On Positive Moderation Decisions

Mattia Samory
GESIS
mattia.samory@gesis.org

Abstract

A crucial role of moderators is to decide what content is allowed in their community. Though research has advanced its understanding of the content that moderators remove, such as spam and hateful messages, we know little about what moderators approve. This work analyzes moderator-approved content from 49 Reddit communities. It sheds light on the complexity of moderation by giving empirical evidence that the difference between approved and removed content is often subtle. In fact, approved content is more similar to removed content than it is to the remaining content in a community—i.e., content that has never been reviewed by a moderator—along dimensions of topicality, psycholinguistic categories, and toxicity. Building upon this observation, I quantify the implications for NLP systems aimed at supporting moderation decisions, which often confabulate moderator-approved content with content that has potentially never been reviewed by a moderator. I show that these systems would remove over half of the content that moderators approve. I conclude with recommendations for building better tools for automated moderation, even when approved content is not available.

1 Introduction

The speed and volume at which online spaces are flooded with abusive content are impossible to handle by human moderators. Moreover, prolonged exposure to foul content takes a toll on their mental well-being (Gillespie 2018). The research community proposed technological solutions to these problems, in the form of classifiers for automated moderation that review content at scale and flag it for inspection by human moderators (Salminen et al. 2018). So far the focus has been on abusive content that requires removal to guarantee healthy online spaces. As the stress is on the content to be removed, approved content received little attention. This is at least partly because what moderators explicitly approve is unavailable to researchers, who settle instead on content that remains accessible online (possibly never reviewed by moderators) to compare against removed content. Yet, moderators police content that is controversial: like it is naive to consider blatant offenses as representative of all content to be removed, so it is to assume phatic talk as representative of content to be approved. To replicate the decisions of human moderators accurately, it is equally important that we understand what moderators approve.

This paper offers just such an insight. I collect a complete set moderation of decisions in 49 communities on Reddit, over more than two years. Adopting a mixed-methods approach, I characterize the topics of approved comments that are shared across communities. I find that approved comments challenge community norms, opposing provocative phrasing, hot-button topics, and attacks to moderators. Referring to literature in social computing and natural language processing, I compare and contrast approved comments with comments that moderators removed or never reviewed, along multiple dimensions of semantic similarity, psycholinguistic categories, toxicity, prose quality, and community feedback. I find that approved comments and removed comments share many traits, including high toxicity and frequent insults. In contrast, never reviewed comments appear trivially offensive. In the light of these observations, I ask: how do automated moderation tools, trained on removed and never reviewed content, perform on replicating actual approve/remove moderation decisions? Poorly: not only distinguishing approved and removed comments is an intrinsically harder task, but also that automated moderation systematically would remove over half of what human moderators approve. I unpack model errors and find that misclassifications happen on approved comments that on the surface look offensive. Aware of the general unavailability of ground truth on what moderators approve, I probe several data sampling heuristics to improve model recall for approved comments. This work thus provides both theoretical insights for researchers on community norms, as well as practical implications for practitioners developing automated moderation tools.

The paper is structured as follows. I contextualize this work in related research. Then, I provide the necessary background on moderation on Reddit, before detailing how I collected moderation data. I discuss important ethical considerations related to this study. Then, I describe three main analyses. First, I give a quantitative characterization of the content of approved comments. Next, I quantify and analyze the errors of automated moderation tools. Last, I test heuristics for substituting approved comments when unavailable. In the light of these findings, I conclude by discussing theoretical and practical implications.
2 Related Work

I draw from two existing lines of research: social computing studies on moderation practices, and applications of natural language processing (NLP) for content moderation.

2.1 Moderation Practices and Norms in Online Communities

Research in social computing has focused on moderation practices and community norms in the context of online governance. Moderation serves the crucial function of gatekeeping communal spaces, and is tasked with striking the balance between individual safety and freedom (Blackwell et al. 2018; Wadden et al. 2021; Gillespie 2018). Online platform policies aim at inclusiveness and generality, often to a fault: they neglect that what is allowed in one social circle may be condemned in another (Fiesler et al. 2018; Pater et al. 2016). On the other hand, the norms of specific subgroups remain largely unwritten (Juneja, Subramanian, and Mitra 2020). Research in governance thus turned to studying the decisions of each community’s moderators, as they de facto embody its specific norms (Chandrasekharan et al. 2018; Rajadesingan, Resnick, and Budak 2020). This paper is situated in this theoretical framework. Whereas previous research looked at the enforcement of negative norms (what is not allowed) to guarantee users’ safety, I look at the enforcement of positive norms (what is acceptable) to preserve users’ freedom.

2.2 Computational Methods for Supporting Moderation

A second line of research in NLP tackles the practical challenge of supporting human moderation through automation. The NLP community and online platforms made available data and infrastructure to promote the development of automated moderation tools (Schmidt and Wiegand 2017; Vidgen et al. 2019), notably through shared tasks and sponsored competitions such as SemEval and the Toxic Comment Classification challenges. These benchmarks have the goal of identifying content to be removed (Zampieri et al. 2019). Differently from existing analyses, I tackle the dual problem of identifying approved content, relying on a multi-community and complete dataset of moderation decisions.

Furthermore, the development of automated moderation advanced our understanding of the language of removed content. In particular, it highlighted the need to account for the variety, nuance, and situatedness of abusive content (Salminen et al. 2018). These results call for skepticism of tools that over-rely on simplistic lexical markers, and advocate for studies that span different contexts (Sap et al. 2019). This paper leverages the current understanding of removed content to study approved content across 49 discussion communities on a variety of topics. I show that approved and removed content share several linguistic properties, clarifying challenges for NLP practitioners.

3 Background: Reddit and Moderation

The day-to-day administration of subreddits is left to users who volunteer as moderators. Reddit offers several tools to help moderators accomplish their tasks. A key tool is the “moderator action log” (also known as modlog), which maintains a record of all actions undertaken by moderators in the past 90 days. The modlog includes the type of action performed, be it a sanctions like banning a user, or community work like customizing the appearance of the subreddit. Among other essential details reported in the modlogs are the moderator who performed the action, the user or the content that has been sanctioned, and the reason for the sanction.

The modlog is accessible exclusively to the moderators of the corresponding subreddit. Because of the privileged position of moderators, users started several grassroots initiatives to make moderation more transparent and accountable. One such initiative is “public modlogs,” which aims at maintaining subreddits censorship-free. Subreddit moderators can opt-in and invite a bot as a moderator with read-only permissions. The bot then shares its view of the modlogs publicly, thus allowing regular users to read the modlog by following a link. Several third-party tools leverage this and similar bots, e.g. ceddit.com and modlogs.fyi.

4 Data

4.1 Data Collection

I scraped the modlogs of the subreddits joining the initiative. The complete dataset includes over 4 million moderation actions, it spans over 2 years and covers over 400 subreddits. Subreddits range in topic—from news and politics to technology and games—as well as in size—from r/memeswithnomods with 7 subscribers to r/MurderedByWords with over 2 million at the time of writing. The dataset is a longitudinal and complete set, which means that it includes content removals as well as approvals, and all other actions performed by moderators in their official role during the data gathering period.

In addition to the modlog data, I obtained all submissions and comments in the same subreddits using the official Reddit API and the archives offered by pushshift.io.

4.2 Filtering and Cleaning Comments

In this work I focus on comments that have been explicitly approved or removed by moderators. I then take multiple steps to filter and clean data. To avoid ambiguity, I discard all comments on which moderators took multiple contrasting actions, e.g. first removing and then re-approving the comment. Upon preliminary inspection, those comments belong to one of two categories: either actions of a moderation bot that are overruled when revised by human moderators, or controversial comments that lay at the boundary of what

---

1https://sites.google.com/site/offensevalsharedtask/
2https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
3https://mods.reddithelp.com/hc/en-us/articles/360022402312
4https://www.reddit.com/r/publicmodlogs/
5e.g. the modlog for the subreddit r/conspiracy is accessible at https://www.reddit.com/r/conspiracy/about/log?feed=7e9b27126097f51ae69cd5b049af34891da6baa6&user=publicmodlogs
Figure 1: Subreddits in the study. I report the number of approved and removed comments for each subreddit (log-scaled), together with their overall topical focus.

is admissible in their community and thus elicit contrasting reactions from the moderators.

I further clean the text of the comments. I convert markdown to plain text, remove URLs, and filter out comments shorter than 50 characters. This last step removes comments that are slurs, spam, low-effort, or phatic, and grants a stronger signal for downstream text analyses.

4.3 Selecting Subreddits

I limit this study to English language subreddits. This is due to current limitations in text analysis methods that would make cross-lingual comparison difficult; I settle on English as the most common language in the subreddits under study. I estimate the dominant language in a subreddit following a semi-automated approach. I gathered candidate subreddits which self-report English as their primary language through the Reddit API, and combine them with subreddits that have the majority of comments in English, as computationally inferred using `langdetect`. Then, I manually removed from the list of candidate subreddits those whose front page was not in English. To enable robust comparisons between approved and removed comments, I only consider subreddits who have at least 100 approved and 100 removed comments. In all, the final dataset for this paper contains 49 subreddits.

4.4 Annotating Automated Moderators

Human moderators make use of several scripts or bots to facilitate their work, and several of those bots directly perform moderation actions that are recorded in modlogs. I do not remove bot moderators because they supposedly enforce rules that human moderators would have enforced themselves. However I consider bot moderators explicitly in the analyses to avoid confounders.

I employ a semi-automated approach also to identify bot moderators. I start with a list of known bots. To this list I add the names of moderators that contain the substring “bot” (e.g. CryptoModBot) as this is a common naming convention on Reddit. Then, I add to the list moderators that show a suspiciously high performance in the modlogs, according to two heuristics: high volume of moderation actions, and number of subsequent moderation actions that are performed less than a second apart. After manual verification, the list of bots includes 438 accounts.

5 Ethical Considerations

Before moving on to detailing methods and results, I will discuss some important aspects that shaped this work. I follow standard ethical guidelines (Vitak, Shilton, and Ashktorab 2016; Zimmer and Kinder-Kurlanda 2017)—I do not attempt to de-anonymize users, I do not link them across platforms, etc. This study was exempt from IRB approval at my institution because I work with publicly available data. Yet, it is important to consider the premises of the open modlog initiative that are ethically challenging. I will go over three major points: user expectations on data access, data persistency, and data ownership.

First, the open modlog initiative explicitly aims at making moderation data publicly accessible, and is both voluntary and opt-in. Though, it is the moderators who enroll the subreddit in the initiative, and therefore it is worth reflecting on whether the users are aware that their information may be re-shared. The open moderation bot is visible to all users in the list of moderators; subreddits also advertise open moderation in their sidebar description; furthermore, subreddits typically engage with users asking them for feedback on the subreddits’ own open moderation policy. Thus, users should be aware of taking part in the open modlogs initiative.

Another potentially contentious issue is that modlogs maintain information available for 90 days. Although I do not know the motivation for this timeboxed access restriction, one must consider that releasing this dataset would extend access to the modlogs. I accumulated modlogs for a long period of time, and I would be making them accessible for analysis farther in the future. Several other independent sources persist modlogs, therefore releasing this dataset would not facilitate nefarious uses or enable new ones.

---

6From https://www.reddit.com/r/autowikibot/wiki/redditbots
On the other hand, I believe that such a longitudinal and persistent dataset constitutes a valuable asset for research. Yet, it is all-important to persist only essential information that I cannot foresee being misused.

Most importantly, modlogs capture content that was found unacceptable by the quality standard of the subreddit it was posted to—content whose author may not want to be associated with, as well as content that is potentially harmful for others, such as doxing attempts. I believe that the people associated with this dataset should be granted the right to be forgotten. I intend to minimize the negative implications of sharing this dataset while enabling future research. Therefore I only distribute the identifiers associated with content, so that authors have the option to remove the content itself from the platform.

6 What Moderators Accept

We have a general understanding of what content moderators remove (Chandrasekharan et al. 2018). But what do moderators approve? I gain insight through a mixed-methods exploration of the textual content of approved comments. Then, I compare and contrast approved comments with two other classes of comments: removed comments, and comments that have never been reviewed by a moderator. I will refer to the latter class as neverReviewed.

6.1 The Content of Approved Comments

Methods I explore the content of approved comments by surfacing latent topics. Specifically, I aim to surface topics that are shared across communities. I thus forego off-the-shelf topic modeling techniques like LDA, which would be biased towards topics representing communities that are large or of that adopt an idiomatic lingo. For example, r/conspiracy’s modlog is 100 times larger than r/neutralnews’s, which makes it more likely that conspiracy theories would percolate in the computationally-derived topics than news reports. By the same token, communities with different background like r/KotakuInAction and r/socialism address the common topic of corruption using different terms, the former using words common in gaming culture and the latter using words common in political rhetoric. It is important that the topics are not confounded by these surface-level lexical differences.

I modify the topic model introduced in (Demszky et al. 2019), which combines sampling and vector spaces to encourage topics that are independent of subreddit size or focus. First, I follow an iterative coding procedure to identify subreddit foci, such as gaming or cryptocurrencies. I start from the complete list of subreddits adopting open moderation for better context. Informed by the front page and the self-description text in the sidebar of each subreddit, I iteratively propose codes for the foci inductively, cluster subreddits according to the foci, and criticize discrepancies between subreddits within in each cluster and across clusters. I converge to the foci in figure 1.

Then, I remove content that is duplicated, that has been approved by automated moderators, or that was authored by moderators themselves. Next, I obtain a stratified sample so that each topical focus is equally represented with 1000 comments. This number conciliates a high variety of the comments sampled with the skewed distribution of approved comments per subreddit. Then, I extract a vocabulary used across subreddits with different foci. I preprocess comments by lowercasing, removing punctuation, normalizing whitespace, and stemming using NLTK’s Snowball stemmer. I keep stems that occur at least 10 times in at least two foci. I repeat the vocabulary extraction procedure two times and keep the intersection of the vocabularies, to account for sampling effects. While this procedure may exclude terms that may be informative in the context of a specific subreddit, it enables us to surface general topics of approved content across subreddits. Next, I train GloVe embeddings for the word stems in the vocabulary, and use (Arora, Liang, and Ma 2017)’s approach to create comment embeddings. I compute a weighted average of the word stem embeddings, stack the vectors as the rows of a matrix, and remove the projection onto the matrix’s first principal component. This creates a shared vector space where to compare approved comments from different subreddits. Then, I cluster comment embeddings with k-means using cosine distance. I determine the optimal number of clusters to be 8 by comparing the silhouette score of different clusterings. Intuitively, each cluster corresponds to a group of approved comments that are topically similar. I interpret topics with a further iterative coding procedure. Data-driven topics may be ambiguous and not conform to intuitive categories. Thus, I ground my interpretation on the word stems closest to each centroid, and two samples of random and central comments per topic. Further, my interpretation is informed by previous literature on moderated content and norms on Reddit (Chandrasekharan et al. 2018; Fiesler et al. 2018; Juneja, Subramanian, and Mitra 2020).

Common topics in approved comments I summarize the topics in table 1. Approved comments discuss challenging topics for moderation. I surface three main types of content: heated debate, controversial issues, and a meta-discussion on moderation itself. First, approved comments feature lexical features of heated debate, including argumentation strategies like clarifying the speaker’s intent or challenging the understanding of the interlocutor (topic “argument”), sarcasm and insider jokes (topic “sarcasm”), all the way to profanities and accusations (topic “swear”). Research on moderation on Reddit found that non-inclusive sarcasm, personal attacks and swearing, and confrontational attitudes like mocking the interlocutor’s sensitivity are commonly found also in removed comments (Chandrasekharan et al. 2018; Juneja, Subramanian, and Mitra 2020). This goes to show that moderation decisions require well-pondered decisions on what fringe content crosses community norm boundaries, especially in the presence of content that looks adversarial at face value.

A second type of approved content discusses controversial issues, such as racial and gender identity, legality, economy, political ideology and media corruption (topics “iden-
hot-button topics, is marked by the framing devices of producer
the smoke trail of a heated debate. Approved content treats
product of large or specialized communities.

This corroborates that the topics surfaced are not the byprod-

lence of a topic across foci never exceeds a Gini index of 0.3.

its content discussing a single topic. The difference of preva-

tributed across subreddit foci. No focus has the majority of

with no single topic making up more than 16% of all

their distribution. Topics cover the dataset homogeneously,
in approved comments. Thus, I check for potential skews in

ments both contain hard cases for moderation.

2020). I show, once again, that approved and removed com-

drasekharan et al. 2018; Juneja, Subramanian, and Mitra

tent antagonizing moderators is a frequent moderation prac-

moderators and call out abuses of power. Removing con-

thus have a clear interest in the matter. Nevertheless, ap-

study, which participate in open moderation practices and

meta-topic is especially relevant to the subreddits under

cusses censorship and, especially, moderation itself. This

of speech.

The positive outcome of moderation: guaranteeing freedom

ty,” “justice,” “economy,” “ideology” respectively). Such

issues are likely to elicit emotional responses, and thus are

commonly used for trolling (Hardaker 2013). Discerning if

a user is being provocative because of their genuine beliefs

or as an attempt enrage others is a challenge for moder-

ation. Though I expected controversial issues in removed

comments, their presence among accepted comments shows

the positive outcome of moderation: guaranteeing freedom

of speech.

In fact, I find a third type of approved content that dis-

cusses censorship and, especially, moderation itself. This

meta-topic is especially relevant to the subreddits under

study, which participate in open moderation practices and

thus have a clear interest in the matter. Nevertheless, ap-

proved comments on this topic challenge the authority of

moderators and call out abuses of power. Removing con-

tent antagonizing moderators is a frequent moderation prac-

tice with problematic repercussions for transparency (Chan-
drasekharan et al. 2018; Juneja, Subramanian, and Mitra

2020). I show, once again, that approved and removed com-

ments both contain hard cases for moderation.

My goal was to surface topics that are commonly found

in approved comments. Thus, I check for potential skews in

their distribution. Topics cover the dataset homogeneously,

with no single topic making up more than 16% of all

approved comments. Topics are also homogeneously dis-

tributed across subreddit foci. No focus has the majority of

its content discussing a single topic. The difference of preva-

lence of a topic across foci never exceeds a Gini index of 0.3.

This corroborates that the topics surfaced are not the byprod-

uct of large or specialized communities.

To summarize, the content that moderators approve is

the smoke trail of a heated debate. Approved content treats

hot-button topics, is marked by the framing devices of po-

lite and less-than-polite arguments, and circles back onto it-

self through a meta-discourse on censorship and freedom of

speech. This mirrors the hallmarks of removed content, in-

cluding swearing, sarcasm, and challenges to the authority—

especially that of online platforms and of their moderators.

Whereas this section offers an exploratory analysis of ap-

proved comments, I turn to drawing empirical difference

with removed and never_reviewed comments next.

6.2 Approved, Removed, Never Reviewed

How do approved and removed content differ, if at all? More

specifically, approved content does not look as innocuous as

its moniker would suggest. How does approved content dif-

fer from the remaining content in the community, i.e. con-

tent that has never been explicitly reviewed by moderators?

I answer these questions comparing approved and removed

comments along several linguistic dimensions that capture
toxicity, prose quality, psycholinguistic categories, and se-

cmantic similarity, together with how well the community re-

ceives the comments. To establish baseline values for each

linguistic dimension, I sample from each subreddit a control

group of never_reviewed comments that have never been re-

viewed by moderators.

Methods I introduce the linguistic dimensions that I use to
describe moderated content, before discussing procedures to
determine the statistical significance of outcomes.

Toxicity I study how toxic the three classes of comments are

for their communities. Toxic language, being contextual

and dependent on the sensibilities of the individuals involved

(Chang, Cheng, and Danescu-Niculescu-Mizil 2020), es-

capes crisp academic definitions. Nevertheless, several tools

for measuring toxicity achieve high accuracy and correla-

Table 1: Topics of approved comments that are common across subreddits. I report the intuitive name for the topic, as well as the prevalence in the dataset in parentheses. I also report the word stems closest to the topic centroid, and example comments in each topic (redacted for anonymity). The last column shows the subreddit overall foci in which each topic is most prevalent.
tion with human assessments. I use one such state-of-the-art measure, the toxicity score provided by Jigsaw’s Perspective API, to determine the relative toxicity of approved, removed, and never_reviewed comments. In this context, toxic is defined as “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.” High values represent a higher likelihood that a comment is toxic.  

**Prose quality** The complexity of a piece of text is an indication of the quality of its prose. Composing highly sophisticated prose is a task that requires effort from its author. Conversely, adversarial text written on impulse is often syntactically simple. I estimate complexity via the Gunning-Fog index, which estimates the years of formal education a person needs to understand the text on the first reading. For example, a Gunning-Fog index of 12 requires the reading level of a United States high school senior.

**Psycholinguistic categories** The advantage of toxicity is that it is a succinct measure of “unlikeable” comments. However, comments can be unlikeable for several reasons. I unpack this variety through psycholinguistic categories that provide a more nuanced picture of the discourse choices, emotions, and psychological states expressed in the comments. In particular, I focus on a subset of the LIWC lexicons related to toxicity. I include swear words and sexual categories that reflect lexical choices in blatantly offensive use of language. Since toxic language often sports heightened emotionality, I use positive and negative emotions to capture the overall sentiment of the comments, and I further break down negative emotions into its components of sadness, anger, and anxiety, to better understand which emotions are elicited by each class of comments. Beyond sentiment and emotions, I also look at discourse features. I disentangle emotional polarity from positive and negative attitudes in argumentation through the categories assert and negation. To the same effect, I include personal pronouns that have been correlated with argumentation—e.g., you is often associated with accusations, whereas we has been associated with frames of in-group belonging. I disentangle the use of pronouns with general stylistic choices through function word categories. While function words (such as, pronouns, prepositions, articles, conjunctions) denote linguistic style, “content words” (such as nouns and regular verbs) indicate the informative content of a text. Lexicons cannot account for context-dependent use of language: though this a limitation to be acknowledged, it helps us measure across subreddit domains. Furthermore, though lexicons are simplistic measures of abstract constructs, their simplicity lends them easily interpretable.

**Semantic similarity** Next, I investigate how the different classes of comments differ in communicated content. To

---

8Perspective API offers also alternative scores such as “severe toxicity,” which accounts not just for the probability of a comment of being toxic but also for the extent of its toxicity. I replicated the analyses using severe toxicity and found no major differences.  

measure the semantic similarity between them, I make use of state-of-the-art transformer architecture for sentence embeddings. In particular, I use the distilBERT base model trained on the NLI and STS tasks. This model, despite being orders of magnitude smaller than BERT large, achieves similar results on the STS benchmark explicitly designed for assessing semantic text similarity. 

**Community feedback** Finally, I look beyond the content of the comments and gauge how well received they are from their respective community. Reddit lets its users vote on content as a measure of appreciation. The Reddit score summarizes this feedback into a single measure, which can be arbitrarily positive or negative.

**Statistical methods** Community norms vary across subreddits: what is accepted and even praised in one, may be sanctioned in another. I account for these contextual differences and make measurements comparable by rescaling. First, I remove content that is duplicated, that has been approved by automated, or that was authored by moderators, to avoid confounders. Then, for each subreddit, I compute the z-score of each measure. Hence, measures can be interpreted as the standardized deviance from the mean, with respect to typical values in the subreddit of origin. As an example, saying that approved comments have a z-transformed toxicity score of 0.4 means that they have values 0.4 standard deviations higher than the population average in their respective subreddits. For fair comparison between large and small subreddits, I sample an equal number of comments per subreddit and per comment type. I set this number to 500, as it strikes a good balance between the distributions of the data and the size needed for meaningful statistical tests. Then, I assess differences with non-parametric tests, specifically Kruskal-Wallis’s for differences in medians, and Dunn’s multiple comparison with Bonferroni correction for post-hoc pairwise tests. I use non-parametric tests because most variables are not normal or not homoscedastic, as per Shapiro-Wilk’s and Levene’s test. Unless otherwise stated, I take 0.01 as a critical value for statistical significance.

Because I encode semantic content as vectors, instead of single measures, I take a different approach to compute semantic similarity. For each approved comment, I take the closest removed and never_reviewed comments from the same subreddit as determined by cosine similarity. Then, for each subreddit, I assess whether removed or never_reviewed comments are closest to approved comments on average via Wilcoxon’s signed-rank test. I also repeat the test via bootstrapping, sampling random triplets of accepted, removed, and never_reviewed comments.

**Approved comments are similar to removed comments** Figure 2 illustrates the differences between approved, removed, and never_reviewed comments.

The trend across nearly all linguistic categories is that approved comments depart from the values of never_reviewed

---

9Using BERT showed no difference in the results. I report distilBERT results because they are more easily replicated.
Figure 2: Differences in the language and community feedback between approved, removed, and never reviewed comments. Z-scores are computed per-subreddit.

Table 1: LIWC analysis of approved, removed, and never reviewed comments.

<table>
<thead>
<tr>
<th>Category</th>
<th>Approved</th>
<th>Never Reviewed</th>
<th>Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIWC_assent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_negate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_posemo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_negemo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_anx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_anger</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_sad</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_sexual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_swear</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_shehe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_you</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_we</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_they</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_i</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIWC_funct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>toxicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GunningFogIndex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>score_praw</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Approved comments break this general trend in a few categories. They are less self-referential (use of “I” and “she/he” pronouns) than the other two classes of comments. They also asssent less and show fewer positive emotion words, while showing more convoluted prose (as per Gunning-Fog Index). Cross-referencing the topics of approved comments in section 6.1, this suggests that approved comments may be less about disclosures of the personal experiences of the speaker, and instead put forth more elaborate argumentation. In fact, the community feedback that approved comments receive is the highest, a sign of quality of their contribution.

Next, I turn to semantic similarity between the three classes of comments. Contrary to my hypotheses, I do not find consistent differences across subreddits, neither selecting the most similar removed and never reviewed comments, nor selecting them at random. Two thirds of subreddits show higher similarity between approved and never reviewed comments than to removed comments, when selecting the most similar to approved comments. This fraction decreases to one half when selecting removed and never reviewed comments at random. In other words, although there exist good matches for approved comments in both other classes, this similarity is dependent on the context of the community.

In a nutshell, what undergoes review by moderators shares many of the same characteristics, regardless of whether it is ultimately approved or removed. In fact, approved comments mirror several qualities of removed comments, such as higher toxicity and anger than never reviewed comments.

7 The Effect of Neglecting Accepted Comments in Automated Moderation

The previous sections show how approved and removed comments are altogether different from content that never undergoes moderation. This finding goes against common knowledge, in that long-standing lines of research in computer-mediated communication and natural language processing focus on negative moderation outcomes, and therefore model moderation as the problem of identifying content to be removed. These premises are encoded in computational tools aimed at aiding human moderators, which are trained to distinguish removed comments from an under-specified class of “other” comments—that are never reviewed. In technical terms, automated moderation coerces a one-class problem (identifying removed comments) into a two-class problem (distinguishing removed from never reviewed comments). I challenge this two-class formulation, and argue that the real task of moderators is to referee comments that are ultimately approved or removed.

To what extent do automated moderation tools actually replicate both decisions of human moderators, to approve or to remove comments?

7.1 Real Life Automated Moderation

I test this by replicating automated moderation pipelines, training them on the traditional removed vs. never reviewed problem, and testing their performance on the new removed vs. approved problem.

Methods First, I assess the performance of tools that use never reviewed comments as their positive class, and compare and contrast the results when using approved comments instead. I set up a standard classification pipeline. For each
Inflated performance in automated moderation

Figure 3 summarizes the results. The right side of the chart shows the performance of models trained on neverReviewed comments. One can see, in the first box, that they perform remarkably well, with a median in-domain accuracy of 74%. However, their performance drops significantly to 58% on the real-life task of distinguishing between approved and removed comments. This is lower than the in-domain accuracy of models trained on approved comments, with a median of 66%. This tells us on the one hand, that the task of distinguishing removed and approved comments is intrinsically harder than distinguishing removed and neverReviewed comments. On the other hand, this shows that models trained on neverReviewed comments do not generalize to the actual decisions of human moderators.

Conversely, models trained on approved comments generalize comparatively better, with a median 60% out-of-domain accuracy. In other words, approved comments carry more information about neverReviewed comments than the converse. More in detail, I look at the fraction of approved comments that are correctly identified by models trained on neverReviewed comments (last box on the right). One can see that they correctly classify only 46% of approved comments. Given that the classification setup is balanced and that the random-chance recall would be 50%, this means that models trained on neverReviewed comments systematically misclassify approved comments. For comparison, models trained on approved comments identify correctly 59% of the neverReviewed comments.

In short, traditional automated moderation sets out to solve a simpler problem than the one human moderators face. Classifiers that distinguish removed from neverReviewed comments boast an inflated accuracy on the simpler problem, but underperform when replicating real moderation decisions. In particular, computational models systematically remove content that human moderators would approve.

7.2 When Automated Moderation Fails

Next, I analyze under which circumstances automated moderation misclassifies approved comments.

Methods

When would automated moderation tools remove comments that human moderators approved? I look more closely at the classification outcomes of models trained on neverReviewed comments. To do so, I train a further logistic regression classifier that predicts whether an approved comment would be correctly classified. I use all of the features introduced in section 6.2, following the same procedure for imputation of null values and standardization. Moreover, I introduce an indicator variable for approvals performed by automated moderators instead of humans.

Automated moderation conflates controversial with toxic

Which characteristics of approved content mislead automated moderation tools, when trained on neverReviewed comments? I unpack model errors by looking at the characteristics of approved content, and by correlating them with successful and unsuccessful classification.

Figure 4 reports the coefficients for the logistic regression. One can see that models are more likely to correctly classify approved comments, when the comments were approved by deployed automated moderation systems. In other words, automated moderation systems would reinforce decisions of other automated moderators. Similarly, models correctly approve content that sports more elaborated prose and that is better received by the community (“Gunning-FogIndex,” “LIWC_func,” “LIWC_negate,” “pawr_score”). Conversely, the models incorrectly remove comments that are less clearly innocuous, such as when they use toxic language, swear words, sexual references, accusations, othering language, and self-references (“toxicity,” “LIWC_swear,” “LIWC_sexual,” “LIWC_you,” “LIWC_they,” “LIWC_i”).

Thus, automated moderation tools trained on neverReviewed comments are successful at identifying...
reviewed comments be-
reviewed
reviewed comments selected multiple
they are
confounded by the hallmarks of adversarial
effectively deployed in the communities. However, they
comments approved by the other automated moderators
proved comments. I sample never
alternative content that could serve as a stand-in for ap-
ternatives do when approved comments are not available?
2018), but approved content leaves no trace. What can prac-
tition decisions. Several approaches exist to reverse-engineer
employees of the discussion platforms) have access to modera-
sider approved content. However, in practice, this is unreal-
ta tion. Second, the performance of automated moderators is
platforms should carefully consider their problem formula-
sion making process of human moderation. Practitioners and
mation. Finally, the error analysis shows the
highlight the importance of appropriate benchmarks for au-
shortcoming of simplistic moderation strategies, like resort-
language like toxic comments and swear words, which
they systematically remove as blatant offenses.
This offers grounds for reflection. First, the problem for-
mulation in automated moderation does not reflect the deci-
sion making process of human moderation. Practitioners and
platforms should carefully consider their problem formulation.
Second, the performance of automated moderators is
infated with respect to their real-world application, which
highlights the importance of appropriate benchmarks for au-
tomated moderation. Finally, the error analysis shows the
shortcoming of simplistic moderation strategies, like resort-
ing to word blacklists and third-party toxicity scores, even
when accounting for contextual norms.

8 Replicating (Hidden) Moderator Decisions
The straight-forward fix for the problems outlined above is
to reformulate automated moderation tasks to explicitly con-
sider approved content. However, in practice, this is unreal-
estic. By and large, only moderators (in most cases, only em-
ployees of the discussion platforms) have access to modera-
tion decisions. Several approaches exist to reverse-engineer
what content has been removed (Chandrasekharan et al.
2018), but approved content leaves no trace. What can prac-
titioners do when approved comments are not available?
Leveraging the insights of these analyses, I look for al-
heuristics: never_reviewed comments that 1) are authored by
users that have had their comments removed, 2) are highly
toxic, and 3) that have been ill-received by their commu-
nity and thus have low scores. Section 6.2 sheds further light on the relationship between approved, removed, and
never_reviewed comments, and shows that indeed approved
comments have the highest scores among the three. I then
sample 4) never_reviewed comments with high scores. Fi-
ally, section 7.2 shows how most approved comments are
misclassified as removed. Taking inspiration from adver-
sarial learning, I sample 5) never_reviewed comments that
are semantically similar to removed comments. Intuitively,
those comments would be confusing examples (they lay
close to the classification margin, at a small distance from
comments of the opposite class), and introducing them in
training would make for more robust models.

Methods I clarify how I identify comments that are simi-
lar to removed comments. Then, I detail the classification
pipeline used to compare the different heuristics.

Finding comments most similar to removed I describe
the procedure for finding similar comments to removed, be-
cause the orders-of-magnitude difference between removed
and never_reviewed comments calls for some technicalities.
In particular, I do not use all never_reviewed comments be-
cause of the computational and memory demands of this
matching process. Instead, for each subreddit, I start with
a sample of never_reviewed comments 10 times larger than
the number of removed comments. This larger sample better
accounts for the variety of never_reviewed comments, while
demanding a more manageable amount of resources. Next, I
encode both removed and never_reviewed comments as sen-
tence embeddings as I did in section 6.2. For each removed
comment, I keep the n most cosine-similar never_reviewed
comments, starting with n = 10. Then, I solve an opti-
mization problem to guarantee high-quality 1-to-1 match-
ing without duplicate never_reviewed comments. This is to
avoid that few never_reviewed comments selected multiple
times would affect the downstream analyses. In particular, I
solve the integer linear program:

\[
\max \left( \sum_{i, j} \text{sim}_{ij} X_{ij} \right) \\
\text{s.t.} \\
\sum_{j} X_{ij} = 1 \quad \forall i \in 1, \ldots, k \\
\sum_{X_{ij}} \text{instanceof } \chi_a \leq 1 \quad \forall a \in 1, \ldots, |\chi| 
\]

Where \( \chi \) is the set of unique never_reviewed comments
that are among the n most similar for at least one removed
comment. \( \text{sim} \) is the \( k \times n \) matrix of cosine similarities,
where \( k \) is the number of removed comments. \( X_{ij} \) is an
indicator variable for the \( j \)-th most similar never_reviewed
comment to the \( i \)-th removed comment, where \( X_{ij} \) is associ-
ated with exactly one comment in \( \chi \). The objective function
maximizes the overall similarity of the matched comments
in the solution. The first set of constraints selects exactly one

Figure 4: Coefficients of the model predicting if an approved
comment is correctly classified (pseudo-\( R^2 = .05 \)).
Figure 5: Performance of heuristic sampling strategies in discerning removed and approved comments (left), and in retrieving approved comments (right). The dashed line in grey reports the baseline strategy of using never_reviewed comments.

never_reviewed comment per removed comment. The last set of constraints makes it so that each never_reviewed comment is selected at most once. If the program results infeasible, e.g., because there are not enough unique never_reviewed comments in $\chi$ to cover all removed comments, I increase $n$ by doubling it.

**Classification pipeline** I test the effectiveness of the different heuristics on how well they perform on the task of classifying approved and removed comments. I adopt a setup similar to that in section 7.1. For each subreddit, I sample comments according to each heuristic, as well as an equal number of removed comments. I split the sample into training and test sets, in 9:1 proportions. I train a similar logistic regression classifier using TF-IDF features. Then, however, I modify the test set: I replace the comments sampled with the heuristic with a random sample of approved comments. I test classifier performance on this modified test set, measuring its overall accuracy, and recall for approved comments. I repeat the procedure 10 times for stabilizing results.

**In the absence of approved comments, take the ones with the worst scores** Figure 5 shows how different heuristics perform on the task of distinguishing approved and removed comments. It is immediately apparent that sampling toxic never_reviewed comments is the worst-performing strategy (the fourth box, in red in the figures). Toxic comments are similar to removed comments and thus confuse the classifier. Indeed, the most toxic never_reviewed comments may be comments that break community norms and would be removed by moderators if reviewed. Among the other strategies, sampling comments from past norm-breaking offenders is the least accurate strategy (the first box, in blue), with a lower median accuracy than the baseline of sampling random comments (the grey dashed line). Surprisingly, sampling comments with the top or worst Reddit score yield similar results, the former showing lower recall. The best performing strategies are to substitute approved comments with comments with the worst score, or with comments that are most similar to removed comments. Both strategies outperform sampling comments at random in correctly classifying approved comments (a median of 54 and 53% respectively vs. 46% of never_reviewed comments). At the same time, both strategies offer a similar median accuracy of 59% (vs. 58% of never_reviewed comments).

9 Discussion

**9.1 Implications for Research: Abusive Language Detection vs. Automated Moderation**

Abusive language detection and automated moderation share the common goal of identifying content to be removed. Yet, although abusive language detection is undeniably a major challenge for moderation (automated or otherwise), moderation requires considerations on several contextual factors and entails a broader set of practices than removing abusive language (Gilbert 2020). Though it is tempting to conflate the two fields, it is important to tease apart their differences and to identify the grounds on which they can truly be of mutual use.

Due to the variety and nuance of abusive language, research struggles to find general definitions for it (Salminen et al. 2018). Even the act of labeling language as abusive is in fact contentious (Gorwa, Binns, and Katzenbach 2020). Moderation research turns these challenges from deductive to inductive, by grounding them in the norms and practices of specific communities: abusive is that which has been labeled so by a moderator. Moderation of online communities can therefore offer large-scale, high-quality ground truth for abusive language detection. To meaningfully do so, though, it needs to convey the broader scope and more specific context of moderation decisions. Positive moderation decisions offer a useful avenue to reflect on how to do so.

Most datasets for abusive content detection do not provide positive annotation instructions for what content is acceptable, and resort instead to the derivative category of “all that is not abusive” (e.g., (Davidson et al. 2017; Wulczyn, Thain, and Dixon 2016; Chatzakou et al. 2017; Zampieri et al. 2016)).
Even when datasets include ground truth about moderation, they rarely include positive moderation decisions, e.g., (Cheng et al. 2017; Chandrasekharan et al. 2017, 2019). I clarify the pitfalls of ignoring approved content in abusive language detection systems meant for automated moderation. Though abusive language offers some indication of potentially problematic content, not all removed content is offensive. I show, similarly, how not all approved content is clearly inoffensive (like most never_reviewed content), and deciding what is admissible requires nuance and knowledge of its full context. Ignoring approved content results not only in misguiding performance measurements, but also in automated moderation systems that do not work in practice across a diverse set of communities.

Going forward, research should take into account both negative and positive outcomes of moderation decisions in task formulations, data collection, and benchmark infrastructure. For research on moderation, this would result in a better understanding of community norms and practices. In turn, such resources would be valuable for the study of abusive language. In fact, the recent discourse within the NLP research community calls for shifting the focus from content removal to supporting equitable participation online, especially through identifying what content should be allowed (Jurgens, Chandrasekharan, and Hemphill 2019). To this end, the present work offers insights on how to reframe the task of abusive language detection as one of emulating positive and negative moderation decision.

9.2 Implications for Practitioners: The Effects of Positive Outcomes on Automated Moderation

False positives—instances that are erroneously flagged by automated moderators—are one important reason why such systems are not widely deployed. Most platforms lament an underprovision of human moderators, which makes it paramount to minimize false positives to maximize the impact of human effort. One can consider approved content as false positives, and reinterpret results in this light.

Should automated systems be trained on removed content and content that was never_reviewed by moderator, considering approved content as the necessary side effect of false positives? Or, should approved content be included as part of model training, making the models learn from its mistakes? With this work, I offer a realistic benchmark for the performance of automated moderation systems, and my results indicate the latter as the better option. Compared to only learning from never_reviewed content, training on approved content reduces false positives by half. This is due to two reasons. Approved and never_reviewed content appear qualitatively different. Furthermore, automated moderation leverages this difference asymmetrically: never_reviewed content does not help reduce false positives from those produced via random guessing, whereas approved content helps correctly classify never_reviewed content.

Being inaccessible for public scrutiny, it is unclear if proprietary automated moderation systems leverage approved content and how. These systems appear to fall under one of two categories: machine learning-backed models, and pattern matching approaches for comparing against a blacklist or rule set (Gorwa, Binns, and Katzenbach 2020). Approved content can help audit such automated moderation systems and help refine them. This study shows how approved content, due to its similarity to removed content, challenges both pattern-matching blacklists and machine learning-backed toxicity classifiers.

10 Limitations

One could argue that approved content intrinsically surpasses the discriminative capabilities of automated moderation, and therefore should undergo human review. Although I do not address this alternative view, I show that approved content is often controversial. Therefore I argue that conflating approved and never_reviewed content would be problematic also in this framing. Future research can explore more general experimental designs, such as multiclass or cascading classifiers. Furthermore, this work relies on a limited set of subreddits that, though internally diverse in many aspects, is not representative of the variety of moderation practices on Reddit, and includes only communities that self-selected for open moderation.

11 Conclusions

This paper gave an overview of content that moderators approve. Approved content carries the hallmarks of heated conversations, treats controversial topics, and challenges moderation. Through a fine-grained comparative analysis, I found that approved content departs from what is generally found in a community and similar to removed content along several linguistic dimensions, including toxicity and blatant insults. Given the commonplace practice of training moderation classifiers on comments that are potentially never reviewed by moderators, I quantified the errors of those classifiers on actual moderator decision. I discovered that moderation classifiers would systematically remove over half of the content that moderators approved. In particular, through an analysis of model errors, I showed that classifiers are confounded by content that at a surface level appears offensive, although it is acceptable by community norms. Although I showed the advantages of including approved comments for developing automated moderation tools, I acknowledge that approved comments are often inaccessible to researchers and practitioners. I addressed this in two ways. First, informed by the novel results I tested several sampling heuristics to make-do when approved comments are unavailable, and showed that substituting random comments for comments with low score improves performance. Second, I release a gold-standard, large-scale, complete dataset of approved comments spanning 49 communities over 2 years, to enable further research and better support for community moderators. I do not imply that automated moderators should sanction autonomously: flagging systems equally benefit from this framing.

I clarified the pitfalls of ignoring approved content in abusive language detection systems meant for automated moderation. Though abusive language offers some indication of potentially problematic content, not all removed content is offensive. I show, similarly, how not all approved content is clearly inoffensive (like most never_reviewed content), and deciding what is admissible requires nuance and knowledge of its full context. Ignoring approved content results not only in misguiding performance measurements, but also in automated moderation systems that do not work in practice across a diverse set of communities.

Going forward, research should take into account both negative and positive outcomes of moderation decisions in task formulations, data collection, and benchmark infrastructure. For research on moderation, this would result in a better understanding of community norms and practices. In turn, such resources would be valuable for the study of abusive language. In fact, the recent discourse within the NLP research community calls for shifting the focus from content removal to supporting equitable participation online, especially through identifying what content should be allowed (Jurgens, Chandrasekharan, and Hemphill 2019). To this end, the present work offers insights on how to reframe the task of abusive language detection as one of emulating positive and negative moderation decision.

9.2 Implications for Practitioners: The Effects of Positive Outcomes on Automated Moderation

False positives—instances that are erroneously flagged by automated moderators—are one important reason why such systems are not widely deployed. Most platforms lament an underprovision of human moderators, which makes it paramount to minimize false positives to maximize the impact of human effort. One can consider approved content as false positives, and reinterpret results in this light.

Should automated systems be trained on removed content and content that was never_reviewed by moderator, considering approved content as the necessary side effect of false positives? Or, should approved content be included as part of model training, making the models learn from its mistakes? With this work, I offer a realistic benchmark for the performance of automated moderation systems, and my results indicate the latter as the better option. Compared to only learning from never_reviewed content, training on approved content reduces false positives by half. This is due to two reasons. Approved and never_reviewed content appear qualitatively different. Furthermore, automated moderation leverages this difference asymmetrically: never_reviewed content does not help reduce false positives from those produced via random guessing, whereas approved content helps correctly classify never_reviewed content.

Being inaccessible for public scrutiny, it is unclear if proprietary automated moderation systems leverage approved content and how. These systems appear to fall under one of two categories: machine learning-backed models, and pattern matching approaches for comparing against a blacklist or rule set (Gorwa, Binns, and Katzenbach 2020). Approved content can help audit such automated moderation systems and help refine them. This study shows how approved content, due to its similarity to removed content, challenges both pattern-matching blacklists and machine learning-backed toxicity classifiers.

10 Limitations

One could argue that approved content intrinsically surpasses the discriminative capabilities of automated moderation, and therefore should undergo human review. Although I do not address this alternative view, I show that approved content is often controversial. Therefore I argue that conflating approved and never_reviewed content would be problematic also in this framing. Future research can explore more general experimental designs, such as multiclass or cascading classifiers. Furthermore, this work relies on a limited set of subreddits that, though internally diverse in many aspects, is not representative of the variety of moderation practices on Reddit, and includes only communities that self-selected for open moderation.

11 Conclusions

This paper gave an overview of content that moderators approve. Approved content carries the hallmarks of heated conversations, treats controversial topics, and challenges moderation. Through a fine-grained comparative analysis, I found that approved content departs from what is generally found in a community and similar to removed content along several linguistic dimensions, including toxicity and blatant insults. Given the commonplace practice of training moderation classifiers on comments that are potentially never reviewed by moderators, I quantified the errors of those classifiers on actual moderator decision. I discovered that moderation classifiers would systematically remove over half of the content that moderators approved. In particular, through an analysis of model errors, I showed that classifiers are confounded by content that at a surface level appears offensive, although it is acceptable by community norms. Although I showed the advantages of including approved comments for developing automated moderation tools, I acknowledge that approved comments are often inaccessible to researchers and practitioners. I addressed this in two ways. First, informed by the novel results I tested several sampling heuristics to make-do when approved comments are unavailable, and showed that substituting random comments for comments with low score improves performance. Second, I release a gold-standard, large-scale, complete dataset of approved comments spanning 49 communities over 2 years, to enable further research and better support for community moderators.

I do not imply that automated moderators should sanction autonomously: flagging systems equally benefit from this framing.

12https://github.com/gesicss/modlogs_approved_comments
Acknowledgments
I thank Indira Sen and the anonymous reviewers for their helpful comments.

References
Arora, S.; Liang, Y.; and Ma, T. 2017. A simple but tough-to-beat baseline for sentence embeddings. In ICLR.
Gilbert, S. A. 2020. “I run the world’s largest historical outreach project and it’s on a cesspool of a website.” Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. Proceedings of the ACM CSCW.
Hardaker, C. 2013. “Uh. . . . not to be nitpicky...,but...the past tense of drag is dragged, not drug.”: An overview of trolling strategies. Journal of Language Aggression and Conflict 1(1).
Rajadesingan, A.; Resnick, P.; and Budak, C. 2020. Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits. In ICWSM.
Wadden, D.; August, T.; Li, Q.; and Althoff, T. 2021. The Effect of Moderation on Online Mental Health Conversations. In ICWSM.