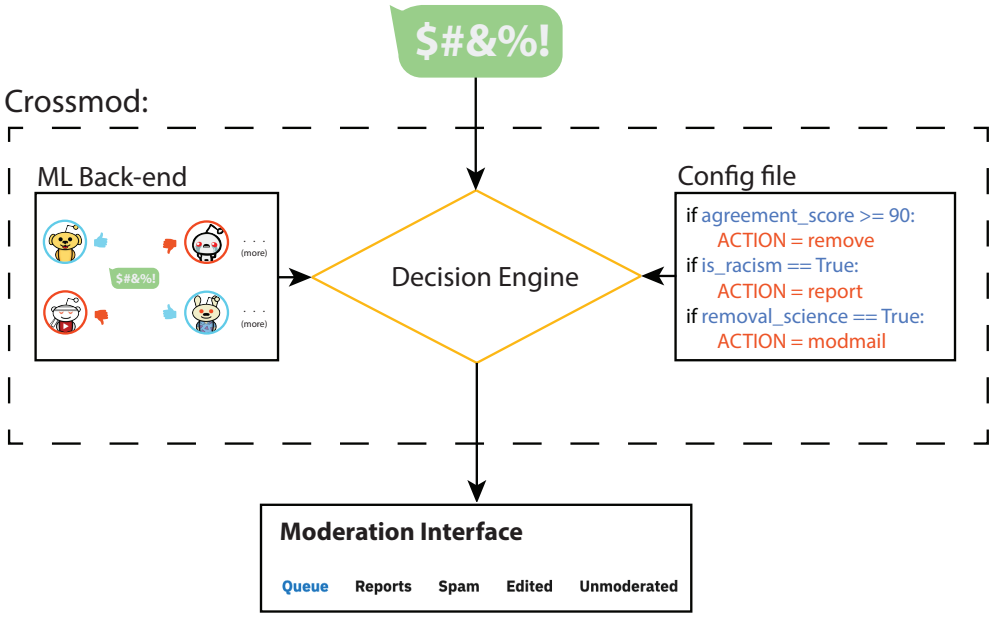

Figure 6.3: Flowchart depicting Crossmod's system pipeline. Crossmod makes its moderation decisions by obtaining predictions from an ensemble of *cross-community learning*-based classifiers. Crossmod wraps this back-end in a sociotechnical architecture that fits into Reddit's existing *Moderation Interface*. My system design allows moderators to easily configure Crossmod using simple conditional statements, and tailor its actions to suit community-specific needs.

leveraging a large corpus of previous moderator decisions via an ensemble of classifiers. The outputs obtained from the ensemble of classifiers are sent further down the pipeline. Currently Crossmod queries the back-end classifiers through a request to a remote server.

Task 3: Finally, the outputs from the ensemble of classifiers in the ML back-end are aggregated to compute *scores*. I describe the different types of scores that can be computed in further sections. Based on these aggregated scores, Crossmod detects comments that violate community norms. Depending on how Crossmod has been

configured by the subreddit's moderators, Crossmod takes the moderation <*actions*> that were triggered based on the <*conditions*> specified in the *configuration file*.

6.3 ML back-end based on cross-community learning

Crossmod's ML back-end comprises an ensemble of cross-community learning-based classifiers that are of two types: subreddit classifiers, and norm violation classifiers. The different classifiers I describe here were generated as part my prior study. Detailed information about the constructions of these classifiers are provided in Chapter 4.

Subreddit classifiers: First, I obtained 100 classifiers trained to detect whether a given comment will be removed by moderators of 100 popular subreddits. These were developed as part of my prior study (described in Chapter 4), and the names of all 100 subreddits I trained classifiers for can be found in Table 4.2 (under the column name *cluster subreddits*). These classifiers were trained using FastText, a state-of-the-art library [146, 147]. The parameters used to tune the FastText classifiers are described in Table 4.1 (see Chapter 4).

Macro norms classifiers: Next, I obtained 8 classifiers trained to detect comments that violate macro norms on Reddit. In my prior work, I found the existence of "macro norms" that are known to be enforced across most parts of Reddit [21]. For example, posting hate speech in the form of homophobic and racist slurs, using misogynistic slurs, graphic verbal attacks, and distributing pornographic material are removed by moderators of most subreddits. As a follow-up to the study, I publicly released a labeled dataset of Reddit comments labeled violating 8 different types of macro norms [159]. Using this labeled dataset of macro norm violations,[1] I trained FastText classifiers to identify comments violating the different types of *macro norms* shared across Reddit [21]. The parameter values described in Table 4.1 were used to tune the FastText classifiers. I tuned these parameters by grid search, trying the range of values described in Table 4.1, and selecting those which maximized the F1 (f-measure) across 10-fold cross-validation in aggregate over all

[1]This dataset is publicly available, and can be found on Github: `https://github.com/ceshwar/reddit-norm-violations`

macro-norms. Given the marginal differences in performance across the parameter space, I decided to use the best performing parameter values from Chapter 4 (shown in Table 4.1) to train all macro-norm classifiers.

This ensemble of 108 pre-trained, cross-community learning-based classifiers—100 subreddit classifiers, and 8 (macro) norm violation classifiers—constitute Crossmod's ML back-end. The main function of the ML back-end is to make predictions about a new query comment using the ensemble of classifiers. The final output from the ML back-end is the list of predictions obtained from the ensemble of classifiers. This is depicted in Figure 6.2. Through Crossmod, I bring the idea of *cross-community learning* to inform online moderation into production. Crossmod is the first open source, AI-backed moderation system to be released publicly, and the system can be easily adopted by new and emerging online communities *off-the-shelf*. The goal of providing the ensemble of classifiers in Crossmod is to provide moderators with the following choice—which subreddits would they like to emulate. It may be possible to use the frame of macro norms to derive normative guidelines for new and emerging online communities. In other words, the ensemble of classifiers used in Crossmod's back-end may serve as sensible defaults for a new online community. Additionally, using predictions obtained from an ensemble of 108 classifiers, instead of relying on a purely in-domain classifier trained on comments that are inherent to the target subreddit only, brings in more diversity and robustness into the decision-making process.

6.4 Configuring the Crossmod: *If This Then That* (IFTTT) format

In my sociotechnical intervention, a machine learning system will intervene in a normally unmediated process. In other words, Reddit would usually immediately post those comments. Issues of agency naturally arise in mixed-initiative systems like the one I build. My approach is to empower the mods in Crossmod. Here, this means that moderators can review and ultimately reject moderation recommendations made by Crossmod's back-end ML (the models can even learn from those corrections). Beyond that, I develop rich sets

of options moderators can use to control and configure Crossmod to meet their subreddit-specific needs.

During the interviews, I learned that configuring AutoModerator (through regexes in a YAML file) was a challenge for most communities. I found that one of the needs of moderators was just to design a moderator tool that simplifies the way it can be configured. P2 said:

> "Well, it also needs to be straight up more user friendly to just even do what we're currently doing. Like even if it was just easier to type in the text without having to get all of the semantics of the Automod configuration properly done, that would be helpful, just in and of itself."

I adopt an *If This Then That* (IFTTT) format where moderators just create chains of simple conditional statements to trigger moderation actions. In other words, the configuration file of Crossmod contains a list of IFTTT commands written in the following format:

> *if* <**condition**>: <**action**>

Through initial rounds of feedback obtained from moderators, I found the use of such simple conditional statements makes configuring Crossmod intuitive and straightforward to moderators without (any) coding experience. Next I present the different types of <**conditions**> that are supported by Crossmod's ML back-end, and then review the different types of <**actions**> that Crossmod can be configured to perform. Figure 6.4 shows an example Crossmod configuration file containing some example conditional statements.

6.5 <conditions> supported by Crossmod's ML back-end

Crossmod makes its decisions based on predictions about a *new query comment* obtained from the ML back-end. Using the output returned by the ML back-end, the decision engine of Crossmod can be configured to evaluate the three types of <**conditions**> shown in Table 6.1. I describe each of them in detail below.

```
if agreement_score >= 95:
    ACTION = remove
if agreement_score >= 80:
    ACTION = report
if is_racism == True:
    ACTION = modmail
if is_misogyny == True and is_homophobic == True:
    ACTION = report
if removal_science == True:
    ACTION = report
if agreement_score >= 90 and removal_The_Donald == False:
    ACTION = remove
```

Figure 6.4: Example configuration file for Crossmod. In this config file, the mod is auto-removing comments with very high agreement scores, and reporting those with moderate scores. In addition to using specific macro norm and subreddit scores, on the last line the mod has exempted r/The_Donald from the agreement score.

Table 6.1: Different types of conditional statements that can be used to configure Crossmod. The values for agreement_score, is_racism, removal_science are computed using the predictions returned by the different classifiers present in the ML back-end.

<**condition-type**>	**Example conditional statements**
removal agreement score	if (agreement_score >95%):
macro norm violation	if (is_racism = True):
specific subreddits would remove	if (removal_nfl = True):

6.5.1 <**condition**>: *agreement_score*

First, I use the subreddit-specific classifiers built for detecting comments that would be removed by moderators of 100 popular subreddits. Using the predictions obtained from all of the 100 source subreddits, for a given unseen comment from the target community, I obtain an *agreement_score*—the percentage of source subreddits that consider the particular comment to be a norm violation (i.e., "If the comment were hypothetically posted on the subreddit, would moderators remove it?"). In other words, if the *agreement_score* computed for a comment is high (say 95%), it denotes a majority agreement to remove

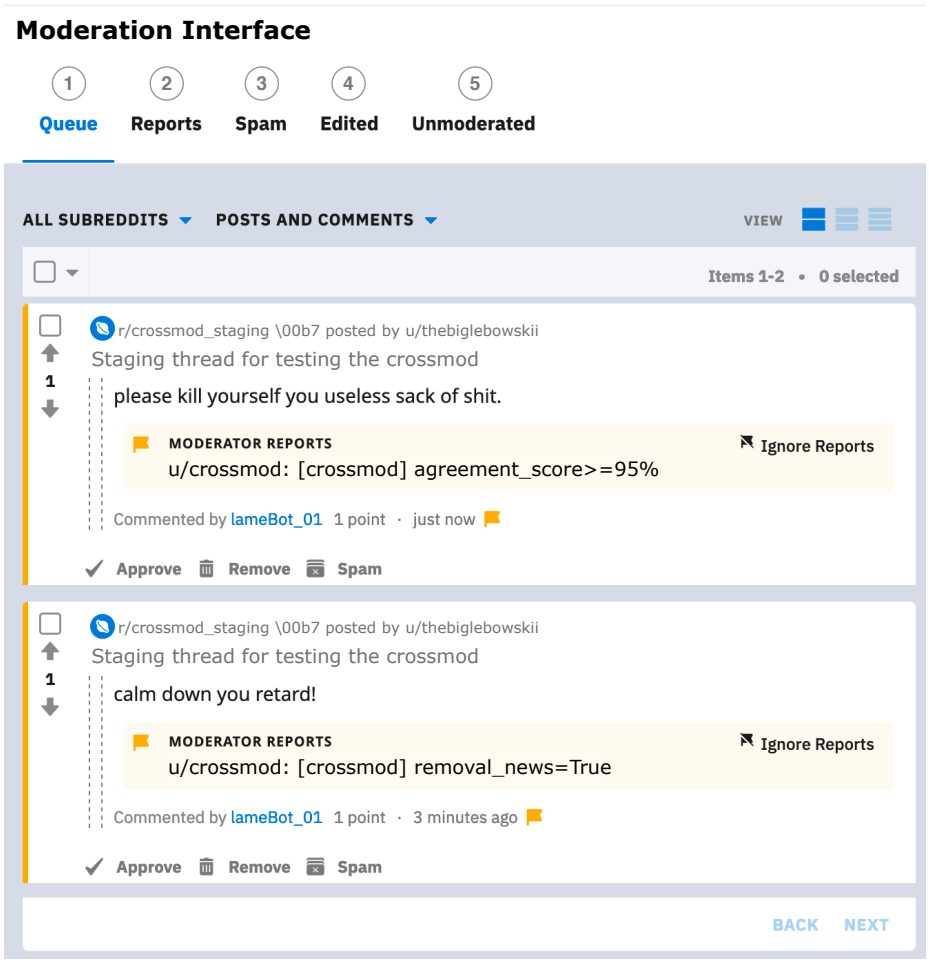

Figure 6.5: How Crossmod is integrated into Reddit's existing moderation interface to support *triaging*. When Crossmod is configured to triage (as opposed to outright remove), comments flagged by Crossmod will be *reported*, and sent for further review by human moderators. In this example, the first comment in the moderation queue is reported for obtaining an $agreement\_score$ over 95%, while the second comment in the queue is reported for because the classifier trained for r/news predicts removal (i.e., $removal\_news = True$). All comments pushed to the moderation queue are reviewed by moderators, and the moderator can perform three different actions based on whether they agree/disagree with Crossmod's report. If they disagree with the report, and feel that the comment does not violate community norms, they can "approve" the comment. Instead, if they agree that the comment does indeed violate community norms, then they can either "remove" the comment, or mark it as "spam".

the comment, among most of the subreddit classifiers present in the ensemble. An example conditional statement of this type would be: $if(\text{agreement\_score}) > 0.95$. Given this <**condition**>, Crossmod will only detect comments that at least 95% of subreddit classi-

fiers in the ensemble predict to remove.

When interviewing moderators for design feedback on my system prototype, I observed that moderators liked the idea of using Crossmod to make moderation recommendations within their subreddit by simulating moderation decisions by other moderators from 100 popular subreddits. P4 said that conditional statements based on $agreement\_score$ would definitely be useful for the subreddits they moderate (i.e., r/news, and r/todayilearned): "*Most useful for me would be the agreement_scores returned by the 100 subreddit classifiers*". Along similar lines, P8 said: "*Agreement is super helpful where you can straight up remove stuff without intervention or flagging it.*"

6.5.2 <**condition**>: $removal\_\{subreddit_i\}$

Next, I use the predictions obtained from individual source subreddits or a group of specific source subreddits for a given unseen comment from the target subreddit. Moderators can configure Crossmod to detect target community comments based on whether a subset of source communities would consider a given comment as a norm violation (if it were hypothetically posted on the source subreddit, a moderator would remove it). An example condition of this type would be: $if(\text{removal\_nfl}) = True$, where we are checking if the classifiers trained for a specific subreddit, r/nfl, predicts to remove the given comment.

In addition to the $agreement\_score$ described previously, moderators asked for this ability to configure Crossmod to make moderation recommendations based on only certain subsets of classifiers (instead of all of them). P0, P1, and P4 said that such conditional statements help smaller, and topically-similar subreddits. In particular, P1 said that,

> "That seems useful because let's say, sports subreddits, for example, r/soccer. If you wants to make a subreddit for a local football team that's not so broad, then learning from what r/soccer does would be useful for moderating."

Figure 6.6: A comment found to violate community norms is proactively removed by Crossmod, and leaves a *tombstone* in place of the comment. All comments that are proactively removed by Crossmod will be sent to a *Spam* tab in the Moderation Interface (denoted by tab #3 in Figure 5.1).

6.5.3 <**condition**>: $is\_\{violation_{norm}\}$

The last type of <**conditions**> supported by Crossmod is based on predictions obtained from the 8 classifiers trained to detect specific types of macro-norm violations (described in Table 4.3). The goal of each of these classifiers is to identify whether an unseen comment from the target subreddit is a macro-norm violation, or not.

An example conditional statement of this type would be: $if(\text{is\_racism}) = True$, where we are checking if the classifier trained to detect macro norm violations predicts the given comment to be racist in nature. P3 (r/news) and P4 (r/science) said that Crossmod's ability to automatically detect such harmful content would be very useful for the subreddits they moderate. In particular P4 said,

> "(Detecting) macro norm violations would be very helpful for my subreddit (r/science). Automod can capture obvious use of hate speech, but a tool that can recognize less obvious speech is a very useful tool."

6.6 <**actions**> that Crossmod can perform

Currently, Crossmod's design allows the system to perform three types of <**actions**> with varying levels of autonomy, shown in Table 6.2. Depending on the severity of removal <**conditions**>, and the target subreddit's preferences, moderators can configure Crossmod

Figure 6.7: In this example, the comment triggered the *modmail* <action> supported by Crossmod. As a result, a modmail was sent to alert the moderators, along with the corresponding <condition> violated by this comment. Moderators can review this comment by clicking on the URL (i.e., *permalink*) pointed to in the modmail from Crossmod. The mods can then choose what to do next.

to take specific types of moderation actions. The goal here is to provide moderators with the flexibility in choosing the <**conditions**> and <**actions**> based on their specific requirements, or the level of autonomy they would like to grant Crossmod. Though there were a few other specific moderation <**actions**> that were requested by individual moderators (e.g., locking the post, adding a spoiler, or marking as not safe for work), I focus on three types of <**actions**> in Crossmod's current design. I found that almost all of the moderators I spoke to agreed on the need for these three <**actions**>.

6.6.1 <**action**>: remove

As illustrated in Table 6.2, the first type of moderation *action* that Crossmod can perform is *direct removal*. Once a comment violates the <**condition**> which triggers the direct removal <**action**>, Crossmod automatically removes the comment from the subreddit. As shown in Figure 6.6, the comment is replaced with a *tombstone* indicating that it was removed by a moderator, in this case Crossmod.

6.6.2 <**action**>: report

The next type of moderation action that Crossmod can perform is reporting a comment, and thereby sending a comment for further review by moderators. Through the report <**action**>, Crossmod can serve as a *triaging* tool for subreddits, automatically detecting

Table 6.2: Different types of moderation actions that are supported by Crossmod.

<**action**>	**Description**
remove	Remove from the subreddit, and move to *Spam* tab in moderator interface
report	Send to mod queue for review by human moderator
modmail	Alert human moderators through a modmail or message

comments that are likely to be undesirable and sending them for moderator review. Figure 6.5 illustrates this. All reported comments continue to remain online until a moderators decides to remove it. Instead of taking moderation decisions autonomously, Crossmod can help augment automated predictions with human judgment. P5 and P9 said that,

> "If there's a rule you're not sure about, then don't remove the comment directly. Instead, just triage it so that mods can remove/approve it after review."

6.6.3 <**action**>: modmail

The final type of <**action**> that Crossmod can perform is alerting the moderators about the presence of new undesirable content on their subreddit. Crossmod can be configured to send alerts over modmail, or in the form of direct messages to specific moderators. An example of Crossmod sending an alert over modmail is shown in Figure 6.7. P0 said,

> "Sending a Modmail would be a good one for sure. That's a different way of reporting. So it to be, oh, I detected a comment that appears like I'm 90% confident is hostile, you should take it and look at it, here's the link. That can be a more detailed way of reporting things, since it's difficult to put much information in a report."

For instance, though most moderators liked the idea of alerting through *modmail*, P1 said they preferred the *remove* and *report* <**actions**> over *modmail*:

> "For r/photoshopbattles, sending comments that violate community norms to Mod queue would be more helpful than sending it to the Modmail directly."

6.7 Design elements to track Crossmod <**actions**>

Under <**action**>: remove, Crossmod has full autonomy over its moderation decisions and it can directly alter conversations within a target subreddit by removing comments (without the need for any human review). As a result of this, I included design elements in the system that can help moderators easily track Crossmod's moderation actions, and reverse Crossmod removals if necessary.

Send removals to Spam tab in Moderation Interface. All direct removals by Crossmod are automatically sent to the *Spam* tab in the moderation interface. This allows mods the flexibility to audit direct removals made by Crossmod, and even reverse Crossmod removals by approving the comment, in case of false positives. This also enables a post-hoc analysis of Crossmod's moderation decisions in the future, and monitor false positive rates (i.e., how many times did human moderators overturn Crossmod's moderation actions?).

Provide removal reasons. I designed Crossmod so that it can be configured to provide a *removal reason* to moderators and/or the user (i.e., author of the comment), in the form of a *reply* and/or a *modmail*. P3, P6 and P9 mentioned that providing removal reasons allows users and other mods to understand why a certain comment was removed by the system:

> "You'd have to spend time either figuring out or guess and approve, remove it based on your own quick judgment. So you should have a removal reason, this has been removed because X, Y, Z." (P3)

Alert moderators after removing content. Finally, Crossmod can be configured to send out a *modmail* informing the moderators about a direct removal made by the system. An example of this is shown in Figure 6.8. Additionally, I am also brainstorming design ideas with moderators on ways to use community feedback to reverse Crossmod removals when necessary. P10 suggested an idea to use community-feedback on the Crossmod's reply containing the removal reason to identify "false positives", and reverse an automated removal. For example, if Crossmod's reply receives over 5 downvotes (indicating that at least 5 users

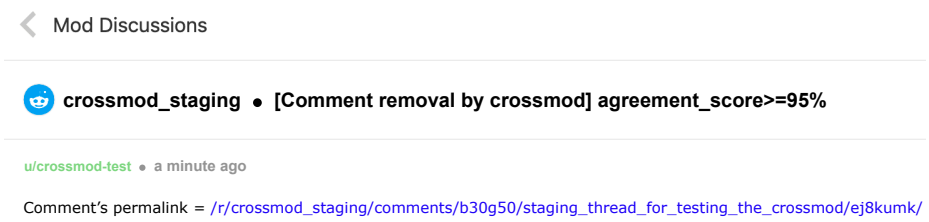

Figure 6.8: A comment was directly removed by Crossmod, and a modmail was sent to alert the moderators about this action. Moderators can review this automated decision by clicking on the URL pointing to the removed comment. If they disagree with Crossmod, they can reverse the decision by "approving" the comment.

disagree with Crossmod's decision that the comment is a violation), Crossmod could be configured to reverse its <**action**> by *approving* the removed comment. This functionality is not supported currently, but I plan to explore this in future iterations of Crossmod. But this idea might be vulnerable to "brigading"; for example, a bunch of users could collude and downvote content they disagree with. As a counter-measure, we would need to take user reputation into account (e.g., [94]) and think of ways to identify "reliable feedback" computationally.

6.8 Deploying Crossmod in a controlled environment

As a proof-of-concept summative evaluation, I deployed Crossmod in a controlled environment, simulating real-time conversations from two large test subreddits with over 10M subscribers each. I created a staging instance for each test subreddit, and simulated actual conversations that took place within the test subreddit, in an unobtrusive manner. When I interviewed Chad (P0), the creator of Automod, suggested a similar deployment study:

> I think even if you could just come up with something that like hypothetically shows, "hey the bot is watching all the comments in our subreddit, here's the one who's it would have acted on." And if they (moderators) can see that and just see, oh yeah, that actually catches a ton of stuff we're doing manually right now, then that would be a huge thing to make them more confident in

implementing it. And that gives an upfront way to see what things it would do instead of saying, "here, put this in and only then we can see what it does."

6.8.1 Test subreddits: r/science and r/Futurology

I deployed Crossmod in a controlled environment simulating real-time conversations from r/science and r/Futurology. r/science is one of the largest communities on Reddit with over 20 million subscribers, and it is a "place to share and discuss new scientific research." r/Futurology is another large community with over 13 million subscribers, and it is "devoted to the field of Future(s) Studies and speculation about the development of humanity, technology, and civilization."

Creating staging instances to mirror test subreddits

First, I created a staging instance for each of my test subreddits. The staging instance is a controlled environment created to mimic conversations within test subreddits, and I obtained test subreddit comments in an unobtrusive manner. For example, the staging instance created for r/science was a subreddit mirroring r/science, and therefore contained a copy of all comments posted to r/science. I created the staging instances by streaming all test subreddit comments posted during a 4-month period, from September 2018 to December 2018. I used Google BigQuery to obtain these comments, and then (re-)posted all of them on the staging instances using the Reddit API.

Deploying Crossmod in the staging instances

I deployed Crossmod in the staging instances so that it would monitor all comments streamed from r/science and r/Futurology. My goal was to simulate the real-time conditions under which Crossmod is intended to be used by moderators. For the purpose of this evaluation, Crossmod was configured to "report" comments that obtained an $agreement\_score >= 85\%$ from the system's ML back-end. Under real-world circumstances, mods would con-

figure the Crossmod with their own custom config file; here, I tested a baseline scenario to see how well Crossmod could perform relative to existing tools. (See Future Work for longitudinal deployment plans.)

6.8.2 Evaluating reported comments with the help of moderators

I recorded the moderation recommendations made by Crossmod when deployed in this controlled environment. In particular, I stored the *id* (i.e., unique identifier for the comment) and *body* (i.e., actual text in the comment) of all comments reported by Crossmod, along with the $agreement\_score$'s computed for each comment. In order to evaluate Crossmod's moderation recommendations, I asked 2 moderators from each test subreddit to decide whether they would allow the comment on their subreddit or not. I conducted human review of comments in 2 phases.

Phase 1

In Phase 1, two moderators from r/science independently rated a set of 100 comments collected from their subreddit. Out of the 100 comments shown to both moderators, 50 of these comments received an $agreement\_score >= 85\%$ (i.e., high scoring), while 50 of these comments received an $agreement\_score <= 1\%$ (i.e., low scoring). Moderators were not told about the distribution of high scoring to low scoring comments in order to avoid any cognitive biases in the decision-making process. Similar to the above process, two moderators from r/Futurology also independently rated a set of 100 comments collected from their subreddit.

Phase 2

In Phase 2, the four moderators were asked to rate a larger set of comments obtained from their respective subreddits. In this phase, I asked moderators to only review a random sample of 650 comments scored highly by Crossmod (i.e., $agreement\_score >= 85\%$).

Table 6.3: Evaluation of Crossmod in Phase 1 where I compare labels provided by 4 moderators and predictions from Crossmod on an equal distribution of high-scoring and low-scoring comments.

	Accuracy	**Precision**	**Recall**
r/science	0.92	0.98	0.875
r/Futurology	0.86	0.72	1.0

In Phase 1, I found that Crossmod achieved a low false negative rate (less than 0.125). As a result, I chose to have the moderators only focus on the comments that were flagged by Crossmod (i.e., high scoring) in Phase 2, as these resemble the type of comments that would be reported by Crossmod upon real-time deployment. In order to prevent bias from the manual review process, I did not disclose the classifier scores (or sampling strategy) to the moderators. As a result, the moderators were only told that the comments were obtained from their respective subreddits, and they did not know whether a comment was actually reported by Crossmod or not.

I would like to note that moderators on Reddit perform large amounts of human labor on a voluntary basis, manually regulating content in order to help maintain healthy conversations on the platform. As mentioned earlier, moderators are struggling to keep up with the vast amounts of content generated within their communities [93]. In addition to this manual labor, the task of moderating content also involves emotional labor as well, with moderators having to view gruesome and disturbing content regularly [88]. Out of respect for the moderators' time and to keep the amount of effort required to review content manually, I asked each moderator to review a random sample of 170 comments scored highly by Crossmod in Phase 2.

6.8.3 Results

Phase 1

In Phase 1, a total of 200 comments obtained from 2 large-scale subreddits were labeled by 4 moderators from these subreddits, ensuring that an equal distribution of high-scoring and low-scoring comments were reviewed by each moderator. I found high inter-rater agreement among the two moderators from each subreddit when deciding whether they would allow the set of 100 comments on their subreddit or not—r/science moderators disagreed on 3 comments, and r/Futurology moderators disagreed on just 1 comment. In the case of disagreements, I asked moderators to discuss and provide a consensus label for the comments (i.e., remove or not). Using the labels assigned by moderators upon review as ground truth, I evaluate Crossmod's performance and the results are shown in Table 6.3. In Phase 1, Crossmod achieved high recall of over 87.5% for both r/science and r/Futurology, with an overall accuracy of over 86%. Moderators told me that they preferred that Crossmod achieve higher recall over precision. This is because moderators' intend to use Crossmod as a reporting tool to triage norm violating comments. Therefore a system that is able to detect as many violations as possible (i.e., higher recall at the cost of precision) is desirable because every reported comment will be eventually be going through human review—moderators can correct false positives during the review. In future iterations of Crossmod, I plan to explicitly give moderators the ability to tune the precision-recall trade-off (e.g., by adjusting the threshold for $agreement\_score$ based on which comments are flagged).

Error analysis: Next, I examined the comments that were misclassified by Crossmod in Phase 1, and found that *false positives* were comments that contained URLs (e.g., links to Wikipedia, or Imgur), and comments that excessively used swear words for exclamation. Some actual Reddit comments that were false positives (i.e., flagged by Crossmod, but not removed by moderators) are provided below:

- *for the love of god wash your fucking bags you gross fucks sincerely people that bag*

your shit

- *www.reddit.com/r/[id omitted]/iama_vacuum_repair_technician_and_i_cant_believe*
- this is hoarding wealth - https://forbes.com/[rest of the URL omitted]
- commenting on reddit aka saying dumb shit

False negatives included comments that were off-topic and anecdotal, and prior work has found that these are considered to be micro norm violations within scientific communities like r/science [21]. One approach to account for community-specific norms would be to use Crossmod along-side a purely in-domain classifier trained on moderated comments from a target community (e.g., r/science). This would allow moderators to make decisions based on moderation recommendations obtained from an ensemble of classifiers trained to encode *Reddit-wide* norm enforcements (e.g., macro norms) along with *community-specific* recommendations (e.g., micro norms). I plan to explore this line of work in the future.

Phase 2

In Phase 2, the 4 moderators reviewed a total of 680 unique comments (each mod reviewed 170 unique comments found within their subreddit) reported through Crossmod. Given the high inter-rater agreement observed in Phase 1, I decided to show a non-overlapping set of unique comments to each moderator from r/science (and r/Futurology) in Phase 2. This allowed me to evaluate Crossmod's performance by reviewing a larger sample of comments. Overall, moderators decided that they would have removed 648 (95.3%) of the comments detected by Crossmod. The r/science moderators decided that 338 out of the 340 comments detected by Crossmod would have been removed from r/science. The r/Futurology moderators decided that 310 out of the 340 comments detected by Crossmod would have been removed from r/Futurology.

Furthermore, for all of the reported comments that were decided to be violations by moderators, I queried the Reddit API (by their unique *id*). My goal was to examine whether

any of these comments were still online on r/science and r/Futurology. I found that out of the 338 comments decided to be violations by r/science moderators, only 10 comments were actually removed from r/science (i.e., taken off the site). In other words, Crossmod was able to detect 328 comments that moderators would have removed, but were previously *missed* by existing moderation tools like AutoModerator. Similarly, I found that 309 out of the 310 comments decided to be violations by r/Futurology moderators were still present online, but were detected by Crossmod.

6.9 Discussion

I deployed Crossmod in a controlled environment, simulating real-time conversations from two large subreddits with over 10M subscribers each—r/science and r/Futurology. Two moderators from each subreddit evaluated Crossmod's moderation recommendations by manually reviewing comments scored by Crossmod that had been drawn randomly from existing threads. In Phase 1, Crossmod achieved an overall accuracy of 86% when detecting comments that would be removed by moderators, with high recall (over 87.5%). In Phase 2, moderators decided that they would have removed 648 of the 680 (95.3%) comments detected by Crossmod; however, 637 of 648 (98.3%) were still online at the time of this writing (i.e., not removed by current moderation tools). While necessarily incomplete and proof-of-concept, these results indicate that Crossmod will significantly extend the capabilities of moderators when deployed in the wild.

6.9.1 Addressing the gap in current moderation practices

As mentioned earlier, human moderation struggles to keep up with the immense volume of content generated within large-scale platforms—plenty of content that violates site guidelines remains online for days, sometimes even years [22]. Additionally, I found that current automated tools on Reddit are not robust, and prone to mistakes through my interview study (see Section 5.4). In particular, it is hard to quantify the rate of *false negatives* (i.e., how

many actual norm violations are being missed by existing tools) due to the reasons mentioned above. My findings provide some clarity on this issue. The results from Phase 2 of my evaluation quantifies the gap in currently deployed moderation approaches (i.e., what are current approaches missing out on), and how Crossmod can help address this gap by extending their capabilities.

6.9.2 Towards real-time deployment on Reddit

My ultimate goal is to push Crossmod into production across Reddit as a real-time moderation system. As the creator of AutoModerator, Chad, said, "*A lot of moderators are quite disappointed in how few moderators tools there are. So when something new comes out, they're pretty quick to adopt that.*" I am currently in conversations with moderators from several subreddits, including r/Futurology, about deploying my system real-time on Reddit. As the first step in this direction, Crossmod is currently deployed as a real-time *reporting* tool to triage norm violating comments for further moderator review on r/Futurology.[2] I hope that by releasing Crossmod publicly, Crossmod can be adopted by moderators and researchers going forward.

6.9.3 Future work: Next steps for Crossmod

I plan to extend Crossmod in the future through the following ways:

Graphic UI to help mods configure the system.

Moderators found configuring Crossmod using simple conditional statements straightforward. In addition to this, I plan on introducing a graphical user-interface that simplifies configuring Crossmod even further.

[2] `https://www.reddit.com/r/Futurology/comments/eceiqz/using_an_ai_bot_trained_on_human_mod_actions_to/`

Incorporate community feedback into Crossmod

I am currently brainstorming design ideas with moderators on ways to use community feedback in Crossmod. One idea is to reverse Crossmod removals when necessary. For example, P10 suggested an idea to use community-feedback on the Crossmod's reply containing the removal reason to identify "false positives", and reverse an automated removal. For example, if Crossmod's reply receives over 5 downvotes (indicating that at least 5 users disagree with Crossmod's decision that the comment is a violation), Crossmod could be configured to reverse its <**action**> by *approving* the removed comment. This functionality is not supported currently, but I plan to explore this in future iterations of Crossmod.

Ensuring fairness and transparency in moderation decisions.

I plan to conduct a systematic analysis of any algorithmic biases [160, 113] inherent in Crossmod's machine learning back-end, and incorporate design changes that ensure fairness, accountability and transparency in the moderation system [161, 162, 163]. As a first step, I added "removal_reasons" to the system's design, facilitating better sense-making of Crossmod's actions. Additionally, using an ensemble of classifiers to make moderation decisions introduces diversity into Crossmod's decision-making process, potentially reducing biases that are likely.

Publicly release Crossmod with API keys.

I plan to release Crossmod publicly through an API, and this would allow us to evaluate the entire system, in addition to the core algorithm that was already evaluated in this work. Prior work has identified that norms are shared across different Reddit communities (e.g., macro norms are found to be universal to most parts of Reddit) [21]. Though I evaluated Crossmod only on a couple of subreddits in this paper, I believe that the different <**conditions**> supported by the system would allow Crossmod to identify norm violations within other subreddits as well. But moderators asked me to take caution before releasing

the data and code for Crossmod publicly since bad actors may easily find ways to circumvent the system. In order to prevent such situation, I will be using API keys to control who has access to Crossmod.

Default configurations for Crossmod's back-end.

In future iterations of Crossmod, I plan to include default configurations for the ensemble based on prior configs used by previous moderators, or automatic tuning based on historical moderated data obtained from the target subreddit under consideration (i.e., automatically selecting which classifiers to use, in addition to asking the moderators to choose manually).

Dynamic re-training of classifiers.

Currently, Crossmod's ML back-end consists of an ensemble of classifiers that are pre-trained. Over time, adversaries might develop ways to circumvent static classifiers, and the norms for what is acceptable (and unacceptable) might evolve with time. In order to keep up with the evolving nature of online behavior, I plan to explore online learning approaches to dynamically re-train the classifiers based on real-time decisions from target communities. One idea for maximizing performance gains from re-training is to over-sample comments that are near Crossmod's *decision boundary*. This sampling strategy would enable avenues for targeted increases in the system's overall precision without losing out too much on recall.

6.10 Conclusion

Through interviews with several moderators, I found that the majority of moderators use automated tools to make the task of curating a large-scale platform like Reddit managable. However, there are very few automated tools that can assist moderators, and existing tools are limited in their scope. I developed *Crossmod*, a sociotechnical moderation system built to extend the capabilities of moderators, while also fitting into their existing workflows. I

would like to emphasize that Crossmod is not intended to replace human moderation or any of the existing automated tools used by moderators. Instead, Crossmod aims to extend the current capabilities provided by existing social and automated tools. My summative evaluation shows that Crossmod can improve moderation by detecting undesirable content which is currently undetected by the existing sociotechnical infrastructure.

CHAPTER 7
CONCLUSION

I began this dissertation aiming to develop a deep understanding of abusive online behavior via statistical machine learning techniques to build tools that help counter it, with the goal of making the Internet a more welcoming place. I have developed computational approaches to model abusive online behavior, aiming to address two of the major gaps in this line of research—the scarcity of labeled ground truth required to train effective ML models, and the contextual nature of online moderation by accounting for community-specific norms. First, I introduced a new class of machine learning tools that are based on cross-community linguistic similarity. Next, I discovered the existence of widely overlapping norms, across distinct online communities, suggesting that new automated tools for moderation could find traction in borrowing data from communities which share similar values. The abuse models that I build will enable a brand new class of interactive machine learning systems that can sidestep the need for site-specific classifiers. My thesis brings these pieces together in the form of open source software to detect abusive behavior online through cross-community learning, and thereby socio-algorithmically govern speech on large-scale Internet platforms like Reddit.

In this chapter, I conclude with some open questions that emerge from my work and suggest directions for future research.

7.1 Online moderation as a sociotechnical phenomenon

Sociotechnical research is premised on the interdependent and inextricably linked relationships among the features of any technological object or system and the social norms, rules of use, and participation by a broad range of human stakeholders [164]. This mutual constitution of social and technological is the basis of the term sociotechnical. Mutual

constitution directs scholars to consider a phenomenon without making a priori judgments regarding the relative importance or significance of social or technological aspects (e.g., [165, 166]). In this thesis, I examined currently deployed approaches to content moderation on Reddit, shedding light on the sociotechnical nature of online governance. Prior work on online moderation has focused on just the social system (e.g., [32]), or the technological system (e.g., [167]), or even the two side by side (e.g., [114, 64]). This thesis extends this line of work by investigating the phenomena that emerges when the two interact, and aims to ameliorate the *socio-technical* gap [48] in moderation systems. Instead of focusing on the behavioral or the technological, I studied the emergent sociotechnical phenomena that sets this thesis apart. I hope to see more theoretical work exploring online moderation along these lines in the future.

7.1.1 Participatory methods to inform the design of socio-algorithmic governance

Machine learning algorithms have become core design elements in modern social computing systems. Most often, they are designed and implemented by corporations who control access to an algorithm's core data and code. Here, I have followed emerging research in exploring a different path: working in dialog with users and key stakeholders to design socio-algorithmic systems [168, 169] that govern large social spaces. Crossmod is one data point in the space of possible socio-algorithmic governance systems that could result with participatory input [170]. Participatory methods should continue to inform the design of socio-algorithmic governance in the future.

7.2 Going beyond detection towards enforcement in online moderation

Current research on automated approaches is focused on the detection-side of online moderation, but there is a gap in the study of the enforcement-side, and the considerations to be made during this process [112]. Depending on the platform and the specific community under consideration, enforcement strategies vary significantly in intent and execution [21],

and these nuances should inform detection strategies in the future. This thesis aims to bridge this gap, and explored how to take the next step by developing a new AI-based moderation system [170] that can be easily customized to detect content that violate a target community's norms and enforce a range of moderation actions.

7.2.1 Agency and configurability in AI-backed sociotechnical interventions

Crossmod turns over control and oversight to empowered mods who can direct the underlying algorithms as they see fit. In sociotechnical interventions like Crossmod, an algorithm (e.g., machine learning) intervenes in a normally unmediated process. Issues of agency naturally arise in mixed-initiative systems like the one presented here. I turned to building configurability and oversight into Crossmod as a solution, by developing rich sets of options mods can use to control and configure Crossmod. As a consequence, Crossmod permits a great deal of mod control over internal machine-generated predictions, and can be tailored to meet subreddit-specific needs. Moderators can easily configure Crossmod using simple conditional statements, and also track moderation actions taken by Crossmod. In cases where moderators disagree with Crossmod, they can overrule its judgments.

7.3 Norms matter in online moderation

Social norms are rules and standards that are understood by members of a group, and that guide and constrain social behavior without the force of laws [68, 65]. Norms can be nested, in that they can be adopted from the general social context (e.g., use of pejorative adjectives are rude), and more general internet comment etiquette (e.g., using all caps is equivalent to shouting). Yet, norms for what is considered acceptable can vary significantly from one community to another, making it challenging to build one abuse detection system that works for all communities [21].

Current ML methods are largely context- and norm-agnostic, which leads to situations where content is removed unnecessarily when deemed inappropriate (i.e., false positives),

eroding community trust in the use of computational tools to assist in moderation [112]. A common failure mode for sociotechnical interventions like automated moderation is failing to understand the online community where they are being deployed [36]. Such community-specific norms and context are important to take into account, as ML researchers are doubling down on context-sensitive approaches to define (e.g. [159]) and detect abuse (e.g., [171]).

Moderators play a key role in governing online communities, and some of them have been doing their jobs for an extended period of time. By examining what the moderators actually remove, we can build better tools for triaging content that violates the community's norms. In Chapter 4, I examined what this space of online norms looks like empirically, by analyzing actual comments removed by moderators on Reddit. I observed that not all of the comments that get moderated are abusive or hateful in nature. There exist many other non-trivial, community-specific norms that get violated, resulting in moderator removals.

7.3.1 Designing automated moderation tools that account for community-norms

The mixed-initiative moderation system I developed for Reddit is the first exploration into a brand new, transformative class of technological artifacts, based on cross-community learning. In the end, the norms of an online community are best understood by its members and moderators, as the community defines which types of speech should be valued, or considered as norm violations. By working closely with Reddit moderators, I explored how they may regulate their communities by augmenting automatic predictions from my cross-community classifiers with human judgement. The socio-technical intervention I developed has been released publicly and relies on a public API I made available as part of this work, offering rich broader impacts to the broader Internet public.

7.4 Developing new and effective approaches to moderation

New AI-based approaches to content moderation must clearly define who is the end-user of the classification labels. For example, will moderators use the system to triage abusive content, or is the goal to automatically remove abusive content on the platform? Current automated solutions are often trained and evaluated in a static manner, only using pre-existing data; whether these solutions are effective upon deployment remains relatively unexplored. Evaluation must go beyond just traditional measures of performance like precision and recall, and instead begin optimizing for metrics like reduction in moderator effort, speed of response, targeted recall for severe types of abuse, user trust and fairness in predictions.

Proactive approaches to abuse detection

Existing computational approaches to handle abusive language are primarily reactive and intervene only after abuse has occurred. A complementary approach is developing *proactive* technologies that prevent the harm from occurring in the first place.

Interventions that occur after a point of escalation may have little positive effect in some circumstances. For example, when two individuals have already begun insulting one another, both have already become upset and must lose face to reconcile [172]. At this point, de-escalation may prevent further abuse but does little for restoring the situation to a constructive dialog [173]. However, interventions that occur *before* the point of abuse can serve to shift the conversation. Recent work has shown that it is possible to predict whether a conversation will become toxic on Wikipedia [174] and whether bullying will occur on Instagram [175]. These predictable abuse trajectories open the door to developing new ML systems for preemptive interventions that directly mitigate harm.

7.4.1 Towards promoting healthy online behavior

The research community has responded primarily by developing technologies to identify certain types of abuse and to facilitate automatic or computer-assisted content moderation. Traditional approaches tend to define healthy behavior as the absence of toxicity, i.e., $P(Healthy) = 1 - P(Toxicity)$. By focusing on detecting and discouraging overt forms of abuse and hate speech, important components of healthy behavior are often overlooked. Specifically, I believe that a comprehensive approach to measuring and addressing healthy online behavior should incorporate techniques to encourage and promote pro-social behavior in a sustainable manner. I encourage researchers to explore ways to quantify healthy behavior in online communities that go beyond traditional definitions.

I argue for a broad re-aligning of our community goals towards promoting healthy behavior, rather than simply eliminating abusive behavior. We would not judge a community to be healthy simply because it managed to eliminate the most overt forms of abuse and harassment. Depending on the goals of the community, we might hope that it fosters healthy behavior like participation , social support , and information exchange. I believe that this shift in perspective offers challenging and exciting new research directions for promoting well-being and safety on the Internet.

REFERENCES

[1] B. Buffington, *Personal communication*, April 4, 2015.

[2] N. Tiku and C. Newton, "Twitter ceo:we suck at dealing with abuse.i," *The Verge*, February 4, 2015.

[3] N. Lapidot-Lefler and A. Barak, "Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition," *Computers in human behavior*, vol. 28, no. 2, pp. 434–443, 2012.

[4] R. Spears, M. Lea, and S. Lee, "De-individuation and group polarization in computer-mediated communication," *British Journal of Social Psychology*, vol. 29, no. 2, pp. 121–134, 1990.

[5] J. Suler, "The online disinhibition effect," *Cyberpsychology & behavior*, vol. 7, no. 3, pp. 321–326, 2004.

[6] F. B. Viégas, M. Wattenberg, and K. Dave, "Studying cooperation and conflict between authors with history flow visualizations," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2004, pp. 575–582.

[7] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi, "He says, she says: Conflict and coordination in wikipedia," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2007, pp. 453–462.

[8] K. Coe, K. Kenski, and S. A. Rains, "Online and uncivil? patterns and determinants of incivility in newspaper website comments," *Journal of Communication*, vol. 64, no. 4, pp. 658–679, 2014.

[9] A. Hermida and N. Thurman, "A clash of cultures: The integration of user-generated content within professional journalistic frameworks at british newspaper websites," *Journalism practice*, vol. 2, no. 3, pp. 343–356, 2008.

[10] B. Choi, K. Alexander, R. E. Kraut, and J. M. Levine, "Socialization tactics in wikipedia and their effects," in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, ACM, 2010, pp. 107–116.

[11] A. Halfaker, A. Kittur, and J. Riedl, "Don't bite the newbies: How reverts affect the quantity and quality of wikipedia work," in *Proceedings of the 7th international symposium on wikis and open collaboration*, ACM, 2011, pp. 163–172.

[12] A. Halfaker, R. S. Geiger, J. T. Morgan, and J. Riedl, "The rise and decline of an open collaboration system: How wikipedias reaction to popularity is causing its decline," *American Behavioral Scientist*, vol. 57, no. 5, pp. 664–688, 2013.

[13] A. A. Anderson, D. Brossard, D. A. Scheufele, M. A. Xenos, and P. Ladwig, "The "nasty effect:" online incivility and risk perceptions of emerging technologies," *Journal of Computer-Mediated Communication*, vol. 19, no. 3, pp. 373–387, 2014.

[14] S. LaBarre, "Why were shutting off our comments," *Popular Science*, vol. 24, pp. 2013–09, 2013.

[15] N. A. Diakopoulos, "The editor's eye: Curation and comment relevance on the new york times," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ACM, 2015, pp. 1153–1157.

[16] R. Kang, L. Dabbish, and K. Sutton, "Strangers on your phone: Why people use anonymous communication applications," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ACM, 2016, pp. 359–370.

[17] B. Drake, "The darkest side of online harassment: Menacing behavior," *Pew Research Center, http://www.pewresearch.org/fact-tank/2015/06/01/the-darkest-side-of-online-harassment-menacing-behavior/*, 2014.

[18] M. Duggan, "Online harassment: Summary of findings," *Pew Research Center,* `http://www.pewinternet.org/2014/10/22/online-harassment/`, 2014.

[19] S. Herring, K. Job-Sluder, R. Scheckler, and S. Barab, "Searching for safety online: Managing "trolling" in a feminist forum," *The Information Society*, vol. 18, no. 5, pp. 371–384, 2002.

[20] S. Machkovech, "No fooling: Reddit's r/games goes silent for one day to call out hate, bigotry, april 2019," `https://arstechnica.com/gaming/2019/04/no-fooling-reddits-rgames-goes-silent-for-one-day-to-call-out-hate/`, 2014.

[21] E. Chandrasekharan, M. Samory, S. Jhaver, H. Charvat, A. Bruckman, C. Lampe, J. Eisenstein, and E. 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*, vol. 2, no. CSCW, p. 32, 2018.

[22] T. Gillespie, "Governance of and by platforms," *Sage handbook of social media. London: Sage*, 2017.

[23] S. Kiesler, R. Kraut, P. Resnick, and A. Kittur, "Regulating behavior in online communities," *Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA*, pp. 125–178, 2012.

[24] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, "Design lessons from the fastest q&a site in the west," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2011, pp. 2857–2866.

[25] A. Schlesinger, E. Chandrasekharan, C. A. Masden, A. S. Bruckman, W. K. Edwards, and R. E. Grinter, "Situated anonymity: Impacts of anonymity, ephemerality, and hyper-locality on social media," in *Proceedings of the 2017 CHI conference on human factors in computing systems*, ACM, 2017, pp. 6912–6924.

[26] S. T. Roberts, "Commercial content moderation: Digital laborers' dirty work," 2016.

[27] J. Preece and D. Maloney-Krichmar, "Online communities: Focusing on sociability and usability," *Handbook of human-computer interaction*, pp. 596–620, 2003.

[28] R. L. Williams and J. Cothrel, "Four smart ways to run online communities," *MIT Sloan Management Review*, vol. 41, no. 4, p. 81, 2000.

[29] S. T. Roberts, "Behind the screen: The hidden digital labor of commercial content moderation," PhD thesis, University of Illinois at Urbana-Champaign, 2014.

[30] A. Chen, "The laborers who keep dick pics and beheadings out of your facebook feed, october 2014," *https://www.wired.com/2014/10/content-moderation/*, 2014.

[31] C. Buni and S. Chemaly, "The secret rules of the internet, apr. 2016," *http://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech*, 2016.

[32] C. Lampe and P. Resnick, "Slash (dot) and burn: Distributed moderation in a large online conversation space," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2004, pp. 543–550.

[33] M. Bickert, "Publishing our internal enforcement guidelines and expanding our appeals process, apr. 2018," *https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/*, 2018.

[34] Google, "Youtube community guidelines enforcement in google's tranparency report for 2018," *https://transparencyreport.google.com/youtube-policy/removals*, 2018.

[35] T. P. Policy, "Evolving our twitter transparency report: Expanded data and insights, december 2018," *https://blog.twitter.com/official/en_us/topics/company/2018/evolving-our-twitter-transparency-report.html*, 2018.

[36] R. Krishna, "Tumblr launched an algorithm to flag porn and so far it's just caused chaos, dec 2018," *https://www.buzzfeednews.com/article/krishrach/tumblr-porn-algorithm-ban*, 2018.

[37] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert, "The bag of communities: Identifying abusive behavior online with preexisting internet data," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, 2017.

[38] L. Lessig, *Code and other laws of cyberspace*. Basic books New York, 1999, vol. 3.

[39] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," *Proc. ACM Hum.-Comput. Interact.*, vol. 1, no. CSCW, 31:1–31:22, Dec. 2017.

[40] M. S. Bernstein, A. Monroy-Hernández, D. Harry, P. André, K. Panovich, and G. G. Vargas, "4chan and/b: An analysis of anonymity and ephemerality in a large online community.," in *ICWSM*, 2011, pp. 50–57.

[41] J. A. Pater, Y. Nadji, E. D. Mynatt, and A. S. Bruckman, "Just awful enough: The functional dysfunction of the something awful forums," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, ACM, 2014, pp. 2407–2410.

[42] A. M. (VP News Feed), "Addressing Hoaxes and Fake News," *Facebook Newsroom*, December 15, 2016.

[43] E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, ACM, 1999, pp. 159–166.

[44] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[45] S. H. Appelbaum, "Socio-technical systems theory: An intervention strategy for organizational development," *Management decision*, vol. 35, no. 6, pp. 452–463, 1997.

[46] E. Mumford, "Socio-technical design: An unfulfilled promise or a future opportunity?" In *Organizational and social perspectives on information technology*, Springer, 2000, pp. 33–46.

[47] R. Kling, R. Lamb, *et al.*, "It and organizational change in digital economies: A sociotechnical approach," *Understanding the Digital Economy. Data, Tools, and Research. The MIT Press, Cambridge, MA*, 2000.

[48] M. S. Ackerman, "The intellectual challenge of cscw: The gap between social requirements and technical feasibility," *Human–Computer Interaction*, vol. 15, no. 2-3, pp. 179–203, 2000.

[49] Z. Papacharissi, "Democracy online: Civility, politeness, and the democratic potential of online political discussion groups," *New media & society*, vol. 6, no. 2, pp. 259–283, 2004.

[50] J. Dibbell, "A rape in cyberspace or how an evil clown, a haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society," *Ann. Surv. Am. L.*, p. 471, 1994.

[51] J. Preece, "Etiquette online: From nice to necessary," *Communications of the ACM*, vol. 47, no. 4, pp. 56–61, 2004.

[52] N. Z. Shapiro and R. H. Anderson, *Toward an Ethics and Etiquette for Electronic Mail.* ERIC, 1985.

[53] R. Fredheim, A. Moore, and J. Naughton, "Anonymity and online commenting: The broken windows effect and the end of drive-by commenting," in *Proceedings of the ACM Web Science Conference*, ACM, 2015, p. 11.

[54] C. Fiesler, J. A. Jiang, J. McCann, K. Frye, and J. R. Brubaker, "Reddit rules! characterizing an ecosystem of governance," in *Proceedings of the 2018 AAAI International Conference on Web and Social Media*, ICWSM, 2018.

[55] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying.," in *The Social Mobile Web*, 2011, pp. 11–17.

[56] J.-M. Xu, B. Burchfiel, X. Zhu, and A. Bellmore, "An examination of regret in bullying tweets.," in *HLT-NAACL*, 2013, pp. 697–702.

[57] S. O. Sood, E. F. Churchill, and J. Antin, "Automatic identification of personal insults on social news sites," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 2, pp. 270–285, 2012.

[58] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[59] S. Chancellor, Z. J. Lin, and M. De Choudhury, "this post will just get taken downi: Characterizing removed pro-eating disorder social media content," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, 2016, pp. 1157–1162.

[60] J. Seering, R. Kraut, and L. Dabbish, "Shaping pro and anti-social behavior on twitch through moderation and example-setting," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ACM, 2017, pp. 111–125.

[61] S. Chancellor, J. A. Pater, T. Clear, E. Gilbert, and M. 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*, ACM, 2016, pp. 1201–1213.

[62] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "How community feedback shapes user behavior," in *Eighth International AAAI Conference on Web and Social Media*, 2014.

[63] S. Jhaver, L. Chan, and A. Bruckman, "The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action," *First Monday*, vol. 23, no. 2, 2018.

[64] S. Jhaver, S. Ghoshal, A. Bruckman, and E. Gilbert, "Online harassment and content moderation: The case of blocklists," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 25, no. 2, p. 12, 2018.

[65] R. B. Cialdini and M. R. Trost, "Social influence: Social norms, conformity and compliance.," 1998.

[66] R. B. Cialdini, R. R. Reno, and C. 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*, vol. 58, no. 6, p. 1015, 1990.

[67] R. B. Cialdini, C. A. Kallgren, and R. R. Reno, "A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior," in *Advances in experimental social psychology*, vol. 24, Elsevier, 1991, pp. 201–234.

[68] H. C. Triandis, "Culture and social behavior," 1994.

[69] C. Civility, "A natural experiment examining the effects of distributed moderation in online forums/c. lampe, p. zube, j. lee et al," *Government Information Quarterly*, vol. 31, no. 2, pp. 317–326, 2014.

[70] C. D. Van Blarcum, "Internet hate speech: The european framework and the emerging american haven," *Wash. & Lee L. Rev.*, vol. 62, p. 781, 2005.

[71] E. Bleich, "Freedom of expression versus racist hate speech: Explaining differences between high court regulations in the usa and europe," *Journal of Ethnic and Migration Studies*, vol. 40, no. 2, pp. 283–300, 2014.

[72] D. Bamman, B. O'Connor, and N. Smith, "Censorship and deletion practices in chinese social media," *First Monday*, vol. 17, no. 3, 2012.

[73] C. Hiruncharoenvate, Z. Lin, and E. Gilbert, "Algorithmically bypassing censorship on sina weibo with nondeterministic homophone substitutions," in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[74] T. Quynh, "Dissecting facebooks censorship,"

[75] K. Langvardt, "Regulating online content moderation," *Geo. LJ*, vol. 106, p. 1353, 2017.

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

[77] J. C. York and E. Zuckerman, "6 moderating the public sphere," *Human Rights in the Age of Platforms*, p. 137, 2019.

[78] J. Rachels, *The elements of moral philosophy new york*, 2007.

[79] W. G. Sumner, *Folkways-A study of the sociological importance of usages, manners, customs, mores and morals*. Read Books Ltd, 2011.

[80] M. J. Quinn, *Ethics for the information age*. Pearson, 2017.

[81] K. Klonick, "Does facebooks oversight board finally solve the problem of online speech?" *Models for Platform Governance*, p. 51, 2019.

[82] R. Price, "Facebook's new oversight board will help with content moderation and check mark zuckerberg's power," *Business Insider*, September 18, 2019.

[83] U. Senate, "Stop enabling sex traffickers act of 2017," 2017.

[84] H. J. Committee *et al.*, "Allow states and victims to fight online sex trafficking act of 2017," 2017.

[85] D. S. Ardia, "Free speech savior or shield for scoundrels: An empirical study of intermediary immunity under section 230 of the communications decency act," *Loy. LAL Rev.*, vol. 43, p. 373, 2009.

[86] A. M. Sevanian, "Section 230 of the communications decency act: A good samaritan law without the requirement of acting as a good samaritan," *UCLA Ent. L. Rev.*, vol. 21, p. 121, 2014.

[87] A. Romano, "A new law intended to curb sex trafficking threatens the future of the internet as we know it," *Vox. com*, vol. 2, 2018.

[88] B. Plackett, "Unpaid and abused: Moderators speak out against reddit, aug 2018," *https://www.engadget.com/2018/08/31/reddit-moderators-speak-out/*, 2018.

[89] C. Fiesler, J. McCann, K. Frye, J. R. Brubaker, *et al.*, "Reddit rules! characterizing an ecosystem of governance," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[90] E. Rader and R. Gray, "Understanding user beliefs about algorithmic curation in the facebook news feed," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, ACM, 2015, pp. 173–182.

[91] C. Li, Y. Lu, Q. Mei, D. Wang, and S. Pandey, "Click-through prediction for advertising in twitter timeline," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2015, pp. 1959–1968.

[92] J. Seering, T. Wang, J. Yoon, and G. Kaufman, "Moderator engagement and community development in the age of algorithms," *New Media & Society*, 2019.

[93] R. Peck, "The punishing ecstasy of being a reddit moderator, mar. 2019," *https://www.wired.com/story/the-punishing-ecstasy-of-being-a-reddit-moderator/*, 2019.

[94] R. Farmer and B. Glass, *Building web reputation systems*. " O'Reilly Media, Inc.", 2010, pp. 243–276.

[95] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman, "Reputation systems," *Communications of the ACM*, vol. 43, no. 12, pp. 45–48, 2000.

[96] P. Resnick and R. Zeckhauser, "Trust among strangers in internet transactions: Empirical analysis of ebays reputation system," *The Economics of the Internet and E-commerce*, vol. 11, no. 2, pp. 23–25, 2002.

[97] J. Beckett, "We need to talk about the mental health of content moderators, september 2018," *http://theconversation.com/we-need-to-talk-about-the-mental-health-of-content-moderators-103830*, 2018.

[98] K. Saha, E. Chandrasekharan, and M. De Choudhury, "Prevalence and psychological effects of hateful speech in online college communities," in *Proceedings of the 11th ACM Conference on Web Science*, 2019.

[99] D. D. Wray, "The companies cleaning the deepest, darkest parts of social media, june 2018," *https://www.vice.com/en_us/article/ywe7gb/the-companies-cleaning-the-deepest-darkest-parts-of-social-media*, 2018.

[100] R. S. Geiger, "Bot-based collective blocklists in twitter: The counterpublic moderation of harassment in a networked public space," *Information, Communication & Society*, vol. 19, no. 6, pp. 787–803, 2016.

[101] "Perspective api," *https://conversationai.github.io/*,

[102] J. Im, E. Chandrasekharan, J. Sargent, P. Lighthammer, T. Denby, A. Bhargava, L. Hemphill, D. Jurgens, and E. Gilbert, "Still out there: Modeling and identifying russian troll accounts on twitter," *arXiv preprint arXiv:1901.11162*, 2019.

[103] H. M. Team, July 16, 2015.

[104] S. Sood, J. Antin, and E. Churchill, "Profanity use in online communities," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2012, pp. 1481–1490.

[105] S. Brody and N. Diakopoulos, "Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: Using word lengthening to detect sentiment in microblogs," in *Proceedings of the conference on empirical methods in natural language processing*, Association for Computational Linguistics, 2011, pp. 562–570.

[106] A. Sukumaran, S. Vezich, M. McHugh, and C. Nass, "Normative influences on thoughtful online participation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2011, pp. 3401–3410.

[107] A. Kasunic and G. Kaufman, "" at least the pizzas you make are hot": Norms, values, and abrasive humor on the subreddit r/roastme," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[108] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2016, pp. 145–153.

[109] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2017, pp. 1391–1399.

[110] J. Im, S. Tandon, E. Chandrasekharan, T. Denby, and E. Gilbert, "Synthesized social signals: Computationally-derived social signals from account histories," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ACM, 2020.

[111] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "Deep learning for user comment moderation," *arXiv preprint arXiv:1705.09993*, 2017.

[112] D. Jurgens, L. Hemphill, and E. Chandrasekharan, "A just and comprehensive strategy for using nlp to address online abuse," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3658–3666.

[113] L. Blackwell, J. Dimond, S. Schoenebeck, and C. Lampe, "Classification and its consequences for online harassment: Design insights from heartmob," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, p. 24, 2017.

[114] K. Mahar, A. X. Zhang, and D. Karger, "Squadbox: A tool to combat email harassment using friendsourced moderation," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, 2018, p. 586.

[115] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, "A computational approach to politeness with application to social factors," in *Proc. ACL'13*, 2013.

[116] E. Chandrasekharan and S. Chakraborti, "Footprints on silicon: Explorations in gathering autobiographical content.," *Int. J. Comput. Linguistics Appl.*, vol. 6, no. 2, pp. 29–42, 2015.

[117] N. Cao, C. Shi, S. Lin, J. Lu, Y.-R. Lin, and C.-Y. Lin, "Targetvue: Visual analysis of anomalous user behaviors in online communication systems," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 22, no. 1, pp. 280–289, 2016.

[118] R. Baeza-Yates, B. Ribeiro-Neto, *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[119] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Elsevier, 2011.

[120] C. D. Manning, P. Raghavan, H. Schütze, *et al.*, *Introduction to information retrieval*, 1. Cambridge university press Cambridge, 2008, vol. 1.

[121] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[122] P. Shuman, *Fox 11 investigates: anonymous*, `https://www.youtube.com/watch?v=DNO6G4ApJQY`, Jul 2007.

[123] F. News, *4chan: The rude, raunchy underbelly of the internet*, `http://www.foxnews.com/story/2009/04/08/4chan-rude-raunchy-underbelly-internet.html`, Apr. 2009.

[124] R. Sorgatz, *Macroanonymous is the new microfamous*, `http://fimoculous.com/archive/post-5738.cfm`, 2009.

[125] S. Biddle, "Reddit (finally) bans coontown," *Gawker*, August, 5, 2015.

[126] A. Robertson, "Reddit bans 'fat people hat' and other subreddits under new harassment rules," *The Verge*, June, 10, 2015.

[127] Reddit, "Removing harassing subreddits (self announcement)," June, 10, 2015.

[128] J. W. Moyer, "coontown': A noxious, racist corner of reddit survives recent purge," *The Washington Post*, July, 17, 2015.

[129] L. Hockenson, "What is voat, the site reddit users are flocking to?," July, 9, 2015.

[130] E. Newell, D. Jurgens, H. M. Saleem, H. Vala, J. Sassine, C. Armstrong, and D. Ruths, "User migration in online social networks: A case study on reddit during a period of community unrest," in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[131] L. Silva, L. Goel, and E. Mousavidin, "Exploring the dynamics of blog communities: The case of metafilter," *Information Systems Journal*, vol. 19, no. 1, pp. 55–81, 2009.

[132] Reddit, *Remember the human*, https://www.reddit.com/r/blog/comments/1ytp7q/remember_the_human/, 2014.

[133] ——, *Reddiquette*, https://www.reddit.com/wiki/reddiquette, 2015.

[134] ——, *Subreddit rules*, https://www.reddit.com/r/AskHistorians/wiki/rules, 2015.

[135] N. Shuyo, "Language detection library for java," 2010.

[136] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[137] S. O. Sood, J. Antin, and E. F. Churchill, "Using crowdsourcing to improve profanity detection.," in *AAAI Spring Symposium: Wisdom of the Crowd*, 2012.

[138] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[139] H. Daumé III, "Frustratingly easy domain adaptation," *arXiv preprint arXiv:0907.1815*, 2009.

[140] H. Daume III and D. Marcu, "Domain adaptation for statistical classifiers," *Journal of Artificial Intelligence Research*, pp. 101–126, 2006.

[141] N. Syed and B. Smith, *A first amendment for social platforms*, https://medium.com/@BuzzFeed/a-first-amendment-for-social-platforms-202c0eab7054, Jun. 2015.

[142] R. González-Ibánez, S. Muresan, and N. Wacholder, "Identifying sarcasm in twitter: A closer look," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Association for Computational Linguistics, 2011, pp. 581–586.

[143] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation.," in *EMNLP*, vol. 13, 2013, pp. 704–714.

[144] E. Ostrom, *Governing the commons*. Cambridge university press, 2015.

[145] S. L. Bryant, A. Forte, and A. Bruckman, "Becoming wikipedian: Transformation of participation in a collaborative online encyclopedia," in *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, ACM, 2005, pp. 1–10.

[146] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[147] A. Joulin, E. Grave, P. Bojanowski, and T. 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*, vol. 2, 2017, pp. 427–431.

[148] E. Gilbert, "Widespread underprovision on reddit," in *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM, 2013, pp. 803–808.

[149] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[150] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal component analysis*, Springer, 1986, pp. 115–128.

[151] R Lletı, M. C. Ortiz, L. A. Sarabia, and M. S. Sánchez, "Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes," *Analytica Chimica Acta*, vol. 515, no. 1, pp. 87–100, 2004.

[152] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[153] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[154] L. Shifman, *Memes in digital culture*. Mit Press, 2014.

[155] K. Long, J. Vines, S. Sutton, P. Brooker, T. Feltwell, B. Kirman, J. Barnett, and S. Lawson, "Could you define that in bot terms?: Requesting, creating and using bots on reddit," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, 2017, pp. 3488–3500.

[156] B. Boe, *Python reddit api wrapper (praw)*, 2016.

[157] S. Jhaver, I. Birman, E. Gilbert, and A. Bruckman, "Human-machine collaboration for content regulation: The case of reddit automoderator," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 26, no. 5, p. 31, 2019.

[158] L. Garton, C. Haythornthwaite, and B. Wellman, "Studying online social networks," *Journal of computer-mediated communication*, vol. 3, no. 1, JCMC313, 1997.

[159] E. Chandrasekharan and E. Gilbert, "Hybrid approaches to detect comments violating macro norms on reddit," *arXiv preprint arXiv:1904.03596*, 2019.

[160] M. Eslami, "Understanding and designing around users' interaction with hidden algorithms in sociotechnical systems," in *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ACM, 2017, pp. 57–60.

[161] N. Diakopoulos, "Algorithmic accountability: Journalistic investigation of computational power structures," *Digital journalism*, vol. 3, no. 3, pp. 398–415, 2015.

[162] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE symposium on security and privacy (SP)*, IEEE, 2016, pp. 598–617.

[163] T. Zarsky, "The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making," *Science, Technology, & Human Values*, vol. 41, no. 1, pp. 118–132, 2016.

[164] S. B. Sawyer and M. H. Jarrahi, "Sociotechnical approaches to the study of information systems," in *Computing handbook, third edition: Information systems and information technology*, CRC Press, 2014, pp. 5–1.

[165] W. E. Bijker, "The social construction of bakelite: Toward a theory of invention," *The social construction of technological systems: New directions in the sociology and history of technology*, pp. 159–187, 1987.

[166] J. Law and J. Hassard, "Actor network theory and after," 1999.

[167] S. Kwon, P. Liang, S. Tandon, J. Berman, P.-j. Chang, and E. Gilbert, "Tweety holmes: A browser extension for abusive twitter profile detection," in *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ACM, 2018, pp. 17–20.

[168] Z. Ashktorab and J. Vitak, "Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, 2016, pp. 3895–3905.

[169] A. X. Zhang and J. Cranshaw, "Making sense of group chat through collaborative tagging and summarization," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, p. 196, 2018.

[170] E. Chandrasekharan, C. Gandhi, M. W. Mustelier, and E. Gilbert, "Crossmod: A cross-community learning-based system to assist reddit moderators," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–30, 2019.

[171] L. Gao and R. Huang, "Detecting online hate speech using context aware models," in *Proceedings of RANLP*, 2017.

[172] J. Z. Rubin, D. G. Pruitt, and S. H. Kim, *Social conflict: Escalation, stalemate, and settlement*. Mcgraw-Hill Book Company, 1994.

[173] J. M. Gottman, *The marriage clinic: A scientifically-based marital therapy*. WW Norton & Company, 1999.

[174] J. Zhang, J. P. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, N. Thain, and D. Taraborelli, "Conversations gone awry: Detecting early signs of conversational failure," in *Proceedings of ACL*, 2018.

[175] P. Liu, J. Guberman, L. Hemphill, and A. Culotta, "Forecasting the presence and intensity of hostility on instagram using linguistic and social features," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.