Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)

Perceptions of Censorship and Moderation Bias in Political Debate Forums

Qinlan Shen, Michael Miller Yoder, Yohan Jo, Carolyn P. Rosé
Language Technologies Institute, Carnegie Mellon University
{qinlans, yoder, yohanj, cprose}@cs.cmu.edu

Abstract

Moderators are believed to play a crucial role in ensuring the quality of discussion in online political debate forums. The line between moderation and illegitimate censorship, where certain views or individuals are unfairly suppressed, however, is often difficult to define. To better understand the relationship between moderation and censorship, we investigate whether users' perception of moderator bias is supported by how moderators act, using the Big Issues Debate (BID) group on Ravelry as our platform of study. We present our method for measuring bias while taking into account the posting behavior of a user, then apply our method to investigate whether moderators make decisions biased against viewpoints that they may have the incentive to suppress. We find evidence to suggest that while moderators may make decisions biased against individuals with unpopular viewpoints, the effect of this bias is small and often overblown by the users experiencing bias. We argue that the perception of bias by itself is an issue in online political discussions and suggest technological interventions to counteract the discrepancy between perceived and actual censorship in moderation.

Introduction

Online discussion forums create space for communities with similar interests to share thoughts and debate issues. However, the technological facilitation of conversation on these forums does not ensure that high-quality deliberation takes place. Discussion forums are vulnerable to problems such as trolling, flaming, and other types of nonconstructive content (Pfaffenberger 2003). Furthermore, when the topic is controversial, such as religion or politics, discussions can become toxic or inflammatory. Perceived anonymity in many online forums often exacerbates this problem by weakening self-censorship, as people are less likely to regulate their own behavior if they believe that it is difficult to trace back what they say (Chadwick 2006; Davis 1999).

To address these issues, online political discussion forums often rely on moderators to enforce rules and boundaries for how users behave and what they can say. However, the line between legitimate forms of regulation, which are used to discourage behavior defined as inappropriate, and *illegitimate censorship*, where particular individuals, opinions, or forms of communication are unfairly suppressed, is often difficult to define (Wright 2006). Censorship is usually defined subjectively, and in cases where there is room for interpretation, the unconscious biases of regulators may affect their judgments. On the other hand, a user's own bias may lead them to perceive unfair treatment where there is none.

In this paper, we contribute new insight into the differences between perceived and actual bias in an online community's attempt to facilitate productive exchange on controversial issues. Fair moderation without illegitimate censorship is fundamental for creating safe, engaging online spaces for deliberation on controversial topics (Carter 1998). Research in this area not only can improve the quality of discussion in online political forums but also can allow insight into the process of developing norms of behavior and effective moderation in online communities. Regardless of whether censorship actually takes place, the perception of illegitimate censorship itself can create an atmosphere where users feel unfairly treated and trust in the forum is undermined (Wright 2006). Thus, it is important to understand the sources of perceived censorship and recognize when and how perceived censorship is actually manifested.

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

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

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

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

To address these questions, we examined the perception of moderation bias against users with unpopular viewpoints in the Big Issues Debate forum on Ravelry. Using a probabilistic graphical model to identify speech acts, we identified high-risk behaviors associated with rule-breaking, then examined the effect of viewpoint on the likelihood of moderation, controlling for high-risk behavior. This allows us to investigate whether users with minority viewpoints are being unfairly moderated, given the behaviors they exhibit. We find that moderators are significantly more likely to moderate posts from users that hold unpopular viewpoints, though the effect size of this bias is small. While this supports the perception of minority-view users that the moderation is un-

fair, we argue that the perception of bias within the group is an issue by itself, as the perception of illegitimate censorship can lead to tension between the moderators and users within a community.

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

Moderation Issues in Political Discussion

Moderators play an important role in many online forums by helping to maintain order and facilitate discussion within their community (Kittur, Pendleton, and Kraut 2009; Lindsay et al. 2009). While conventional wisdom suggests that moderators positively influence the quality of discussion in forums (Hron and Friedrich 2003), the role of a moderator is often diverse (Maloney-Krichmar and Preece 2005), unclear (Wright 2006), or emergent (Huh 2015) across different communities. Thus, it is important to consider how moderators operate within the context of the community that they are trying to maintain. In online political forums, moderators are considered critical in ensuring quality discussions by creating and enforcing regulations for proper behavior (Edwards 2002), as useful debates require that participants maintain order, respect, and civility towards each other (Carter 1998; Wilhelm 2000).

However, when these political discussions are facilitated by interested groups, moderation can quickly be labeled as censorship. These claims are common on online political forums administered by national governments, a focus of research on the potential for new forms of deliberative democracy (Wright and Street 2007; Khatib, Dutton, and Thelwall 2012). Wright (2006) reviews the process for moderation in two of the UK government's online political discussion forums. They find that moderation must be done carefully to avoid the "shadow of control", the perception that some entity of power can control what is said (Edwards 2002). Ideally, rules for censorship must be detailed, openly available, and enforced by an independent party (Wright 2006). Moderation should also be done in a way that explicitly facilitates the goals of the forum.

In non-governmental political discussion forums, the concept of a "shadow of control" is less obvious, as these forums are not explicitly run by a centralized entity with particular goals. Nevertheless, unconscious cognitive biases may arise from the structural organization of political discussion forums and from cognitive tendencies. Bazerman et al. (2002), in their investigation into why accountants make biased decisions, noted that ambiguity in interpreting information gave accountants the room to make self-serving decisions. In the context of political discussions, ambiguity in the rules for how to engage appropriately in a debate may allow moderators to make unfair decisions against particularly troublesome users or viewpoints they disagree with. Another (more surprising) condition that often promotes unconscious cognitive biases is the belief in one's personal impartiality (Kaatz, Gutierrez, and Carnes 2014). While moderators are expected to act impartially, as they are often removed from debate, they may unconsciously make more biased decisions because they are primed to believe that they are genuinely impartial, instead of recognizing these biases.

In the following section, we describe our platform of study, the Big Issues Debate group on Ravelry, and discuss the organizational elements that make it prone to perceived and actual unconscious biases.

Ravelry and Big Issues Debate

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

Big Issues Debate

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

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

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

Debates on BID begin with a user creating a thread and posting their view on an issue. Other users post responses to the original user's post or to other posts in the thread. An example BID post is given in Figure 1. Every post in the thread, including the original posts, has a set of six associated *tags* (Figure 1, A) that users can interact with: *educational*, *interesting*, *funny*, *agree*, *disagree*, and *love*. Clicking on one of the tags allows a user to anonymously increase the value of a particular tag once per post, though these values do not affect the order in which posts are displayed.

There are three officially recognized and regulated formats of debate on BID: *Order* (default debate format), *Rigor* (stronger standards for sourcing/citations), and *BID* (discussion about policies and practices on BID). Thread creators can choose which format they want their debate to be in by

[1] https://www.ravelry.com/statistics/users

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

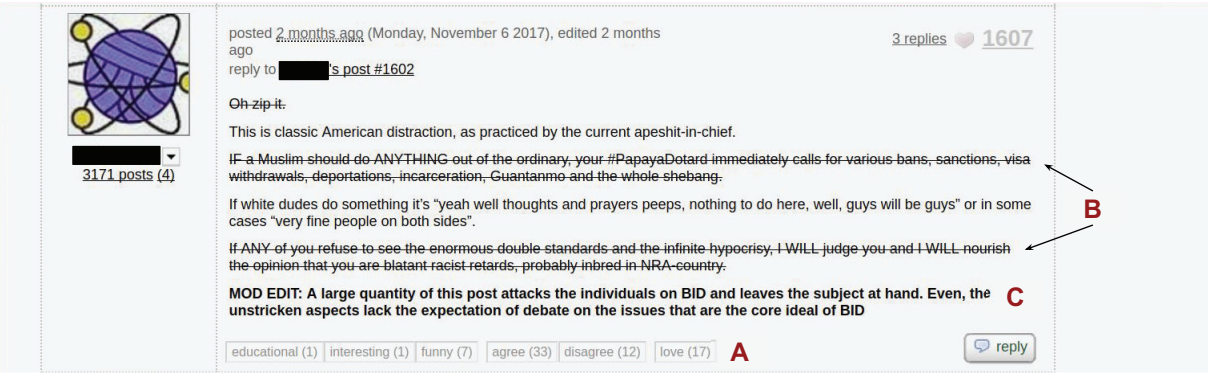

Figure 1: Example of a BID post that was also moderated. (A) shows the *tags* associated with the post. The text of the post that was crossed out (B) was not crossed out by the original poster but by the moderators after judging the text as a violation of the rules of BID. (C) gives the moderators' reasoning for how the post violates the rules of BID. Note that although the post was moderated, more users in the group *agree* with the post than *disagree*.

tagging it in the thread title (e.g. "ORDER - Media Responsibility in Politics", "RIGOR: Bigotry and the 2016 US presidential race"). If not tagged, the thread is assumed to be in the Order format. In all of the recognized formats on BID, users are expected to follow these rules:

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

Within a discussion thread, users can flag another user as being in violation of one of the 6 main rules. Whether or not a post is flagged is only public to the moderation team, the user who made the flag, and the user who received the flag. Moderators then judge whether flagged posts are in violation of the BID rules. If the post is judged to be in violation of the rules, it is hereinafter referred to as *moderated*. In almost all cases, moderated posts are kept visible, but the offending part of the post is crossed out with a strikethrough (Figure 1, B). Moderators are also expected to give reasons for why a post was moderated (Figure 1, C), though they do not post their username. Users who repeatedly make offensive posts may have posting privileges suspended for a period of 24 hours or banned from the group for a longer period of time based on severity of the offense. Moderators may also delete posts, but this is only practiced in the Ask the Mods thread (where only specific types of posts are allowed) or in cases of "extreme spam" [3].

One key limitation on moderator privileges is that moderators cannot participate in debate threads they moderate, which prevents moderators from making explicit decisions against users they are debating.

Issues with Moderation

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

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

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

> *The one thing we can say with some certainty is that a lot of conservative voices have come forward saying they're not being treated fairly. I don't think that's true, but then I wouldn't, would I?*

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

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

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

I think there are people who draw a moderation when others wouldn't. I don't think it has anything to do with political leanings. It's embarrassingly apparent at times.

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

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

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

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

If my posts were deleted why not XXXs?.

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

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

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

Contrasting Views of Bias

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

Method

To assess whether the moderation team is actually making biased decisions based on the viewpoints of users, we present an approach for evaluating moderator decisions alongside users' actual behavior in posts considered for moderation. In order to determine whether or not user viewpoint plays a role in moderation decisions, we need to characterize viewpoints on BID. We also need to identify the behaviors that may put a user at risk of being moderated, as certain types of users may contribute offensive content more often. If users of a certain group more often behave inappropriately, they may be deserving of more moderation. After operationalizing these relevant variables of viewpoint and behavior, we include them in a binary logistic regression model with odds ratios (OR) to predict whether a given post is moderated. This model allows interpretation of the factors that may increase the likelihood that a post would be moderated; odds ratios allows us to estimate the effect of a variable on the probability that the post is moderated.

Dataset

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

Model Specification

Our model is designed to measure the effect of user viewpoint on the likelihood of being moderated. To control for the effect of users' histories with moderation, in addition to our main effect variables indicating viewpoint and high-risk behavior, we include an additional lag variable *mod_prev*. We include this variable to see the extent to which the history of being moderated increases the chance of being unfairly moderated again, as users have argued that moderators tend to repeatedly target the same user for moderation.

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

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

Table 1: Correlation and multi-collinearity checks for main effect variables.

Dependent Variable

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

Independent Variables

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

Correlation and multi-collinearity checks for the main effect variables are found in Table 1.

Assigning Viewpoint

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

The general perception on BID is that right-leaning, conservative users and viewpoints are in the minority while left-leaning, liberal users and viewpoints make up the majority. To verify that the agree and disagree tags align with this liberal-conservative conception of majority-minority on BID, we sampled 20 posts with higher agree than disagree tag values and 20 posts with higher disagree than agree tag values. Posts were sampled across threads to determine the general trend of views on BID on a variety of issues. We then presented the posts, along with the title of the relevant thread and the preceding post in the reply structure as context, to two native English speakers with moderate political knowledge and asked them to separately determine whether the opinion expressed in a post leaned more towards a liberal viewpoint or a conservative viewpoint. We define *liberal* viewpoints as those that favor social progressivism and government action for equal opportunity and *conservative* viewpoints as those that favor limited government, personal responsibility, and traditional values.

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

Assigning viewpoints to users To label the viewpoint of a particular user, we first find every thread they have participated in on BID. For each thread, we sum the agree tag values for each post the user made in that thread. We repeat the same process for the disagree tag values in the same thread. As threads on BID are intended to be centered around a particular issue of debate (e.g. gun control, immigration, tax reform), the summed agree and disagree tag values should indicate how much the other users on BID agree or disagree with the user on that particular issue. If the total disagree tag value is greater than the total agree tag value for a user on a particular thread, we label that user as having the minority viewpoint on the issue discussed in the thread. This thread-level notion of viewpoint is analogous to the *issue-oriented viewpoint* described in Kelly et al. (2005).

However, simply holding a minority view on one thread does not indicate that a user holds the minority viewpoint across BID – users may have particular issues where their viewpoints do not align with the ideological group closest to their general beliefs (e.g. primarily liberal user who is pro-life). Thus, in order to get a general viewpoint for each user, we compare the number of threads where they hold the majority viewpoint with the number of threads where they hold the minority viewpoint. If the number of threads where they hold the minority viewpoint is greater, we label that user as a *minority-view user*. This notion of viewpoint is analogous to the *ideological viewpoints* described in Kelly et al. (2005), which are coherent systems of positions across issues. We focus on ideological viewpoints in our analyses because users participate across threads and recognizably carry their ideological positions with them. This is apparent in BID meta-discussion threads where users will refer to each other with ideological labels (e.g. "conservative", "liberal"). Thus, we predict that moderator impressions of users are based on their activity beyond the level of single-issue threads.

Characterizing Behavior in BID Posts

In the section "Issues with Moderation", we presented evidence that the primary sources of the perception of bias in BID are false negative judgments. Thus, in our analyses, we want to control for the case where users make high-risk, potentially offensive acts in their posts.

In order to identify the types of behavior that are associated with getting moderated, we choose to focus on speech

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

Table 2: Speech acts/foreground topics learned by CSM.

acts within posts. While previous work has characterized offensive behavior using lists of curated terms associated with hate speech or profanity (Chandrasekharan et al. 2017; Hine et al. 2017), we found that this method is unsuited for identifying the types of behavior associated with moderation. First, lists of unacceptable words or phrases will not fully capture more subtle, implicit ways of attacking or offending other users, such as sarcasm or passive aggressive statements. Second, the use of offensive terms is acceptable behavior on BID in certain contexts. Profanity is generally accepted (e.g. "We do not mod for profanity, no matter what people have tried to flag for.", "I have no issues whatsoever with profanity and often sprinkle my posts with it just for my own amusement."), while hateful terms are often quoted or referenced in debates about language use (e.g. "I nearly blew a gasket when my stepmother referred to Obama as 'that nigger in the White House'", "Do you think homosexual people are bullying others when they speak up about people using 'gay' and 'faggot' as insults?").

We instead focus on the intent behind each utterance. The literature on speech acts argues that utterances in discussions function to achieve some conversational goal, called a speech act (Bach and Harnish 1979; Searle 1969). Communicative intents present in discussions and the intents considered to be harmful depend on the norms in the community being examined. Therefore, we use an unsupervised model to capture the speech acts present in BID. Specifically, we use the Content Word Filtering and Speaker Preferences Model (CSM) (Jo et al. 2017), which has been demonstrated to separate the intentions of utterances from their content. CSM identifies dialogue acts in conversation by assuming that the conversation takes place against a backdrop of underlying topics that change more slowly in the conversation than dialogue acts. With the assumption that these two processes have different transition speeds, CSM learns a set of fast-transitioning *foreground topics* that capture dialogue act-related words and slower-transitioning *background topics* that capture more content-related words. This property of the model is desirable because we are interested in speech acts uncorrelated with topics being discussed.

Each thread in BID is considered a conversation in CSM, and each post in the thread as an utterance in the conversation. CSM assumes that the given data has a set of sentence-level speech acts, each of which is defined as a probability distribution over words, like traditional topic models. Thus, we segment posts into sentences using *sent_tokenize* from NLTK (Bird, Klein, and Loper 2009). We set the number of sentence-level speech acts to 10, and the number of background topics in the data to 10, as it gave the highest log-likelihood over the data. The number of states (soft clusters of sentence-level speech acts) is set to 5 [4].

Identifying High-Risk Behaviors After running CSM, we identified the learned speech acts most heavily associated with being moderated as our high-risk behaviors. It is difficult to interpret a speech act by examining the words with the highest weights, as is commonly done for topic models, because speech acts are highly associated with function words that reflect the style and intention of a speaker. Thus, we had two native English speakers interpret the learned speech acts for consistency by examining the 10 sentences with the highest weight for each speech act and looking for common themes and trends in user intention. Though this method has limitations, the speech acts were generally consistent between annotators and such interpretation is commonly used for topic models. The interpreted speech acts are displayed in Table 2.

Many of these identified speech acts are expected in a debate forum: speech acts F0 and F1 are typical moves in argumentation, F5 establishes a user's credibility, while F4 could be used to build rapport with other users. Talk classified as F3, F7, or F8 negotiates the reliability of information presented in the debate. On the other hand, speech acts for expressing a personal perspective (F2) and giving short exclamations (F6) are more surprising in the domain of political argumentation, as they are emotional in nature and primarily

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

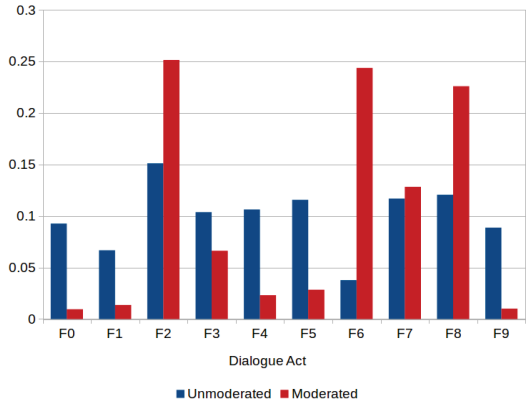

Figure 2: Comparison of the distributions of speech acts between moderated and unmoderated posts. The distribution of speech acts between moderated and unmoderated posts are different with statistical significance $p < 0.001$ by Pearson's chi-square test.

express a user's personal state.

From Figure 2, we see that the three speech acts with the greatest difference between moderated posts and unmoderated posts are F2, F6, and F8. These three topics fit with violations of BID's moderation guidelines. F2, which contains many expressions of personal states and opinions, includes examples of harsh personal judgments that were moderated for being uncivil or attacking (e.g. "I do not for one second think you are trying to hide anything", "You would make a great politician"). F6, which is largely made up of exclamations and short comments, contains many snippy statements that could come off as being uncivil and dismissive to another user. "Le sigh", for example, sarcastically dismisses a previous comment as being beneath the author's attention. As a whole, F2 and F6, which reflect more emotional acts, may be more associated with moderation, as the rules of BID espouse argumentation around the topic and not the users participating in the debate. Probing and evaluating other perspectives (F8) is inherently threatening to other users, and statements where a user immediately dismisses or questions another user's claim without reasoning often violate BID's rule against debating the person and not the topic. Though these particular speech acts and their association with moderation may be specific to the norms of BID, CSM is unsupervised and easily applicable to other domains.

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

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

Table 3: Logistic regression results for whether moderators are biased against users holding minority viewpoints. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Findings

Table 3 shows the findings from our regression on which factors contribute to the likelihood of a post being moderated.

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

On the other hand, the interaction term *high_risk* × *minority* is not significant (OR = 0.915, $p = 0.148$). This means that users with minority viewpoints are moderated more even at the same level of high-risk behaviors as their majority-view counterparts. Figure 3, which shows the predictive margins of majority-view vs. minority-view users at different values of *high_risk* on the probability of a post getting moderated, demonstrates that this is the case.

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

How strong is the effect of viewpoint on moderation? The regression model suggests that posts by users who express minority viewpoints are more likely to be moderated than posts by users who express majority viewpoints. However, the effect size in terms of Cohen's d is 0.1324, suggesting that the effect of a user's viewpoint on whether their

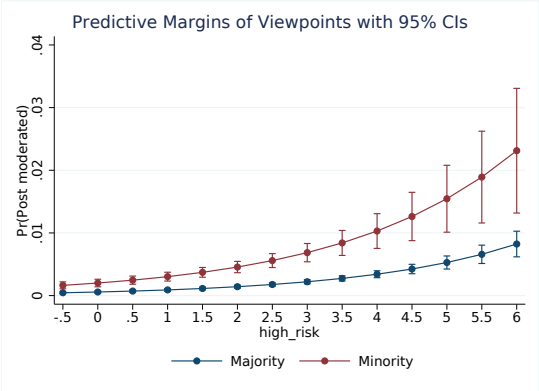

Figure 3: Comparison of predictive margins of minority vs. majority view users over different values of high-risk speech act use (standard deviations from mean) on the probability of a post getting moderated.

posts are moderated is small. Thus, while there is evidence that there is some form of moderation bias against users who express minority views, the impact of bias on these users compared to the level of moderation on BID is negligible.

Discussion

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

Sources of Actual Bias in BID

In the case of BID, moderators can be susceptible to bias against certain viewpoints for a number of reasons. One of the most notable systemic reasons for bias (Bazerman, Loewenstein, and Moore 2002) is ambiguity in how rules and guidelines can be interpreted. Users of BID explicitly raise this issue of rule ambiguity:

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

This type of ambiguity can make moderation susceptible to the cognitive biases of individual moderators (Bazerman, Loewenstein, and Moore 2002) and mask subjectivity in determining who is acting in a "civil" way. When moderators are not aware of these biases and instead believe they are acting objectively, this can make moderation even more biased (Kaatz, Gutierrez, and Carnes 2014).

Specific cognitive biases that could influence moderators to moderate unfairly include the *ecological fallacy*, making assumptions about individuals based on judgments about a group (Kaatz, Gutierrez, and Carnes 2014). In the context of BID, moderators likely recognize users who express conservative viewpoints and make judgments based on that group membership instead of individual behavior. *In-group/out-group bias* (Kaatz, Gutierrez, and Carnes 2014) may also be a factor in moderator bias. Moderators may more easily make negative judgments about users expressing positions that differ from their own group's. Unfortunately, we cannot easily compare the ideological positions of the moderators in BID with the users they judge. Moderators do not give their names with mod edits and the current Ravelry API does not include logs of post edits, so pinpointing the specific moderator who handed down judgment is impossible. Additionally, it is difficult to determine the viewpoints of the moderation team on BID with our current approach for assigning ideology. Though moderators can in theory participate in debate threads they are not moderating, moderators in practice almost never post outside of their moderating duties. This is likely due to the high workload of the moderator role and a previous prohibition against all moderator participation in debate, which some moderators still follow.

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

Sources of Perceived Bias

Ambiguity in the moderator guidelines may also play a role in why users perceive bias against them when they are moderated. Vague rules, such as "Behave civilly" in BID, allow users to make judgments about their behavior in their own self-interest (Bazerman, Loewenstein, and Moore 2002). As it is in their interest not to get moderated, a user may be prone to *blind-spot bias* (Kaatz, Gutierrez, and Carnes 2014) and perceive themselves as being more civil than they actually are. If these users are then moderated, they may be inclined to believe that moderators made an unfair judgment by moderating them for their "civil" behavior. While we saw that most users viewed the main issue of censorship in BID to be false negative judgments, some users do argue that they have been moderated without cause:

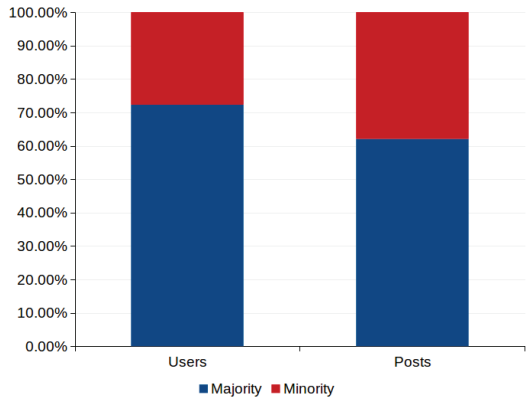

Figure 4: Comparison of viewpoint distributions over users vs. posts. The proportion of majority vs. minority are different between users and posts with statistical significance $p < 0.001$ by Pearson's chi-square test. Note that the distribution of viewpoints over posts is more balanced than the distribution of viewpoints over users.

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

Another possible explanation for the perception of biased moderation from minority-view users in general is that minority users may experience a *halo effect* where their perception of the moderators are shaped by their experiences with other users within the group. Kelly et al. (2005) found that in political Usenet groups, minority-view posts are overrepresented compared to the population of minority-view authors, meaning minority-view users generate more posts per person than majority-view users. We see this same pattern in BID (Figure 4). This pattern suggests that individual minority users must spend more effort on defending their views, as there are fewer people on their side who can help support their arguments. As a result, these minority-view users may feel like they are outnumbered and targeted by majority-view users, who can afford to spend less effort individually. These feelings of unfairness could be transferred to the moderation team, as the moderators are responsible for regulating conversations and maintaining order within the group.

Interventions and Future Work

One way of addressing the image of moderators as biased dictators is to shift both the power and burden of moderation in the group. Studying the political branch of the technology news aggregator Slashdot, Lampe et al. (2014) argue for the success of a distributed moderation system in which users with highly rated comments can become moderators, who in turn are allowed to rate others' comments higher or lower. Along with a "meta-moderation" system that broadly crowdsources the review of moderator actions, they argue that this model can filter out unproductive behaviors as well as develop and pass on community norms. Such a meta-moderation system could not only counter moderator bias, but improve feelings of ownership in the moderation system for users who are not moderators. A danger of these meta-moderation systems that rely on the user base, however, is that minority-view users have fewer protections against the majority. An independent panel of judges may be helpful in protecting minority-view users from the tyranny of the majority, yet these judges should be made aware of their own biases to avoid introducing blind-spot biases (Kaatz, Gutierrez, and Carnes 2014).

Moderators accused of censorship are often criticized for providing little evidence for why a particular post is moderated while others are not. One possible intervention in these cases is an automated system that does not directly classify posts as needing moderation, but instead provides better grounding for the discussions between moderators and those being moderated (Gweon et al. 2005). An example of such a grounding is an automated metric of inflammatory language that also provides comparisons to similar past posts that have been moderated. Making this visible to both the moderators and users could lend greater transparency and objectivity to how moderators operate, though this method would have to be safeguarded against the possibility of reproducing the bias of previous moderation.

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

Conclusion

Moderation in political discussion forums can be controversial, especially when claims of illegitimate censorship of specific views and individuals arise. In this paper, we examined whether perceived unfairness against minority-view conservative users aligns with actual moderation patterns in Ravelry's Big Issues Debate forum. We found that users holding minority views are more likely to be moderated, even after accounting for levels of potentially offensive behaviors across groups. We found, however, that the effect of this bias is much smaller than how the issue is represented. Nevertheless, the perception that there is bias against certain subgroups remains an issue in political forums, as it may lead to tension and conflict over how moderation should be handled. We argue that ambiguity in how guidelines are laid

out allows cognitive biases to slip in, explaining how both actual bias from the moderators and the perception of bias from users arise. We make recommendations for interventions that mitigate these biases by reducing ambiguity and increasing transparency in moderation decisions. While our study focuses primarily on Big Issues Debate, the techniques presented can easily be applied to other political debate forums and it is likely that our findings about the issue of perception of bias are not exclusive to this context.

Acknowledgements

This work was funded by NSF grants IIS 1546393 and DGE1745016 and the Kwanjeong Educational Foundation.

References

Bach, K., and Harnish, R. 1979. *Communication and Speech Acts*. Harvard UP.

Bazerman, M. H.; Loewenstein, G.; and Moore, D. A. 2002. Why good accountants do bad audits. *Harvard Business Review* 80(11):96–103.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Carter, S. L. 1998. *Civility: Manners, morals, and the etiquette of democracy*. Basic Books (AZ).

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

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

Davis, R. 1999. *The web of politics: The Internet's impact on the American political system*. Oxford UP.

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

Gweon, G.; Rosé, C. P.; Wittwer, J.; and Nueckles, M. 2005. Supporting Efficient and Reliable Content Analysis Using Automatic Text Processing Technology. In *Human-Computer Interaction–INTERACT 2005*.

Hine, G. E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM 2017*.

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

Huh, J. 2015. Clinical questions in online health communities: the case of see your doctor threads. In *CSCW 2015*.

Jo, Y.; Yoder, M. M.; Jang, H.; and Rosé, C. P. 2017. Modeling Dialogue Acts with Content Word Filtering and Speaker Preferences. In *EMNLP 2017*.

Kaatz, A.; Gutierrez, B.; and Carnes, M. 2014. Threats to objectivity in peer review: the case of gender. *Trends in pharmacological sciences* 35(8):371–373.

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

Khatib, L.; Dutton, W.; and Thelwall, M. 2012. Public diplomacy 2.0: A case study of the US digital outreach team. *The Middle East Journal* 66(3):453–472.

Kittur, A.; Pendleton, B.; and Kraut, R. E. 2009. Herding the cats: the influence of groups in coordinating peer production. In *WikiSym 2009*.

Lampe, C.; Zube, P.; Lee, J.; Park, C. H.; and Johnston, E. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31(2):317–326.

Lindsay, S.; Smith, S.; Bellaby, P.; and Baker, R. 2009. The health impact of an online heart disease support group: a comparison of moderated versus unmoderated support. *Health education research* 24(4):646–654.

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

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

Searle, J. R. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge UP.

Wilhelm, A. G. 2000. *Democracy in the digital age: Challenges to political life in cyberspace*. Psychology Press.

Wright, S., and Street, J. 2007. Democracy, deliberation and design: the case of online discussion forums. *New media & society* 9(5):849–869.

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