Safe Spaces & Free Speech: Effects of Moderation Policy on Structures of Online Forum Discussions

Anna Gibson
Stanford University
agibson2@stanford.edu

Abstract

Moderation policies of "free speech" and "safe space" have often been equated to low- and high-censorship levels. However, this paper proposes that moderation policies of "safe space" and "free speech" can also be thought of as a design choice that establishes norms of how individuals should treat each other in that discussion space. Analysis of word usage in matched Reddit communities provides evidence that safe spaces do have higher levels of censorship than free speech zones, and, furthermore, moderation also guides standards of politeness, which can be tracked through word frequency analysis.

1. Introduction

The terms "free speech" and "safe space" have become buzzwords in American culture over the last few years, in part due to widely publicized incidents at the University of Missouri and Yale University in 2015 [1] [2]. Both incidents involved student protests about the role of speech on campus; university officials defended their policies of free speech in the name of intellectual rigor, while minority students protested that free speech served simply to maintain a racist status quo. Safe spaces, free from harassing and hateful speech, were held up as important alternatives for minority students. The debate continues, as evidenced by the University of Chicago's welcome letter to its incoming freshmen in the fall of 2016, which expressed the College's "commitment to freedom of inquiry and expression" while refusing to "condone the creation of intellection 'safe spaces'" [3].

Safe spaces are premised on the idea that power relations are inherent within all structures, including speech interactions [4]. In order to prevent the marginalization of voices already hurt by dominant power relations, safe space policies are implemented to prevent exclusion of those groups. This includes a strict no-tolerance policy of "hate speech" or other discussion that would undermine the political project assumed in the space of the community. In practice, this often means that people can be censored or ejected from a space for not properly observing the standards of speech, tone, or style. This includes hateful statements, but also ignorantly prejudiced or unintentionally triggering topics without trigger or content warnings.

Critics allege that such censorship results in echo chambers, intolerant of outside ideas and quick to ban those who disagree with the locally established party line. These critics argue that only a truly free space, in which no one is censored, can engender real, productive conversations.

The debate about free speech and safe space comes to the fore in online discussion forums. Online discussion has been theorized as a new space for discussions of civic importance [5] [6]. Like ancient Athenian forums, online forums can be a public gathering place where citizens can debate issues important to a functioning democracy. The specific features of online discussion, such as design and access, have been the topic of some research [7].

One understudied area of online discussion features is moderation policy. Like other spaces for debate, online forums also use moderators – usually a computer program or a person – to determine the baseline rules of discussion. Moderators play an important role in preventing disruptive users like trolls or spam from taking over forums. However, moderators can just as easily act as a censor of opinions and ideas. Due to how online forums are structured, forum moderators are able to remove users' posts from the forum or even ban users entirely. Consequently, moderators have much more power to affect the discussion in online forums than other users. Therefore, the forum policies established by moderators, and the effects they have on discussion, are important to fully understand.

To date, we know very little about how online discussions guided by moderation policies of self-

URI: http://hdl.handle.net/10125/41439
ISBN: 978-0-9981331-0-2

designated "free speech" zones and "safe spaces" differ. In this paper I use 13,000 comments posted on two comparable discussion communities to quantitatively analyze the effects that these different policies have on censorship, self-other equity, and tone.

2. Theoretical models of moderation

2.1 Moderation as censorship

A basic function of the role of an online moderator is to censor. Moderators have the ability to delete comments and ban users, effectively removing them from the conversation. Free speech moderation policies are based on the idea that little to no censorship of participants is ideal; correspondingly, online free speech policy should show low censorship through low usage of these moderator powers. Safe space policies, on the other hand, reserve the right to protect the political ideas of the space through removing dissenting opinions. Accordingly, online safe spaces should be more willing to use the tools of the moderator to remove counterproductive voices. A common understanding of safe spaces as high moderation and free speech zones as low moderation reflects this common understanding of the moderation-as-censorship model.

2.2 Moderation as establishment of equity rules

Another way to theorize moderation policy is through equity theory, as described by Clark [8]. Policies do not just describe how and when comments will be deleted, but they also lay out the proper way for participants in this forum to show respect for each other.

Equity theory posits that people try to maximize their own outcomes within a socially determined system, but when they find themselves in an inequitable situation, they feel distress. The distress can be explained through Eving Goffman's concept of face: the ideas of self-worth and autonomy within social interactions. Goffman suggests that during an interaction, participants are motivated to maintain both their own and their discussion partner's self-worth and autonomy [9]. For example, if I demean your self-worth by taking one of your belongings without compensating you for it, the situation is also inequitable. This applies to speech as well; if I feel that you are impinging on my autonomy by telling me what to do, the situation also becomes uncomfortably inequitable.

This urge to maintain face is theorized to be universal in social interactions; however, face is defined and maintained through specific social and cultural rituals. Actions that are face-saving in one culture may be deeply offensive in another one. For example, haggling is used in many cultures to maintain the face of both buyer and seller. In the United States, haggling would be offensive by violating norms of autonomy and self-worth.

Because equity theory predicts universal urges towards equity, participants in online interactions should also be motivated to maintain face. However, unlike interactions in physical space, there are often relatively fewer cues for the cultural norms of maintaining equity between participants.

Moderation policies can be theorized to serve the role of cultural indicator in an otherwise culturally non-specific space. Moderation policies, in the context of equity theory, can serve to outline what is considered a violation of face within the context of discussion.

For example, many online discussions require trigger warnings to protect readers from trauma. In these forums, marking comments with trigger warnings maintains face by protecting the self-worth of the poster and respects the autonomy of the reader. However, in communities that regard themselves as free speech zones, trigger warnings may have the opposite effect, and are in themselves offensive. In such spaces, implying that some members would not be able to deal with content lowers the self-worth of those members. As such, marking trigger warnings creates an inequitable situation, which is uncomfortable for participants.

In either situation, the offending party can restore equity by correspondingly lowering their own self-worth through an apology.

Moderation policies can be understood as explicitly defined codes for maintaining the face of all users, i.e. what speech acts constitute a violation of social equity and how the actors must be punished or compensated.

3. Reddit as a Site for Study

3.1 Background

According to Alexa.com, Reddit is currently the 9th-most visited site in the United States, and 30th globally [10]. Reddit's own information page notes that it had over 240 million unique visitors last month from 212 different countries [11].

The site is organized into communities, called subreddits, which users can subscribe to or visit. Users with accounts can post text, links, or images into these subreddits and also comment in response to the posts. Anyone can create an anonymous account without an email address or real name and begin posting and commenting immediately. Redditors can vote posts and comments "up" or "down" which will affect the post or comment's public score and subsequently how easily other redditors and the general public will see that post or comment.

Reddit is an interesting site for study because of its accessibility as well as its organization. Unlike social sites like Twitter, Reddit comments are structured into long discussion threads of users responding to the post or other users. Each subreddit has its own rules and moderators who determine what kind of posts and comments are appropriate in that particular subreddits. Therefore, there is not a uniform policy of moderation across the site. Individual users are subsequently free to seek out and self-select into various subreddits.

The established norms of acceptable discourse in a subreddit are not arbitrary; moderators create and post public rules, and then enforce those policies accordingly. Reddit users can join or leave communities in reaction to those policies. Therefore, it can be posited that moderation policies act as an independent factor in the study of subreddit discourse; users' speech changes according to changes in policy. Studying the differences in discourse between subreddits with differing moderation policies may therefore provide evidence of the effects of those policies on commenter discourse.

Of course, there are many variables that affect discussion within a subreddit: size, topic, and whether users are subscribed to the subreddit by default are some of the most prominent factors. All of these may have much more profound effects on the observed measures of discourse than moderation style. Therefore, any investigation of subreddit discourse will need to account for these confounding variables. To compare the effects of moderation policy, we will need to match subreddits that are alike in almost every way except for moderation policy.

3.2 Subreddits in this study

Very few subreddits can be well matched, as they usually differ not just on moderation policy, but also topic and size. One exemplar pair of matched subreddits is r/lgbt and r/ainbow. Both communities focus on the topic of lesbian, gay, bisexual, trans, and queer issues, and therefore should have the same types of discussion and target audience. They are not politically focused subreddits, and so their discussions have a wide range of topics. Additionally, they are relatively well matched in number of users: r/lgbt has approximately 110 thousand subscribers and r/ainbow has approximately 37 thousand subscribers. To put this in perspective, the thirty most popular subreddits each have upwards of 6 million subscribers, while many have next to none [12].

Well matched in topic and size, these two subreddits differ only in explicit level of moderation. They even link to each other in their descriptions. The moderation policy on r/lgbt specifically describes itself as a safe space:

"This is a safe space. Anyone can make a mistake and accidentally say something hurtful or triggering. If you find yourself corrected for making this error, please try to learn from it. This is not a place to tell people that they need to reclaim a pejorative so you can use it, that they should laugh at jokes about them, or that they otherwise just 'shouldn't be so sensitive.'" [13]

In contrast, r/ainbow describes itself a "free speech zone":

"[C]omments are generally not removed …. This subreddit is a free speech zone …which means that it's up to you the community to downvote offensive posts and comments, and upvote constructive content." [14]

Because these subreddits are so similar in topic and size, but differ explicitly in moderation policy, differences in discourse between the two subreddits can be hypothesized to be a result of that moderation policy.

4. Hypotheses

If the censorship model of moderation policy is correct, we should expect more moderator-removed comments in the safe space community than the free speech community. Reddit's data also allows us to see whether users have deleted their own comments, so we can also compare relative rates of self-censorship between the communities.

H1: There will be more censorship of comments by moderators in the safe space than the free speech zone.

H2: Users will delete their own comments more in the safe space than the free speech zone.

It is slightly more complicated to quantitatively measure whether users have different ways maintaining equity in different discussion forums. In order to do so, I will use LIWC to measure the relative frequency of various words present in the discussions. [15].

Pennebaker describes in his book *The Secret Life of Pronouns* how LIWC was originally used to measure the relative frequency of pronoun usage in depressed patients. Depressed people are more self-focused, and consequently use the word “I” more often than their peers [16]. Subsequent research, described in the same book, has shown that tracking pronoun usage allows researchers to follow the “gaze” of a speaker’s attention. Comparing relative usage of pronouns gives clues about how people regard themselves and others in social situations. For example, a person in a happy relationship uses the word “we” relatively more than a person in an unhappy relationship. In a conversation with someone in a higher status, a lower-status person will use the word “I” relatively more often, suggesting a gaze from down from the higher status person.

Using Pennebaker’s theory of pronouns-as-gaze, the principles of equity outlined in the moderation policies of the free speech and safe space subreddits suggest how pronoun usage between the two subreddits might differ.

In the safe space of r/lgbt, users are encouraged to be highly sensitive to others’ needs and mark trigger warnings when applicable. Participants are urged to think about themselves as individual actors within webs of power and to be highly conscious of their relative positions to others. As the policy reads,

“[d]emonstrate a willingness to learn …. Anyone can make a mistake and accidentally say something hurtful or triggering. If you find yourself corrected for making this error, please try to learn from it. This is not a place to tell people that they need to reclaim a pejorative so you can use it, that they should laugh at jokes about them, or that they otherwise just ‘shouldn't be so sensitive.’” [13]

Consequently, we can theorize that individual identity will be much more salient in the safe space than the free speech zone. Face is maintained through attention and awareness of power dynamics that might devalue someone’s self-worth. We should therefore be able to detect a difference in equity principles by tracking relative usage of the word “I” and “we”.

In safe spaces, referring to “we” rather than “I” will constitute a violation of others’ self worth or autonomy and leave commenters open to corrections by other users. This does not hold true in free speech zones, where the moderation policy does not make such sweeping claims about individual identity. Under a free speech policy, users maintain equity through a more classical understanding of power wherein equal access to expression is paramount. Face is maintained in these communities by respecting the autonomy of individuals to say whatever they would like to say.

H3: There will be relatively more usage of “I” in safe spaces than free speech zones.

H4: There will be relatively more usage of “we” in free speech zones than safe spaces.

Additionally, this should also be borne out through references to marginalized identities. Both r/lgbt and r/ainbow cater to one kind of marginalized identity, LGBTQ. However, across Reddit, females are generally underrepresented; a 2013 Pew poll indicates that American men are twice as likely as women to use the website [17]. For this study, I am assuming that these differences extend to r/lgbt and r/ainbow, and that their demographics are roughly similar. With these assumptions, females would be a minority identity in r/lgbt and r/ainbow.

Within the safe space, users are made much more sensitive to their relative positions with regard to others. This should also extend to minority status, such as being female. Because individual identity is a much more salient part of equity in the safe space than the free speech zone, it follows that marking minority status will be more explicit in the discussions.

H5: The safe space will have more references to females than the free speech community.

Aside from identity, the equity principle also allows us to make predictions about tone in these subreddits. r/lgbt states in its discussion guidelines, “Rule 1**:** No homophobia, bi/panphobia, transphobia, racism, serophobia, or misogyny of any kind” [13]. Aggression, correspondingly, is seen as a violation of equity within this community.

r/ainbow, however, has a much broader guideline: “We encourage you to … engage in robust discussion and interact with the community.” [14] Though the guideline encourages respect, it offers much less rejection of aggression. The word “robust” suggests forceful and spirited discussion. The policy suggests much more acceptance for aggressive remarks in this community.

The features of the discourse in these communities should reflect these different approaches to equity. We should see a higher usage of words indicating anger, including swear words, in the free speech community than the safe space.

H6: The free speech community will have relatively more angry words that the safe space.

H7: The free speech community will have relatively more swear words than the safe space.

5. Method

Reddit's API has made it possible for all data on the site to be aggregated and analyzed. A corpus of all Reddit comments from 2007 to April 2016 is available on Google's BigQuery [18]. This corpus includes data about when and where each comment was posted, whether the user or the moderator deleted it, how many votes it received, and (if the comment was not deleted) the content of the comment.

All of the comments and associated meta-data created in the r/ainbow and r/lgbt subreddits during April 2016 were downloaded from the BigQuery reddit corpus. There were 8,381 comments in the r/lgbt community and 4,619 in r/ainbow, for a total of 13,000 comments in the data set.

LIWC 2015 was used to analyze each comment for frequency of word use [15].

r/lgbt had a mean of 54.2 words per comment and r/ainbow 55.5 words per comment. A t-test was performed to ensure the validity of the comparison, and found no significant difference in mean comment word count, t(13000) = -0.87, *p* = 0.38.

6. Results

H1: There will be more censorship of comments by moderators in the safe space than the free speech zone.

Table 1: Differences in comment removal rates

Subreddit	Removed	Total	% removed
r/lgbt	163	8381	1.9%
r/ainbow	25	4619	0.5%

A chi-square test found a significant difference in removed comments between subreddits, $X^2(1, N = 13000) = 39.18, p < 0.001$. This result suggests that the censorship model of moderation is correct.

H2: Users will delete their own comments more in the safe space than the free speech zone.

Table 2: Differences in comment self-deletion rates

Subreddit	Deleted	Total	% deleted
r/lgbt	466	8381	5.6%
r/ainbow	234	4619	5.1%

A chi-square test found no significant difference in deleted comments between communities, $X^2(1, N = 13000) = 1.19, p = 0.27$. This result suggests that the censorship model of moderation policy does not extend to self-censorship.

H3: There will be relatively more usage of "I" in safe spaces than free speech zones.

H4: There will be relatively more usage of "we" in free speech zones than safe spaces.

H5: The safe space will have more references to females than the free speech community.

Table 3: Differences in mean LIWC scores

LIWC cat.	r/lgbt	r/ainbow	t	*p*
I	4.01	3.41	6.04	< 0.001
we	0.59	0.67	-1.95	0.052
female	0.74	0.58	4.20	< 0.001

There was a significant, relatively higher usage of the word "I" in r/lgbt than r/ainbow, providing support for H3. While there was a relatively higher usage of "we" in r/ainbow than r/lgbt, it was only marginally significant, so we cannot provide sufficient support for H4.

As predicted in H5, there was a significant, higher rate of reference to females in r/lgbt than r/ainbow.

H6: The free speech community will have relatively more angry words that the safe space.

H7: The free speech community will have relatively more swear words than the safe space.

Table 4: Differences in mean LIWC scores

LIWC cat.	r/lgbt	r/ainbow	t	*p*
anger	1.18	1.56	-4.90	< 0.001
swear	0.66	0.94	-4.08	< 0.001

As predicted in H6 and H7, there were significantly more anger-related words and swear words in r/ainbow than r/lgbt per sentence.

To check that angrier word choice was not just an effect of a generally more negative attitude on r/ainbow, scores (the sum of upvotes and downvotes by other Reddit users) of comments in r/lgbt and r/ainbow were compared. The mean score of all comments in r/lgbt was 6.339 and the mean score of all comments in r/ainbow was 6.364. A t-test showed no significant difference between the mean scores, t(13000) = 0.04, p = 0.97.

7. Conclusion

These results suggest that the moderation style of a community does not simply affect which voices will be heard, but in a larger way, how the users will conceive of themselves in relation to the rest of the community.

As predicted, the safe space moderators deleted more comments than the free speech zone moderators. However, the relatively similar rates of self-censorship cannot adequately capture whether users self-censored themselves in the process of deciding whether or not to post in the subreddit.

In addition to censorship, moderation policy has a perceptible impact on how users, in general, interact with each other, by establishing rules for equity. In short, safe space and free speech policies define codes for politeness in online spaces where it may be otherwise difficult to assess what the standards of politeness are.

Moderation, therefore, is an important feature of the architecture of a discussion space. As more and more discussion moves online, moderation will need to be a factor carefully watched and discussed. The subreddits examined here are just two of the countless message boards through which online discussion is happening. It is an open question whether my findings will generalize to other communities on Reddit, much less the rest of the Internet. However, if they do, it would suggest a new theoretical avenue in how free speech and safe space policies are approached.

There are many opportunities to continue this line of research. I have provided here only an incomplete theorization of safe spaces and free speech policies.

Because the moderators of r/lgbt and r/ainbow have describe themselves as safe space and free speech, there is no way to know whether they are truly being enforced in the way they describe, or if their definitions of safe spaces and free speech match more popular conceptions of the terms. A more complete theory of safe space and free speech will need to be able to encompass more than just self-described communities, and perhaps be able to categorize communities through text analysis.

Finally, within the debates about safe space and free speech, there is a huge implicit question that I have not addressed here: which policy leads to better discussions? Although I was able to quantitatively measure several effects of moderation policy, I have not yet determined a way to objectively measure or evaluate discussion quality.

At the end of the day, discussion quality may be the factor that most voices in the debate care the most about, but quality remains in the realm of ethics and philosophy until we can determine how to measure and compare it effectively.

8. References

[1]http://www.thedailybeast.com/articles/2015/11/09/mizzou-protesters-to-media-stay-out-of-our-safe-space-or-we-ll-call-the-cops.html

[2]http://www.theatlantic.com/politics/archive/2015/11/the-new-intolerance-of-student-activism-at-yale/414810/

[3]https://www.insidehighered.com/news/2016/08/25/u-chicago-warns-incoming-students-not-expect-safe-spaces-or-trigger-warnings

[4]http://fusion.net/story/231089/safe-space-history/

[5] Price, V., Nir, L., & Cappella, J. N. (2006). Normative and informational influences in online political discussions. Communication Theory,16(1), 47-74.

[6] Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. New Media & Society,6(2), 259-283.

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

[8] Clark, H. H. (1996). Using language. Cambridge university press.

[9] Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage (Vol. 4). Cambridge university press.

[10]http://www.alexa.com/siteinfo/reddit.com, accessed April 24th, 2016

[11] https://www.reddit.com/about/, accessed April 24th, 2016

[12] http://redditlist.com/all

[13] http://www.reddit.com/r/lgbt

[14] http://www.reddit.com/r/ainbow

[15] Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. UT Faculty/Researcher Works.

[16] Pennebaker, J. W. (2011). The secret life of pronouns: How our words reflect who we are. New York: Bloomsbury.

[17] http://pewinternet.org/Reports/2013/reddit.aspx

[18] https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments

[16] Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage (Vol. 4). Cambridge university press.