Deciding to Delete Posts on Reddit: What Factors Influence Content Removal

Chris Vargo
University of Colorado Boulder,
College of Media, Communication and Information
Christopher.Vargo@colorado.edu

Toby Hopp
University of Colorado Boulder,
College of Media, Communication and Information
tobias.hopp@colorado.edu

Gina M. Masullo
The University of Texas at Austin,
School of Journalism and Media
gina.masullo@austin.utexas.edu

Abstract

This study uses the sociological theories of deviance and social control to explore factors that contribute to post deletion and/or removal on Reddit's most active subreddits. Using a random sample of over 1 million Reddit submissions, we show that the presence of language attacking others' identities is somewhat strikingly associated with both user-initiated content deletion and moderator-initiated content deletion on Reddit. Findings offer insight into what it means to patrol the boundaries of acceptable communication online and what factors may contribute to the decision-making that precedes post deletion on social media platforms.

Keywords:
Content deletion, identity threats, toxicity

1. Introduction

Since social-networking platforms emerged in the late 1990s (boyd & Ellison, 2007), user-generated posts, a ubiquitous feature, have raised concerns because they are frequently mired by toxic attributes, such as insults, profanity, and threats (Chen, 2017). Platforms have several avenues to control this content: the use of volunteer or paid moderators who are in charge of policing online spaces (Roberts, 2019); allowing users to flag offensive speech so moderators can remove it (Crawford & Gillespie, 2016); machine learning that is trained to filter out toxic comments (Wakabayashi, 2017); and permitting original posters (OPs) to remove their own content at any point (Yilmaz et al., 2021). Ultimately, this results in deletion and/or removal of content, but a research gap exists regarding what factors contribute to the decision-making that leads to these deletions when humans—either as moderators or as posters—intervene. Filling this research gap is important because it contributes to our understanding of how users make decisions about what content is permissible and what content should be removed. Thus, understanding content deletion motivations helps illuminate how online users construct their experience in digital spaces. Further, our findings contribute theoretically to understanding how people online enact social control (Hirschi, 1969) by identifying deviant content through their deletions. We theorize deletion and/or removal of content may operate as a form of social control that communicates to other Redditors what type of content is permissible on a specific subreddit. Overall, this study seeks to address one overarching question: How do Redditors exact social control by making decisions about what content to delete or remove?

1.1. The case of Reddit

This study fills a gap in our understanding of decision-making on social media platforms by examining how Redditors police online communication through the deletion and/or removal of posts on the most active subreddits on Reddit. Subreddits are specific user-created communities that can be focused on a particular topic (e.g., "politics") or affinity group (e.g., "butch lesbians") and Reddit uses volunteer moderators who police their own subreddits. Reddit offers a robust platform to study how users regulate communication because, unlike other platforms (e.g., Facebook), it offers both "upvote" and "downvote" reactions that users can employ to give meaning to posts (Giuntini et al., 2019), signal that a post deserves attention (Carr et al., 2018), or communicate an emotional response to a post, similar to how facial cues operate offline (Derks, et al., 2007). Reddit subtracts the "downvotes" on a post from the "upvotes" to generate a "Reddit score," which is essentially a measure of post popularity. Thus, users can use the "downvote" to signal that a post contains problematic content as a form of social control, while "upvotes" indicate that content is favored (Masullo, 2022).

The "downvote" does even more than signal disapproval. It also removes "karma points" from the OP, so users who frequently get downvoted will have lower karma scores than those who consistently get upvoted (Anderson, 2015). This karma score is visible on users' profiles, so downvoting can be seen as a way to influence whether a particular user is "more highly valued within the collective" (Anderson, 2015, p. 8). Thus, downvoting can be seen as a means to police behavior because it offers consequences to the users in terms of having less "karma." This makes Reddit a useful platform to consider for this study because it allows one to consider whether post popularity, along with other factors—toxicity, visibility, and topic—correlate with whether a post is deleted and/or removed, either by a moderator or by the OP.

Gaudette et al. (2020) explored how Reddit's voting algorithm facilitates collective identity formation among members of the extreme right on the subreddit r/The_Donald, finding that the algorithm creates an environment wherein extreme right views are continuously validated through upvoting. Through a thematic analysis of posts, the study provides insight into how the voting algorithm of Reddit can be used to facilitate collective identity formation but is limited by its focus on a single subreddit's data. Cannon et al. (2020) explored Reddit's advice communities, finding that community guidelines often encourage downvoting as a way to minimize deviance from a subreddit's normative community values.

In addition to a thematic analysis of deleted and removed posts, here we examine the role of toxicity, popularity, and visibility in the decisions that moderators and posters make regarding deleting and/or removing posts.

2. Literature review

2.1. Online conversations and toxicity

Online discussions on social media require scholarly inquiry because they have become an essential way much of the public talks about topics in the news and contemporary society. These discussions can help people feel more involved with the news (Oeldorf-Hirsch & Sundar, 2015), foster political engagement (Shah, 2016), and provide a lens through which people understand and mirror societal discourses. Since the early days of these platforms, toxic content within these conversations has been a concern, with estimates showing 20% of online user-generated content exhibits some form of toxicity, such as insults, profanity, or threats (Chen, 2017). Toxic content has been shown to be even more frequent on particularly divisive issues like immigration (Santana, 2015) or race (DiFátima, 2023). Toxic content online can be defined in a variety of ways, but we define it as "impolite" meaning it contains attributes such as insults and profanity but also includes more virulent forms of communication, such as threats against a person or identity (Papacharissi, 2004). We consider all of these types of toxic speech because they provide a continuum of content that may cause harm (Chen, 2017). This type of content raises great concerns because it can emotionally exhaust people, such as the moderators tasked with reading it (Riedl et al., 2020), and also foment negative emotions and even aggressiveness (Rösner et al., 2016).

2.2. Content moderation

Platforms attempt to control these discussions by articulating rules for behavior through terms of service and community guidelines that are designed to foster civil discourse (Gillespie, 2018). They also manage conversations through moderation, where people or algorithms flag inappropriate comments and delete them or even block users who post such content repeatedly (Reidl et al., 2020). In this sense, moderation is defined as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" (Grimmelmann, 2015, p. 42). This moderation can be conducted through paid work outsourced to third-party companies or crowdsourcing platforms (Roberts, 2019), or through volunteer work such as flagging by users (Crawford & Gillespie, 2016). Moderation of posts can happen before publication, where a moderator must approve a post before it goes live, or after publication, where a moderator or the OP removes the post after it is published (Ksiazek, 2018). Moderators can also hide comments or ban users, depending on the platforms.

2.3. Theoretical framework: Deviance and social control

The sociological theory of deviance and social control (Hirschi, 1969) provides the theoretical foundation for our inquiry in the decision-making that relates to post deletion and/or removal on Reddit. The theory, particularly when applied through Johnson's (1991) threefold commitment model, involves how people make decisions and act within situations, constrained by structure (Ulmer, 2000). Hirschi (1969)

developed the theory arguing that people need social bonds, such as a commitment to following rules, or connections with other people, to help them follow social norms and eschew criminal behavior. In essence, he argued that people will act out unless constrained through social bonds or norms, debunking the persistent belief at the time that people needed "criminal motivation" to deviate from norms (Hirschi, 1969).

Johnson (1991) built on Hirschi's work and developed the three-commitment model to understand why people stay in interpersonal relationships and argued that three types of commitment play a role: structural commitment (feeling one must stay); moral commitment, (having a moral obligation to stay); and personal commitment (wanting to stay). We adopt this framework, arguing that these three factors may also play a role in the decision to delete a post. People, either the OPs or volunteer moderators, may feel a requirement that certain content must be removed because it clearly violates the platform's terms of service, the "structure" of the platform. They also may feel a moral obligation to remove certain content even if it doesn't explicitly violate terms of service because it is morally wrong (e.g., identity attacks). Finally, there may be content people just want to remove because if they are the OP, it is unpopular because of "downvotes," or other issues. Thus, we use the theory in a similar way as Ulmer (2000), who applied it to understanding why people participate in or leave social movement activities.

2.4. Better understanding content removal

Deletion offers the permanent removal of a post either by a volunteer moderator or by the OP. Deletion is important because it signals "how people seek to render certain aspects of data invisible" (Jacobsen, 2022). We delineate content removal in this study by focusing on who is deleting the content, either the OP of the content or the moderators in charge of a subreddit.

Our rationale for this question is that users may delete their own posts for different or the same reasons that moderators do. Moderators are tasked with upholding Reddit's overall content policy, as well as enforcing the specific rules for a subreddit. Posters do not have this obligation and may even have differing points of view on Reddit's moderation policies.

2.4.1. OP content deletion. When an OP deletes a post, one may infer that the post contains information the OP, upon reflection, considers damaging, inappropriate, overly personal, or poorly worded (Minaei et al., 2021), although the actual reason is not apparent to other users. For example, they may delete unpopular posts to protect their own reputation because they realize (from comments or "downvotes") that they made a mistake by posting the initial content (Yilmaz et al., 2021). To date, there are no existing thematic analyses of OP-deleted content on Reddit. To better understand commonalities as to what posts are deleted, we ask the following research question:

RQ1: *What specific topics or themes are most frequently present in the posts deleted by OPs on Reddit?*

2.4.2. Content moderation removal. Content moderation policies vary dramatically by platform, and can involve hate speech, misinformation, or explicit content (Buckley & Schafer, 2021; Langvardt, 2017). Additionally, topics related to controversial political issues or divisive social debates may also be subject to increased moderation due to their potential to incite uncivil discussions or violate platform policies (Almerekhi et al., 2020).

Even though Reddit's moderators do not disclose their reasons for removing posts, their motivations are somewhat more clear than OPs. One may infer that the posts violated Reddit's content policy or specific rules for a subreddit. Policies vary across subreddits, with each subreddit having its own set of community guidelines and moderation practices. This decentralized approach to content moderation allows for a diverse range of topics and discussions to flourish on the platform. However, it also raises questions about the consistency and transparency practices across different subreddits (Gibson, 2019).

By examining the topics that are most commonly deleted by volunteer moderators and comparing those to the topics most commonly deleted by OPs, we can gain insights into how subreddit-specific content moderation policies and practices influence the types of discussions that are allowed or disallowed on the platform.

RQ2: *Which specific topics or themes are most frequently present in the posts removed by volunteer moderators on Reddit?*

RQ3: *Which common themes overlap between content deleted by OPs and content removed by moderators on Reddit?*

For both content moderation and OP deleting, we perform thematic analyses by applying unsupervised topic modeling techniques to explore the most

common topics that are found in the deleted data. We aim to infer as much as possible what moderation policies are being most enforced and to better understand what types of content individuals themselves are most commonly compelled to delete.

2.5. Post attributes and likelihood for deletion

Beyond an exploration of deleted and removed content, we also set out to understand the extent to which specific attributes from previous research (toxicity, popularity, and visibility) might predict whether a post was deleted either by an OP or a volunteer moderator.

2.5.1. Toxicity. For toxicity, we focus on four types of uncivil content identified by Google's Perspective. The algorithm was developed by Jigsaw and Google's Counter Abuse Technology, and it uses supervised machine learning to generate a score (ranging from 0 to 1) for the following attributes: identity threats, hateful content that targets someone's identity, such as gender or race; insults, negative inflammatory content directed toward a person or group; profanity, curse words or obscene language; and threats, content that expresses intention to inflict injury or violence against a person or group. Much research shows that uncivil posts are frequently removed by moderators, although uncivil posts may also draw greater engagement from other users (Vargo & Hopp, 2023). This offers evidence that moderators might be more likely to remove toxic posts and that posters might be more likely to remove their own posts if they deem them as violating norms of civility on a subreddit. Indeed, qualitative research shows that other users appreciate when posters delete offensive or abusive content (Yilmaz et al., 2021).

H1a: Posts with higher levels of identity threats are more likely to be deleted or removed, either by OPs or volunteer moderators, on Reddit.

H1b: Posts with higher levels of insults are more likely to be deleted or removed, either by OPs or volunteer moderators, on Reddit.

H1c: Posts with higher levels of profanity are more likely to be deleted or removed, either by OPs or volunteer moderators, on Reddit.

H1d: Posts with higher levels of threats are more likely to be deleted or removed, either by OPs or volunteer moderators, on Reddit.

2.5.2. Popularity. In addition to toxicity, we considered that posts might be deleted that were either more or less popular, as evidenced by the "Reddit score." Our rationale is Lampe et al.'s (2020) natural experimental finding that comments with lower "scores" on Slashdot were more likely to get moderated, meaning volunteer moderators either elevated or demoted less popular content. Slashdot's "comment score" operates similarly to the Reddit score, where users acquire a reputation or karma through activities such as reading, posting, or moderating comments (Lampe et al., 2020). So a higher comment score signals a greater reputation, much as a higher Reddit score does.

OPs may delete posts that end up with a low score because that indicates a high frequency of "downvotes," which could convey that the post is problematic or does not fit within the norms of the specific subreddit. In this case, popularity becomes a proxy for public sentiment on the subreddit, and the OP may respond to that score by removing the post to better fit within the norms of the subreddit. In this way, the OP could be deleting a post in response to the social control enacted by the other users through their use of the "downvote" and "upvote." But it is also possible that an OP may delete a post that receives few votes (either up or down) because that may convey that the post does not resonate with the community on the subreddit.

The literature is mixed on whether toxic posts would be more likely to warrant "upvotes" or "downvotes" that would make these posts more popular. Vargo and Hopp (2023) had a similar finding, although few posts in their sample were actually uncivil. Rajadesingan et al., (2020) analyzed political subreddits on Reddit and found virtually no correlation between the voting scores for comments and the comments' toxicity, suggesting that popularity of a content may have little to do with whether it is toxic. However, Davis and Graham (2020) found that upvotes preceded positive sentiments and "downvotes" preceded negative sentiments, suggesting that the content of comments may influence popularity. Additionally, downvoted content received higher levels of engagement than upvoted content (Davis & Graham, 2020).

H2a: Posts with a higher Reddit score, indicating greater popularity, are less likely to be deleted by the OP on Reddit.

H2b: Posts with a higher Reddit score, indicating greater popularity, are less likely to be removed by volunteer moderators on Reddit.

2.5.3. Community size. An important factor to consider when examining content deletion and removal on Reddit is the size of the subreddit community, measured in terms of the number of subscribers. As a community grows in popularity and gains more subscribers, it is likely to have a more engaging and active community (Baek & Shore, 2019). With an increase in subscribers, the subreddit is also likely to receive more submissions (or posts), resulting in a greater need for content moderation and a more rigorous content removal process.

Moreover, larger subreddits are more likely to be monetized, and, therefore, subject to content guidelines that protect advertisers from inadvertently sponsoring unsavory content (Vargo et al., 2023). In general, the size of a subreddit has been shown to negatively correlate with toxic-related behaviors. Generally speaking, larger subreddits have less toxic discussion and rigorous content moderation policies (Vargo & Hopp, 2023).

All things considered, in larger subreddits, selectivity, or exclusivity, may play a significant role in determining whether a post is removed or deleted. For a post to be left untouched in a growing subreddit, it must be considered exceptional or highly relevant to the community's interests, norms, and values. Considering the potential impact of community size on content removal and deletion, we propose the following hypotheses:

H3a: Posts within larger subreddits, as indicated by a higher number of subscribers, are more likely to be deleted by the OP on Reddit.

H3b: Posts within larger subreddits, as indicated by a higher number of subscribers, are more likely to be removed by volunteer moderators on Reddit.

3. Method

3.1. Sample

We wanted to draw a representative sample of active subreddits. To do so, we collected all top 100 "safe for work" subreddits as calculated by redditlist.com. This list broadly represents subreddits that are known to have relatively high levels of recent activity. To avoid issues with seasonality, we sampled a two-year period from January 1, 2021 to December 31, 2022. From our universe of 39.70M Reddit posts and replies, we randomly sampled 27.51% of the submissions, resulting in a final sample size of 1,092,286 posts. Sampling was done by randomly shuffling the submissions and receiving as many as possible given our quota from Alphabet.

3.2. Measures

3.2.1. Deletion outcomes. Deletion outcomes were retrieved from the post metadata. Around half of the submissions (n = 518,772; 47.5%) were not subject to any moderation action, 23.8% (n = 259,878) were deleted by OP, 23.4% (n = 255,379) were deleted by a subreddit moderator, and 5.3% (n = 58,257) were subject to some other moderation action (e.g., copyright takedown, filtered as spam, etc.).

3.2.2. Toxicity detection. The Perspective API was employed in this study to detect toxicity in user-generated comments. Developed using an extensive dataset of human-provided annotations, the API generates a continuous probability value (P; theoretical range: 0-1.00) representing the presence of a specified attribute in a given text. The Perspective API is a popular and reliable approach to measuring uncivil content, including Reddit. Full information on the development processes used to create the Perspective API can be found at https://developers.perspectiveapi.com/s/?language=en_US. In this particular project, we deployed the Perspective submodels for identity threat, insults, profanity, and threat. The identity threat attribute detects identity-based attacks (IBAs), like racism. The insults measure identifies negative comments directed at individuals, while the profanity algorithm detects vulgar language and its derivatives. Finally, the threat measure captures the intention to harm individuals or groups. Sample-wide means for the Perspective-generated incivility measures were: identity attack: M = 0.02, SD = 0.05; insults: M = 0.05, SD = 0.09; profanity: M = 0.07, SD = 0.12; and threatening language: M = 0.02, SD = 0.06.

3.2.3. Popularity. To measure popularity, we used the Reddit score. Reddit scores for each post are included as part of the metadata package. Reddit scores are calculated as the number of positive votes minus the number of negative votes. To avoid certain types of undesirable behaviors (e.g., "piling on") submissions cannot have a score value lower than 0, even if the post receives more downvotes than upvotes (M = 294.46, SD = 3,012.98; min = 0, max = 265,929).

3.2.4. Community size. Community size was measured by assessing the number of subreddit

subscribers at the time of the submission's creation. The community size variable was a metadata component. Because the evaluated subreddits were quite large (M = 15,435,443, SD = 13,735,455; min = 8,908, max = 46,442,576), this variable was scaled by 100,000 to aid coefficient interpretability (M = 154.35, SD = 137.35; min = 0.09, max = 464.43).

3.2.5. Covariates. In addition to the toxicity, popularity, and community size variables, a handful of covariates were assessed. First, because sexual content is a frequent reason that content is removed from popular Reddit forums (which are typically denoted as "safe for work" places), Perspective's sexual content submodel was included as a control (M = 0.05, SD = 0.10). We also assessed the number of comments associated with each submission (M = 22.47, SD = 279.10; min = 0, max = 92,768) and the number of community awards associated with each submission (M = 0.42, SD = 9.12; min = 0, max = 5,154; see: https://www.reddit.com/r/announcements/comments/chdx1h/introducing_community_awards/

3.3. Thematic analysis

We explored what common themes could be surmised from popular topics in the text of deleted submission titles. To do so, we leveraged a combination of unsupervised deep learning, generative learning and most importantly, our own intuition as researchers, to better qualitatively understand the types of content most commonly deleted.

First, we fit supervised topic modeling using BERTopic, an advanced topic modeling algorithm based on the transformer architecture (Grootendorst, 2022). We created two data segments that were then modeled separately: 1) posts that had been removed by moderators, and 2) posts deleted by users.

We fine-tuned the topic representations using a representation model called KeyBERTInspired and fit the BERTopic model to the dataset of deleted posts. This model uses machine learning algorithms to identify patterns in large volumes of text data and group them into topics that share similar content. To evaluate the goodness of fit of our topic model, we calculated coherence scores, a widely used metric in topic modeling literature. Coherence measures the semantic similarity between the top words within a topic, indicating the interpretability and quality of the resulting topics. Specifically, we used the Content Vector (CV) coherence measure, which is based on a sliding window, a one-set segmentation of the top words, and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity. A higher coherence score suggests a better fit of the topic model, with more semantically related words within each topic. After fitting models on the datasets, we calculated the coherence score and then reduced the number of topics to the top 100. By comparing the coherence scores before and after reducing the number of topics, we can assess whether the model's interpretability improved or worsened. For both models, the initial BERTopic model generated a large number of topics (moderated k = 2,384 topics; deleted k = 2,487) with a coherence score of 0.36 and 0.37 respectively. After reducing the number of topics to the top 100, the coherence score increased to 0.420 in the moderated model, and to 0.415 in the OP deleted model. These improvements in the coherence score suggest that the top 100 topics are more interpretable and semantically coherent than the initial set of 2,384 topics, while also being a reasonable amount of data to interpret qualitatively.

By fitting the model on the dataset of the deleted posts, we inferred the most common topics that were deleted on Reddit. For each of the identified topics, we obtained their associated words and representative documents from the model. These topic words and documents provide insight into the type of content within each topic and allow us to infer likely reasons for their deletion.

After identifying the topics, we used OpenAI's Generative Pre-Trained Transformer version 4 (GPT-4) to assist us in summarizing the themes of Reddit posts being deleted and why such content is frequently deleted by users. GPT-4 is an advanced natural language processing model. For each identified topic, we provided GPT-4 with the topic's keywords and representative documents and instructed it to return a topic description for each along with the likely reason why the content would be often deleted by users. We issued all API calls with 0 temperature to minimize hallucination. Once we had topic descriptions and reasons for deletion, we fed this information back into the GPT-4 model to summarize the topics and reasons into major themes, discuss each theme in detail, and provide examples or specific subreddits for each theme when possible. We compared our notes with reading the 200 topics, and we feel that the model accurately summarized the major themes and possible reasons why content was removed from the platform. To visually inspect the

topics and the most discriminate words associated with them.

4. Results

RQ1 asked what common content themes emerged for posts that were deleted by OPs. Several major themes emerged as reasons for deletion. These included off-topic or repetitive content, sensitive or controversial topics, privacy concerns, violation of subreddit rules, promoting misinformation, copyright issues, financial and economic topics, and illegal activities and substance abuse.

RQ2 asked what common themes emerged for content that was removed by moderators. Major themes included controversial or sensitive content, misinformation, off-topic or irrelevant content, low-quality or repetitive content, explicit content, copyright infringement, and violation of subreddit-specific rules.

RQ3 asked what common themes appear in both moderator and OP deletions. In reviewing the themes for similarities, six common themes emerged in both data. These were off-topic or repetitive content; sensitive or controversial topics; violation of subreddit rules or guidelines; promoting misinformation; inappropriate or offensive content; or copyright issues (see Table 1**)**.

To assess Hypotheses 1–3, a series of binary logistic regression models were estimated. Given that the current dataset included more than 1 million observations, frequentist p-values were judged to be non-informative and were not generated. Each logistic regression model contained Perspective API attributes (identity attack, threatening language, insulting language, profane language), community-generated score for the submission (i.e., popularity), the number of subreddit subscribers (i.e., community size), number of comments associated with the submission, and number of awards issued to the submission.

Model 1 predicted the likelihood that a given submission was deleted by the author while Model 2 predicted the likelihood that a post was deleted by a subreddit moderator. Finally, Model 3 explicitly compared self- vs. moderator-deletions. Because Model 3 only assessed posts that were deleted by an OP or moderator, it focused on only a portion (n = 515,257; 47.1%) of the total corpus. Models 1 and 2 used the entire corpus (n = 1,092,286). For all models, odds ratios (OR) are reported.

Model 1, shown in Table 2, showed that posts featuring identity attacks are substantially more likely to be deleted by the OP. Specifically, we see that a 1-unit change on the identity attack variable (i.e., P = 0 vs. P = 1) is associated with a 576% (OR = 6.76) increase in the probability that an author initiates submission deletion. This can be contrasted with more moderately-sized odds ratios for the threatening language (OR = 1.65) and insulting language (OR = 3.30) and a negative relationship between author-initiated deletion and profanity (OR = 0.34). Neither the size of the subreddit (OR = 1.00) nor the submission score (OR = 1.00) demonstrated a substantial relationship with the likelihood of author-initiated post deletion.

A similar pattern emerged in Model 2 (see Table 2). We again found that posts featuring language attacking the identity of others were substantially more likely to incur moderator-initiated deletion (OR = 5.58) than posts featuring threatening language (OR = 1.34), insulting language (OR = 0.68), or profane language (OR = 1.10). Finally, the results of Model 2 did not indicate that subreddit size (OR = 1.00) or post score (OR = 1.00) played an outsized role in the prediction of moderator-initiated post deletion.

Model 3 compared deletion outcomes initiated by the author with deletion outcomes initiated by the moderator. Author deletions were coded as 0 and moderator deletions were coded as 1. The results of Model 3 indicate that moderators are substantially more likely than post authors to delete profane posts (OR = 2.31). Alternately, we see that authors are more likely to delete posts containing identity attacks (OR = 0.95), threatening language (OR = 0.85), and insulting language (OR = 0.31). We failed to find evidence that subreddit size (OR = 1.00) or post score (OR = 1.00) played a substantially meaningful role in distinguishing between moderator and author-initiated post deletions.

Taken as a whole, the results seemed to provide somewhat consistent evidence in support of H1a as we observed fairly straightforward evidence that the use of language that threatens the identities of others is strongly associated with both OP deletions and volunteer moderator deletions. Likewise, in the case of H1d, we observed positive associations between the occurrence of threatening language and both OP and moderator removals. The evidence in support of H1b and H1c is, comparatively speaking, a bit more mixed. Insulting language, for instance, was robustly and positively associated with OP deletions but weakly and negatively predictive of moderator deletions. Alternately, profanity was negatively associated with OP deletions but positively (albeit marginally so)

associated with moderator deletions. For these reasons, we failed to find support for H1b and H1c. As it pertains to H2 and H3, we failed to find evidence that either Reddit score (i.e., post popularity) or subreddit size (i.e., community size) played a meaningful role in OP or moderator decisions to remove content from Reddit.

5. Discussion

This study set out to better understand the factors that underlie deletion and/or removal decisions on Reddit. The study aimed to fill a gap in the literature regarding how these decisions may operate as a form of social control that communicates to other Redditors what type of content is permissible on a specific subreddit. Using a randomly selected sample of over 1 million submissions, our findings suggest that both OP and subreddit moderators are sensitive to the presence of language that attacks the identity of individuals and groups, and, therein, that the presence of such language is increasingly likely to result in content deletion. We found a similar pattern of results for threatening language as posts containing threats were comparatively likely to be taken down by either the originating author or the subreddit's moderator. Notably, democratic incivility theory has long focused on the damaging effects of identity- or group-based attacks, as such attacks often serve as a discursive means of delegitimizing participation in the public sphere (e.g., Hopp, 2019). While the current research context doesn't necessarily refer to the sort of explicit political and/or social communication typically addressed by political communication scholars, our results do suggest that both moderators and OPs recognize the inherently damaging nature of group-based attacks, and, as such, may be more likely to undertake moderation actions as a means of preserving community harmony or protecting one's reputation. Theoretically, our findings suggest that both moderators and OPs see group-based attacks as violating norms, so deleting/removing these posts exerts social control.

Our failure to straightforwardly support Hypotheses 1b and 1c points to some potentially interesting discrepancies between OP and moderator deletion decisions. For instance, in Model 1, we found that posts containing high levels of insulting language were nearly 250% more likely to be deleted by their originating authors. Conversely, in Model 2, we found that posts containing insults were less likely to be deleted by moderators. One explanation for this finding might lie in the fact that OPs and moderator deletion decisions are guided by differing concerns. Moderators are typically tasked with ensuring that community discussions conform to Reddit's broader community and moderator guidelines (e.g., https://www.redditinc.com/policies/moderator-code-of-conduct), and with the specific group norms that manifest themselves in guidelines specific to the subreddit. Reddit users, on the other hand, use the platform to generate and pursue social linkages and share information of joint interest. As such, it may be the case that subreddit moderator decisions around deletion are primarily driven by a sort of socially-detached administrative responsibility to ensure adherence to a set of rule-based logics. OPs, on the other hand, recognize that productive and enjoyable social interactions rest upon adhering to norms of politeness, and are then more likely to engage in the self-censorship of posts that might contain unnecessary levels of discourtesy and disrespect. This suggests that OPs may be exerting more social control with their deletions/removal by trying to create a hospitable digital space.

Overall, these findings suggest some broader insights into how moderators and OPs are making decisions on whether to delete and/or remove content. Both moderators and OPs are clearly focusing efforts on identity attacks, which scholars have identified as more problematic than other forms of toxicity (e.g., Chen, 2017; Rossini, 2020). Notably, however, moderators were more likely to remove profanity than OPs. A similar trend was found with insulting language, with insults factoring more into social media deletion decisions for OPs. These findings suggest that OPs and moderators approach decision-making differently for social media deletion and/or removal. While our study design does not allow us to know for certain what motivated post-deletion and/or removal, our findings suggest that both OPs and moderators may see identity attacks as requiring deletion or even creating a moral obligation (Johnson, 1991) for removal, thus prioritizing this content as most normatively problematic.

We also note that the community size and Reddit score variables did not appear to be meaningfully associated with deletion and/or removal decisions. In regard to the former, we had assumed that economic and related considerations might encourage both moderators and OPs in large communities to be somewhat more judicious when making decisions around content suitability. Given that this was not the case, it may be that removal and deletion decisions are

more narrowly targeted to enforce explicit and established community rules and norms, and, therein, factors such as economic viability or broader visibility are not relevant to Reddit users and moderators. In regard to the latter, our findings suggest that community feedback, in the form of Reddit karma scores, has limited utility when it comes to guiding deletion and removal decisions. Note, we do not believe that community feedback is unimportant in OP or moderator deletion/removal decisions. Instead, it may be the case that the qualitative content of the user comments associated with the submission may play a stronger role in decision-making than quantitative up/down voting outcomes.

5.1. Limitations and future research

While we used the most popular subreddits, decision-making regarding deletion and/or removal of posts may differ in less popular subreddits or those that focus on specific populations (e.g., minoritized groups). Future research should attempt to replicate our findings on different types of subreddits that may have varying populations and, as a result, different decision-making criteria for deletion and/or removal of posts. Additionally, we focused on the role of post popularity, toxicity, and community size in deletion and/or removal decisions, but other factors may also be worth considering. From an empirical perspective, we note that Models 1–3 were associated with somewhat small pseudo R-squared values (range: 0.004 - 0.02). While pseudo R-squared coefficients cannot be straightforwardly interpreted as a measure of accounted-for variance (e.g., Jorgensen & Williams, 2020), they do provide an approximate measure of model fit. Our findings, as such, point to value in surveying or interviewing OPs and moderators to see what factors they consider in their own minds before making decisions on what content to delete or remove. Finally, the present results are derived from associational data, and don't provide clear insights into the causal mechanisms that may or may not be driving deletion decisions on Reddit.

12. References

Almerekhi, H., Jansen, B. J., & Kwak, H. (2020). Investigating toxicity across multiple Reddit communities, users, and moderators. *WWW '20: Companion Proceedings of the Web Conference 2020*, 294–298. https://doi.org/10.1145/3366424.3382091

Anderson, K. E. (2015). Ask me anything: What is Reddit? *Library Hi Tech News, 32*(5), 8–11. https://doi.org/10.1108/LHTN-03-2015-0018

Baek, J., & Shore, J. (2019). Forum size and content contribution per person: A field experiment. *Boston University Questrom School of Business Research Paper No. 3363768*. https://doi.org/10.2139/ssrn.3363768

boyd, d., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Buckley, N., & Schafer, J. S. (2022). "Censorship-free" platforms: Evaluating content moderation policies and practices of alternative social media. *For(e)dialogue*, *4*(1). https://doi.org/10.21428/e3990ae6.483f18da

Cannon, E., Crouse, B., Ghosh, S., Rihn, N., & Chua, K. (2022). "Don't downvote a$$$$$$s!!": An exploration of Reddit's advice communities. *Proceedings of the 55th Hawaii International Conference on System Sciences*, 2940–2949. https://doi.org/10.24251/hicss.2022.363

Carr, C. T., Hayes, R. A., & Sumner, E. M. (2018). Predicting a threshold of perceived Facebook post success via likes and reactions: A test of explanatory mechanisms. *Communication Research Reports, 35*(2), 141–151. https://doi.org/10.1080/08824096.2017.1409618

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Palgrave Macmillan.

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society, 18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Davis, J. L., & Graham, T. (2021). Emotional consequences and attention rewards: The social effects of ratings on Reddit. *Information, Communication & Society, 24*, 649–666. https://doi.org/10.1080/1369118X.2021.1874476

Derks, D., Bos, A. E. R., & Grumbkow, J. (2007). Emoticons and social interaction on the Internet: The importance of social context. *Computers in Human Behavior, 23*(1), 842–849. https://doi.org/10.1016/j.chb.2004.11.013

DiFátima, B. (2023). *Hate speech on social media*. LabCom Communication & Arts.

Gaudette, T., Scrivens, R., Davies, G., & Frank, R. (2020). Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society, 23*, 3491–3508. https://doi.org/10.1177/1461444820958123

Gibson, A. (2019). Free speech and safe spaces: How moderation policies shape online discussion spaces. *Social Media + Society, 5*(1). https://doi.org/10.1177/2056305119832588

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press. https://doi.org/10.12987/9780300235029

Giuntini, F. T., Ruiz, L. P., Kirchner, L. D., Passarelli, D. A., Dos Reis, M. D. J. D., Campbell, A. T., & Ueyama, J. (2019). How do I feel? Identifying emotional expressions on Facebook reactions using clustering mechanism. *IEEE Access, 7*, 53,909–53,921. https://doi.org/10.1109/ACCESS.2019.2913136

Grimmelman, J. (2015). The virtues of moderation. *Yale Journal of Law and Technology, 17*(42), 42–109.

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv preprint. arXiv:2203.05794.

Hirschi, T. (1969). *Causes of delinquency*. University of California Press.

Hopp, T. (2019). A network analysis of political incivility dimensions. *Communication and the Public, 4*(3), 204–223.

Jacobsen, B. N. (2022). 'You can't delete a memory'; Managing the data past on social media in everyday life. *Sociological Research Online, 27*(4), 1003–1019. https://doi.org/10.1177/13607804221110237

Johnson, M. P. (1991). Commitment to personal relationships. In W. H. Jones, & D. W. Pelman (Eds.), *Advance in personal relationships* (pp. 117–143). Jessika Kingsley.

Jorgensen, A., & Williams, R. A., (2020). Goodness-of-fit measures. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, & R. A. Williams (Eds.), *SAGE Research Methods Foundations*. https://doi.org/10.4135/9781526421036946001

Ksiazek, T. B. (2018). Commenting on the news: Explaining the degree and quality of user comments on news websites. *Journalism Studies, 19*(5), 650–673.

Lampe, C., Zube, P., Lee, J., Park, C. H., & Johnston, E. W. (2014). Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly, 31*(2), 317–326. https://doi.org/10.1016/j.giq.2013.11.005

Langvardt, K. (2017). Regulating online content moderation. *Georgetown Law Journal, 106*(5), 1353–1388. https://doi.org/10.2139/ssrn.3024739

Masullo, G. M. (2022). Facebook reactions as heuristics: Exploring relationships between reactions and commenting frequency on news about COVID-19. *First Monday, 27*(8). https://doi.org/10.5210/fm.v27i8.12674

Minaei, M., Mouli, S. C., Mondal, M., Ribeiro, B., & Kate, A. (2021). *Deceptive deletions for protecting withdrawn posts on social media platform*. arXiv. https://doi.org/10.48550/arXiv.2005.14113

Oeldorf-Hirsch, A., & Sundar, S. S. (2015). Posting, commenting, and tagging: Effects of sharing news stories on Facebook. *Computers in Human Behavior, 44*, 240–249. https://doi.org/10.1016/j.chb.2014.11.024

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society, 6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. *Proceedings of the International AAAI Conference on Web and Social Media, 14*(1), 557–568. https://doi.org/10.1609/icwsm.v14i1.7323

Riedl, M. J., Masullo, G. M., & Whipple, K. N. (2020). The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior, 107*, 106262. https://doi.org/10.1016/j.chb.2020.106262

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior, 58*, 461–470. https://doi.org/10.1016/j.chb.2016.01.022

Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research, 49*(3), 399–425. https://doi.org/10.1177/0093650220921314

Santana, A. D. (2015). Incivility dominates online comments on immigration. *Newspaper Research Journal, 36*(1), 92–107. https://doi.org/10.1177/073953291503600107

Shah, D. V. (2016). Conversation is the soul of democracy: Expression effects, communication mediation, and digital media. *Communication and the Public, 1*(1), 12–18. https://doi.org/10.1177/2057047316628310

Ulmer, J. T. (2000). Commitment, deviance, and social control. *The Sociological Quarterly, 41*(3), 315–336. https://doi.org/10.1111/j.1533-8525.2000.tb00081.x

Vargo, C., & Hopp, T. (2023). Incivility on popular politics and news subreddits: An analysis of in-groups, community guidelines and relationships with social media engagement. https://hdl.handle.net/10125/102930

Vargo, C., Hopp, T. & Agarwal, P. (2023). Inside a social media brand safety algorithm: A computational investigation of subreddits, toxicity, and advertising inventory. *Proceedings of the 2023 American Academy of Advertising Annual Conference*.

Wakabayashi, D. (2017). *Google cousin develops technology to flag toxic online comments*. The New York Times. https://www.nytimes.com/2017/02/23/technology/google-jigsaw-monitor-toxic-online-comments.html.

Yılmaz, G. S., Gasaway, F., Ur, B., & Mondal, M. (2021). Perceptions of retrospective edits, changes, and deletion on social media. *Proceedings of the International AAAI Conference on Web and Social Media, 15*, 841–852. https://doi.org/10.1609/icwsm.v15i1.18108

Table 1. Common themes for removed posts in both moderator and OP deletion/removal data

Theme	Definition
Off-topic	Is not relevant to the subreddit's main theme or does not contribute to meaningful discussions.
Sensitive/controversial	May incite heated debates or arguments, violate community guidelines, or spread misinformation.
Rule/guideline violation	Does not adhere to the specific rules or guidelines.
Misinformation	Promotes or spreads misinformation, conspiracy theories.
Inappropriate/Offensive	Hate speech, explicit content, offensive humor, or content that incites racial tensions.
Copyright issues	Copyright infringement, unauthorized leaks, or self-promotion.

Table 2. Summary of logistic regression models predicting various moderation outcomes

	Model 1	Model 2	Model 3
	OP Deletion OR	Moderator Deletion OR	OP vs. Moderator Deletion OR
Identity Attack	6.76	5.58	0.95
Threatening Language	1.70	1.34	0.85
Insulting Language	3.45	0.68	0.31
Profanity	0.34	1.10	2.31
Community Size	1.00	1.00	1.00
Post Score	1.00	1.00	1.00
Sexually Explicit Language	4.82	0.85	0.29
Number of Post Comments	0.99	1.00	1.00
Total Post Awards	1.00	1.00	1.00
N	1,092,286	1,092,286	515,257
Nagelkerke R^2	0.03	0.004	0.01