Conversational Resilience: Quantifying and Predicting Conversational Outcomes Following Adverse Events

Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan

University of Illinois at Urbana-Champaign
{cjl8, apr5, eshwar}@illinois.edu

Abstract

Online conversations, just like offline ones, are susceptible to influence by bad actors. These users have the capacity to derail neutral or even prosocial discussions through adverse behavior. Moderators and users alike would benefit from more resilient online conversations, i.e., those that can survive the influx of adverse behavior to which many conversations fall victim. In this paper, we examine the notion of conversational resilience: what makes a conversation more or less capable of withstanding an adverse interruption? Working with 11.5M comments from eight mainstream subreddits, we compiled more than 5.8M comment threads (i.e., conversations). Using 239K relevant conversations, we examine how well comment, user, and subreddit characteristics can predict conversational outcomes. More than half of all conversations proceed after the first adverse event. Six out of ten conversations that proceed result in future removals. Comments violating platform-wide norms and those written by authors with a history of norm violations lead to not only more norm violations, but also fewer prosocial outcomes. However, conversations in more populated subreddits and conversations where the first adverse event's author was initially a strong contributor are capable of minimizing future removals and promoting prosocial outcomes after an adverse event. By understanding factors that contribute to conversational resilience we shed light onto what types of behavior can be encouraged to promote prosocial outcomes even in the face of adversity.

Introduction

Online communities of all types are often victims of bad behavior. These types of negative influences are inevitable given the widespread nature of online communities and the anonymity afforded to users (Bernstein et al. 2011). Reddit, for example, is consistently working to abate the spread of bad behavior by imposing sanctions like bans, quarantines, and comment removals (Chandrasekharan et al. 2017). Within communities, volunteer moderators regulate behavior and deal with adversity on a daily basis. A common approach used by moderators to tackle norm violations is comment removals (Jhaver et al. 2019a), and researchers continue to explore ways to assist moderators in identifying norm-violating content (Jhaver et al. 2019b). Such context-sensitive information regarding what is considered to be adverse or unacceptable behavior within a given community is hard to replicate using out-of-domain methods to detect antisocial behavior like toxicity, hate speech, and so on (Jurgens, Chandrasekharan, and Hemphill 2019). As a result, we consider moderator comment removals to be indicators of adverse events within communities. By removing a comment, moderators who enforce and understand the norms within their respective communities have indicated that the comment is in violation of community norms, and therefore an instance of adverse behavior (Chandrasekharan et al. 2018).

Prior work has explored the detection of antisocial behavior (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015), prosocial outcomes (Bao et al. 2021), and early signs of conversational failure (Zhang et al. 2018), but how a conversation is affected by the presence of adverse behavior remains to be explored in detail. For example, what happens to conversations after the first instance of adversity? Are all instances of adverse behavior likely to lead to similar outcomes? In this paper, we introduce the task of predicting from the first instance of adverse behavior in a conversation whether it will end the conversation altogether, lead to more undesirable behavior, or even result in prosocial behavior.

As a motivating example, Figure 1 illustrates the three different ways a Reddit conversation can proceed after the first removal. A branch is said to *proceed* if at least one comment was posted in the conversation after the first removal. While some conversations will not be able to withstand the influx of adverse behavior and will adopt similar adverse behaviors, others will recover from the interruption without falling victim. Through examining online interactions at the fine-grained level of comment threads, it becomes clear that the interruption by an adverse event is not always a death sentence. We define this idea of resisting the influence of adverse events as *conversational resilience*.

We explore the concept of conversational resilience in this paper through the lens of three research questions.

RQ1: What factors contribute to whether a conversation proceeds after the first adverse event?

RQ2: How often does an adverse event lead to further removals in the conversation?

RQ3: What factors encourage prosocial outcomes in conversations despite encountering adverse events?

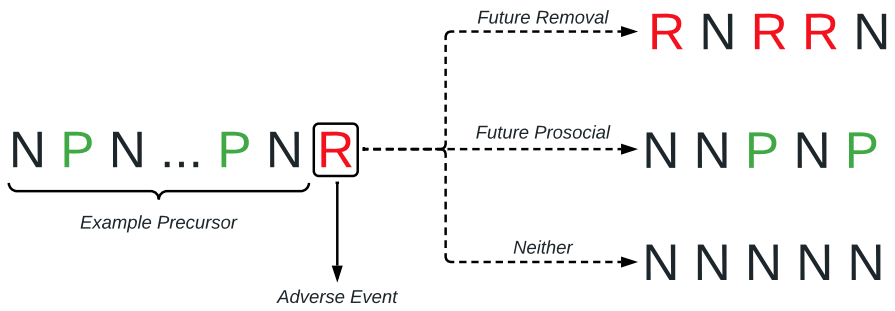

Figure 1: Example conversation thread sequences abstracted to the type of behavior in each comment. In this image, *P* represents a prosocial comment, *R* a comment removal, and *N*, a neutral comment (i.e., neither prosocial nor adverse). Three example outcomes are shown based on the outcome types discussed in the methods section.

The foundational question that must be answered in this research is whether it is possible for conversations to proceed after being interrupted by an adverse event. Conversations that continue after an adverse event might demonstrate the presence of conversational resilience. The next question that must be asked to capture resilience is whether conversations that proceed after an adverse event result in adverse outcomes. Finally, we ask what it takes for a conversation to recover from adversity and lead to prosocial outcomes.

To answer RQ1, we consider all conversation threads with at least one adverse event (i.e., removed comment) and explore factors that may affect whether a conversation proceeds after the event. To answer RQ2 and RQ3, we define three types of conversational outcomes: future removals, future prosocial, and neither (see Figure 1). RQ2 aims to determine what specific factors contribute to a branch ending up with future removals as the conversational outcome. We examine a subset of our data from RQ1 (i.e., only branches that proceed after the first removal), and employ statistical methods to explore factors related to adverse outcomes. We use the same statistical methods to explore correlations between prosocial outcomes and features of events, precursors, authors, and subreddits to answer RQ3. These associations can uncover what types of behavior can be encouraged to promote prosocial outcomes even in the face of adversity.

Related Work

This section will elaborate on pertinent research related to the primary themes of this paper: antisocial behavior, norms and norm violations, and prosocial behavior. These themes are used to define various features used in the statistical analyses intended to answer our research questions.

Antisocial Behavior

There is an abundance of research related to antisocial behavior in online communities (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015; Chandrasekharan et al. 2017; Seering, Kraut, and Dabbish 2017; Chandrasekharan et al. 2018; Zhang et al. 2018; Hessel and Lee 2019; Chang and Danescu-Niculescu-Mizil 2019). Researchers have explored different methods for handling the moderation of Reddit communities with significant amounts of adverse behavior through the use of explicit subreddit bans and quarantines (Chandrasekharan et al. 2017). These researchers explore specifically how effective these interventions are at diminishing adverse behavior, such as hate speech. In other words, such work investigating the efficacy of moderation essentially examines the resilience of bad behavior to community-level interventions.

Content removal. Other research provides support for moderator decisions in online communities by examining the language used in removed content versus non-removed content. For example, Chancellor, Lin, and Choudhury (2016) explore pro-eating disorder Instagram posts and compare their language features with those of posts that were not removed by moderators. By building classifiers to distinguish between these two types of content, they identify fundamental differences between adverse content that was moderated from the rest. This research helps support decisions made in our research, particularly concerning the focus on comment removals as indicators of adversity.

We consider comment removals by community moderators to be one of the most context-sensitive indicators of adversity. In all cases of removed content on Reddit, either a human moderator or an automated moderator[1] applies moderator-generated rules to take down a comment. In both cases, moderators, who are typically also participating members of the subreddit, are dictating what defines a norm violation within their specific community. As a result, comment removals have previously been considered indications of violations of community norms (Chandrasekharan et al. 2018). We use comment removals in much of our statistical analyses to indicate the first instance of adversity within a conversation thread (i.e., *the first adverse event*).

Norms and Norm Violations. Generally speaking, norms are an important part of any online community. Prior work includes predicting how a user will react after being blocked for a norm violation on Wikipedia (Chang and Danescu-Niculescu-Mizil 2019) and an exploration of how moderation tools can discourage antisocial behavior and encourage

[1] Automod is the most popular automated moderation tool on Reddit - https://www.reddit.com/wiki/automoderator

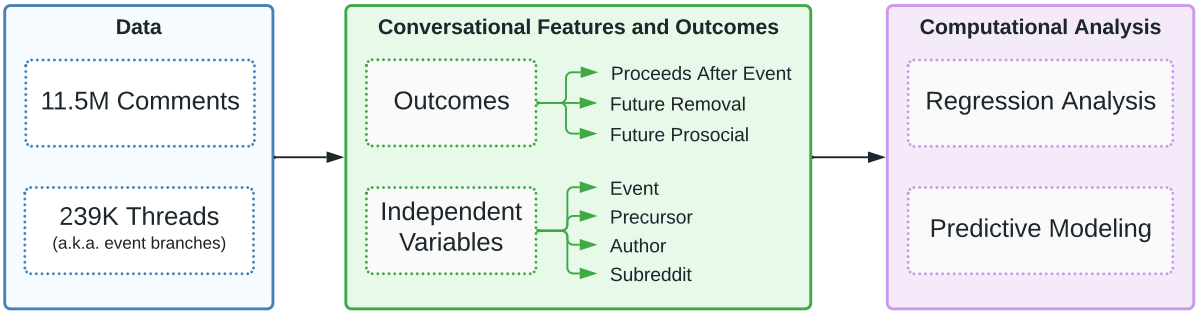

Figure 2: Visualization of the data processing and analysis pipeline. Beginning with data collection and pre-processing, we combine comments into their original conversation thread structure, extract various features and outcomes from the data, and finally statistical analyses to answer the research questions.

prosocial behavior on Twitch (Seering, Kraut, and Dabbish 2017). Chandrasekharan et al. (2018) break down the concept of community norms by defining three separate types of norms: macro, meso, and micro norms. These categories correspond to norms that are Reddit-wide, relevant to some subreddits, and specific to unique subreddits respectively.

In this paper, we apply this understanding of norms and the comment moderation process on Reddit to define adverse behavior. Chandrasekharan et al. (2018) also developed 100 classifiers trained on the comment removals made by 100 subreddits to predict whether each of the studied subreddits would remove a given comment. We employ these classifiers to calculate a *macro norm violation score* for each Reddit comment.[2] This score is defined by the proportion of the 100 classifiers that predict the comment would be removed from its specific subreddit. It allows us to augment micro-level estimates of adversity—comment being removed by a given subreddit's moderators— with a macro-level estimate of adversity—how many other communities would consider this comment to be adverse.

Comment Toxicity. Another common method of evaluating whether a comment represents undesirable behavior is using the Perspective API,[3] which assigns *toxicity scores* to text. These scores indicate what percentage of human annotators would consider a comment toxic (Zhang et al. 2018; Bao et al. 2021). Perspective API returns a score between 0 and 1 for a given comment, where a smaller value indicates that a lower percentage of people would consider the comment toxic. We define a toxicity threshold to indicate the score above which a comment should be flagged as toxic. We employ the thresholds used in Bao et al. (2021): 0.5 to indicate toxic and 0.8 to represent highly toxic comments.

Prosocial Behavior

Our research questions also consider prosocial behavior, another important component of online social interactions (Danescu-Niculescu-Mizil et al. 2013; Seering, Kraut, and Dabbish 2017; Bao et al. 2021). One interesting related work is the research done by Danescu-Niculescu-Mizil et al. (2013) to identify linguistic features of politeness and build a model to classify Wikipedia talk page and StackExchange data on various components of politeness. Their work explores how politeness can often be a decisive factor in whether a social interaction will go poorly, which ties back to the idea that prosocial behavior may help conversations be more resilient in the face of adversity. Bao et al. (2021) define a task to predict whether an online conversation will have a prosocial outcome based on the first few comments of the conversation thread. Their research identifies many categories of prosocial behavior, providing six quantitative metrics for us to use in order to assess how prosocial a conversation is. These metrics come in two categories: lexicon-based counts and BERT scores.

Lexicon-Based Counts. Bao et al. (2021) identify prosocial behavior using raw counts of three types of events. The first type of event is a *donation* event which counts the number of fundraising URLs present in the dataset. The second metric is for *gratitude*, which counts the number of gratitude words in a comment. Gratitude words include "thank you" and other words/phrases with similar intention. The final metric pertains to *laughter* and uses a lexicon of common laughter words (e.g., "haha") to count laughter events. We consider a comment to be prosocial if it contains at least one event out of any of the three types described above.

BERT Scores. Bao et al. (2021) provide models to calculate scores in three categories: *agreement*, *politeness*, and *support*. Scores range from 1 to 5, where 5 indicates a strong presence of a given category. While a score of 1 represents disagreement, impoliteness, and un-supportiveness in each category respectively. A score of 3 indicates neither extreme is present. We consider a comment to be prosocial if it has a score of 4 or more in at least one of the three categories.

Methods

Next we describe our data collection process, report relevant descriptive statistics relating to the data, and detail the analysis methods. An overview of this process is visualized in Figure 2. Additionally, this section explains the computation

[2]Note that prior work referred to this as an "agreement score." We use "macro norm violation score" instead to avoid confusion with the agreement BERT scores described in subsequent sections.

[3]https://www.perspectiveapi.com/

Subreddit	Total Comments	Removed Comments	Total Branches	Event Branches	Avg. Length
r/Games	480,246	55,071	209,236	21,904	3.11
r/legaladvice	578,014	18,716	225,921	10,650	2.79
r/books	589,468	9,795	372,923	8,732	3.83
r/science	561,380	159,636	312,997	118,243	3.15
r/AskWomen	702,912	20,292	200,493	3,772	2.79
r/PoliticalDiscussion	851,603	36,696	244,800	18,873	3.59
r/relationships	2,511,361	58,677	1,436,794	47,581	3.44
r/nba	5,250,635	9,458	2,893,611	9,593	4.30
$\mathcal{D}$	11,525,619	368,341	5,896,775	239,348	3.29

Table 1: Descriptive data statistics for comments and branches by subreddit, sorted by subreddit size. We report how many comments were collected and removed from each subreddit. For branches, we report the total number of branches, how many are event branches (i.e., contain at least one adverse event), and the average length of event branches. The final row summarizes the entire dataset, $\mathcal{D}$.

of features related to three distinct parts of a conversation: the first adverse event (i.e., the *event*), the precursors to the event, and the outcome following the event. We also compute statistics related to authors and subreddits themselves. We use this data in our analysis which consists of multiple logistic regression tasks and various prediction models to answer our research questions.

Data

Reddit is a platform designed to facilitate discussion among users of similar interests. These interests are split up into different sub-communities, referred to as *subreddits*. Reddit inherently promotes threads of conversation in the form of sequential replies, thereby providing an interesting case study to examine our research questions.

Specifically, we focus on all comments posted between May 10, 2016 and February 4, 2017 (study period) for eight subreddits: r/AskWomen, r/books, r/Games, r/legaladvice, r/nba, r/PoliticalDiscussion, r/relationships, and r/science. We select these eight from the top 100 subreddits[4] based on the number of comment removals that were publicly released by Chandrasekharan and Gilbert (2019). These subreddits range in subscriber count from 1.5M (r/PoliticalDiscussion), to 26.9M (r/science). In addition to removals, we collect all comments posted during the study period using Pushshift API (Baumgartner et al. 2020). We refer to the collection of all eight study subreddits as $\mathcal{D}$.

Comment Threads. Next, we introduce new structure to our data to facilitate sequential analysis of conversations. We compile Reddit comments into *branches*, which we define as a single thread of comment replies starting at the post's top-level comment and ending in a single descendent leaf. This structure is visualized in Figure 3. A branch is essentially a tree where each node has at most one child and one parent.

Table 1 provides descriptive statistics related to the data. For each subreddit, we report total number of comments posted within the study period and how many of them were removed. Before constructing branches, we filter out comments based on certain rules. We drop comments with no available body, non-alphanumeric text, comments deleted by their author, and comments posted by human and automated moderator bots (e.g., u/AutoModerator). After preprocessing the data, we construct branches (or conversations) as described earlier. We define *event branches* as branches containing at least one removed comment. Table 1 reports subreddit-level statistics like total number of branches compiled from comments, number of event branches, and average length of event branches. The final row reports statistics aggregated over all subreddits ($\mathcal{D}$). Overall, $\mathcal{D}$ consists of 152,948 unique first removals that lead to 239,348 event branches.

Subreddits in Table 1 are sorted from smallest to largest based on how many comments are in our dataset. We observe that the number of removed comments is not proportionate to the activity level of the subreddit. We can single out two subreddits in particular for their high removal rates: r/science and r/Games. On the other hand, r/nba is the largest subreddit and has the lowest removal rate by far. These observations speak to differences in community characteristics. For example, the high rate of removal for r/science may be attributed to the strict rules and moderator presence on the subreddit. This subreddit demands "no off-topic comments, memes, low-effort comments or jokes."[5] By this rule, a comment like "Thank you!" will be removed in r/science while allowed on most other subreddits.

From this point on, the only branches being considered are those that contain at least one removal. These are event branches that can shed light on the idea of conversational resilience when faced with adverse events. Using these event branches, we analyze the different elements of online conversations in further analyses.

Conversational Outcomes and Features

This section describes all the variables extracted from the branch data. These variables fall into five categories, each representing a component of the visualized regression equation in Figure 4. Whenever possible, we construct variables in terms of raw counts, because proportions raise complications when very few comments are present, and binary vari-

[4]Full list of subreddits at https://github.com/ceshwar/reddit-norm-violations/blob/master/data/study-subreddits.csv.

[5]www.reddit.com/r/science

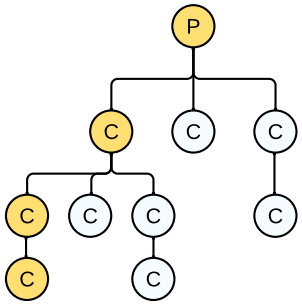

Figure 3: Visualization of a comment thread or branch. The top-level node represents a post and each child node represents a comment reply. Nodes highlighted in yellow form an example of a branch as defined in this paper.

ables using thresholds require validation to ensure reliability. Note that aside from the first removal's text, all described variables were computed for this project.

Outcome Variables. We need to understand the outcome of a conversation in order to evaluate conversational resilience, thus the first category of variables pertains to measures of conversational outcomes. We measure *outcomes* by examining all comments following the first removal (i.e., event) in a branch. Specifically, we examine three outcomes:

1. *Future Removal* represents an outcome that contains at least one additional removed comment.
2. *Future Prosocial* represents an outcome containing at least one prosocial comment. Future prosocial branches can contain some removed comments, but not all comments in these branches may be removed.
3. *Neither* represents an outcome that contains neither removals nor prosocial behavior.

In Table 2 we restate the number of event branches and report how many proceed after the adverse event. Across subreddits, between 42% and 68% of all event branches proceed after the event. This is the first indication that norm violations do not always end a conversation. This lacks the understanding of whether the subsequent comments will violate community norms, but seeing that more than half of all event branches in $\mathcal{D}$ proceed after the first adverse event is nonetheless a step towards quantifying and forecasting resilience.

Table 2 also breaks down what percentage of proceeding branches fall into each outcome category. Note that an outcome can be both future removal and future prosocial, however this only happens in 0.6% of event branches. Future prosocial is consistently the outcome with the fewest branches (5.47% in $\mathcal{D}$) while future removal branches are quite frequent (64.31% in $\mathcal{D}$). This speaks to how challenging it can be for conversations to recover from adverse events. At the subreddit level, we find that r/Games has the largest proportion of future removal branches at 83%. Conversely, only 23% of r/nba event branches proceed to contain a future removal. Further analysis is required to determine if

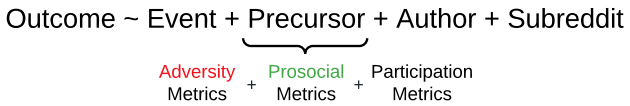

Figure 4: Abstract visualization of our regression analysis. Specific variables used are described in the methods section.

conversations in r/nba are more resilient to adverse behavior, or if r/Games has stricter norm enforcement by moderators.

We use three variables to represent the above mentioned outcomes following the first removal (i.e., event) in a branch:

1. Number of comments after the adverse event.
2. Number of removed comments after the adverse event.
3. Number of prosocial comments after the adverse event.

Event Features (E). These features pertain to the *event* of focus: the first removed comment in an event branch.

1. *Depth:* Depth of the event in its branch. This feature measures "when" the first removal occurs in a conversation.
2. *Toxic Flag:* Binary value representing whether or not the event was toxic. This variable uses the threshold 0.5 to flag a toxic comment.
3. *Highly Toxic Flag:* Binary value representing whether or not the event was highly toxic. This variable uses a threshold of 0.8 to flag a highly toxic comment.
4. *Macro Violation Score:* The macro norm violation score of the event determined by the ensemble of subreddit classifiers trained by Chandrasekharan et al. (2018).
5. *Text:* The removed text of the event using one of four possible word embedding techniques we used for predictors. The removed text was collected by Chandrasekharan and Gilbert (2019). This is the only feature not included in the regressions.

Precursor Features (Pre). The following are computed based on the *precursor* to the event (i.e., all comments before the first removal). Variables include measures of participation, antisocial and prosocial comments before the event.

1. *Avg. Agreement:* Average agreement in precursor.
2. *Avg. Politeness:* Average politeness in precursor.
3. *Avg. Support:* Average support in precursor.
4. *Donation Count:* Number of donation comments.
5. *Gratitude Count:* Number of gratitude comments.
6. *Laughter Count:* Number of laughter comments.
7. *Prosocial Count:* Number of prosocial comments.
8. *Prosocial Proportion:* Proportion of prosocial comments.
9. *Toxic Count:* Number of toxic comments.
10. *Highly Toxic Count:* Number of highly toxic comments.
11. *Non-Toxic Proportion*: Proportion of precursor that is non-toxic, i.e., comments with toxicity score below 0.5.
12. *Event Author Comments:* Number of comments posted by the author of the first removal in the precursor.
13. *Unique Author Count:* Raw count of how many distinct authors posted comments in the precursor.

Subreddit	Event branches	Proceed	Future Removal	Future Prosocial	Neither
r/Games	21,904	11,795 (53.85%)	83.69%	2.13%	14.64%
r/legaladvice	10,650	5,217 (48.99%)	56.01%	5.81%	39.58%
r/books	8,732	3,672 (42.05%)	40.85%	11.38%	49.05%
r/science	118,243	61,450 (51.97%)	78.17%	3.09%	19.81%
r/AskWomen	3,772	1,504 (39.87%)	79.26%	3.39%	18.15%
r/PoliticalDiscussion	18,873	12,408 (65.74%)	48.15%	4.13%	48.60%
r/relationships	47,581	23,154 (48.66%)	42.65%	10.59%	48.84%
r/nba	9,593	6,600 (68.8%)	23.09%	15.05%	63.97%
$\mathcal{D}$	239,348	125,800 (52.56%)	64.31%	5.47%	31.48%

Table 2: This table reports various branch statistics. For each subreddit, we report the number of event branches in the dataset. We also explain the number of branches that proceed after its first removal, and further explore whether the outcome has future removals, future prosocial behavior, or neither. These three columns are expressed as a percentage of the branches that proceed.

Author Features (A). These features pertain to the author of the first removal and are computed using all comments posted by the author during the month prior to the event within the same subreddit as the event.

1. *Monthly Activity*: Number of comments by the author.
2. *Monthly Toxic Count:* Number of toxic comments by the author.
3. *Monthly Highly Toxic Count:* Number of highly toxic comments by the author.
4. *Monthly Non-Toxic Proportion:* Proportion of non-toxic comments by the author (i.e., toxicity less than 0.5).
5. *Monthly Prosocial Count:* Number of prosocial comments by the author.
6. *Monthly Prosocial Proportion:* Proportion of comments by the author that are prosocial.
7. *Monthly Removals*: Number of comments by the author that were removed by moderators of the subreddit.
8. *Monthly Removal Proportion*: Proportion of comments by the author that were removed by moderators.

Subreddit Features (Sub). These features represent characteristics of a subreddit.

1. *Active Authors:* Number of authors who have posted at least five comments in the subreddit. We chose a threshold of five, based on prior work (Chandrasekharan et al. 2017), to account for chance posts from random accounts.
2. *Monthly Avg. Comments:* Average number of comments per month.
3. *Monthly Avg. Removals:* Average number of removals per month.
4. *Subscribers:* Number of subscribers as of December 2021[6] (exact counts during study period are unavailable).

Computational Analysis

We employ two methods of analysis on the data. First, we perform logistic regression analyses to identify associations between conversational features and outcomes. Then, we determine how well these variables can be utilized to build predictive models for the research questions in our study.

[6]https://subredditstats.com

Regression Analysis. Using the outcomes and features previously described, we examine the factors that contribute to conversational resilience, in addition to factors that counteract it. The only feature that is not included in our regression analyses is the text of the first adverse event, which is used exclusively for prediction modeling. All regressions are modeled on the abstract structure illustrated in Figure 4.

We explore three logistic regression tasks driven by our research questions. First, based on RQ1, we examine what variables correlate with whether an event branch proceeds after its first removed comment. Next, we focus our attention on only the event branches that proceed after the first removal. Based on RQ2, we examine associations between conversational features and branches with future removals. Finally, based on RQ3, we explore relationships between conversational features and branches with prosocial outcomes. We run all regressions on each subreddit's data individually, as well as on the combined data from all subreddits to examine how well our findings generalize.

Parameter tuning. For each binary outcome variable, we run Logistic Regression with L1 regularization to shrink unimportant variables to zero for better feature selection. Similar regression analyses with L1 penalties have been employed in the past to identify associations between variables of interest and develop prediction models (Mitra and Gilbert 2014; Chancellor, Lin, and Choudhury 2016). We tune the parameters by varying `alpha`, which serves as a multiplier for the L1 penalty term, between 0 and 1 in increments of 0.01, and select the best model that minimizes the Akaike Information Criterion (AIC) value.

Predictive Modeling. Next, we introduce computational models to forecast conversational outcomes after the first removal. Similar to the structure of our regression task, we construct three prediction tasks based on our research questions. First, we train a predictive model on all event branches from $\mathcal{D}$ to determine whether a branch will proceed after the first comment removal (RQ1). Second, considering only the event branches that proceed after the first removal, we train a predictive model for RQ2 that distinguishes between branches with at least one future removal from those with none. Similarly, the model for RQ3 predicts whether an event branch will contain a future prosocial comment.

We include five categories of features in our predictive models: event (E), text (Text), precursor (Pre), author (A), and subreddit (Sub) features. We build several models per prediction task to explore which feature categories are most informative, specifically distinguishing between text and non-text features. Predictors are trained over the combined data from all eight subreddits ($\mathcal{D}$).

Parameter tuning. Unlike the regression analyses, event features for our prediction task include a representation of the text in first removals. We experiment with three different representations for the first removal's text: bag-of-words (Pedregosa et al. 2011), BERT embeddings (Devlin et al. 2018) using the `sentence_transformers` package (Reimers and Gurevych 2019), and GLoVe sentence embeddings (Pennington, Socher, and Manning 2014) by creating a normalized vector of the sum of each word embedding in a sentence. For GLoVe specifically, we try two different pretrained models. The first was trained on Wikipedia data (GLoVe-W) and the second on Twitter data (GLoVe-T).

Model selection. We test two different types of classifiers: logistic regression with L1 penalty and random forest. During training, we balanced class weights by penalizing training errors made on the less prominent class more heavily. The weights given to each class are inversely proportional to how frequently they appear in the data. All other parameters are set to default values. For each model, we perform 5-fold cross-validation by repeating Stratified K-Fold twice with different randomization in each repetition (Pedregosa et al. 2011).

Findings

This section summarizes the findings from our computational analyses. Results are shown in Table 3 and Table 4.

Regression Analysis

Here we present the results from the regression analysis to answer our research questions and identify factors affecting conversational outcomes using all event branches in $\mathcal{D}$. Additionally, we discuss the generalizability of significant associations observed across study subreddits and highlight key differences, should they arise, found using individual subreddit regression models.

	RQ 1	RQ 2	RQ 3
EVENT FEATURES			
Depth	$-$	$+$	$-$
Highly Toxic Flag	$-$	$-$	$+$
Macro Violation Score	$-$	$+$	$-$
Toxic Flag	$+$	$-$	$+$
PRECURSOR FEATURES			
Avg. Agreement	$-$	$+$	
Avg. Politeness	$-$	$-$	
Avg. Support	$+$		
Donation Count		$+$	
Gratitude Count			$+$
Laughter Count	$+$	$+$	$+$
Prosocial Prop	$-$		
Toxic Count	$-$		
Highly Toxic Count	$+$		
Event Author Comments	$+$	$-$	$+$
Unique Author Count	$+$	$-$	
AUTHOR FEATURES			
Monthly Activity	$+$	$-$	$-$
Monthly Removal Prop	$-$	$+$	$-$
Monthly Non-Toxic Prop	$+$	$-$	$+$
Monthly Prosocial Count	$-$	$+$	$+$
Monthly Prosocial Prop	$-$		
SUBREDDIT FEATURES			
Subscribers	$-$	$-$	$+$
Active Authors (≥ 5)	$-$	$-$	$+$
Monthly Avg. Comments	$+$	$-$	$-$
Monthly Avg. Removals	$+$	$+$	$-$
n	239,348	125,800	125,800

Table 3: L1-regularized Logistic Regression associations for each regression task over all branches in $\mathcal{D}$. The $+$ and $-$ symbols denote positive and negative coefficients respectively. Signs are reported only for variables with significant associations ($\alpha < 0.05$) are reported.

Factors that Allow a Conversation to Proceed (RQ1). We found that 52.56% of all event branches in $\mathcal{D}$ proceed after the first adverse event (see Table 2).

Event Characteristics. Several event features have significant correlations with whether a conversation proceeds after the first adverse event. First, we observed that the deeper into a branch the event occurs, the less likely it is that the branch will proceed afterwards. Given the low average branch lengths (see Table 1), this correlation was expected. This also implies that adverse events are not immediate indicators that a conversation is about to die, even if the adversity occurs early on. Second, we found that highly toxic events and events with higher macro norm violation scores are negatively correlated with the branch proceeding. However, comments that are toxic but not highly toxic are correlated with the conversation proceeding. This indicates that a toxicity threshold of 0.5 may not be high enough to identify removals that halt a conversation.

Precursor Characteristics. We found that more comments by the event author and more unique authors in the precursor are positively associated with a conversation proceeding. It is likely that these conversations include many exchanges between the event author and others, and one norm violation does not necessarily end a conversation. We do note an exception to this correlation in r/science, where more precursor comments posted by the event author are associated with the conversation not proceeding. Due to stricter moderation, r/science users have many restrictions to consider when posting, thus posting multiple times in one conversation may present more opportunities for norm violations.

Additionally, while more toxic comments in the precursor are associated with a conversation not proceeding, more

highly toxic comments have the reverse correlation. This could indicate that highly toxic comments inspire more heated discussion. We also observe disagreement when exploring whether more prosocial behavior encourages a conversation to continue. Agreement, politeness, and proportion of prosocial comments are negatively associated with a conversation continuing, but support and laughter have positive associations. Participants in the precursor to an event who are behaving politely and agreeably do not seem to reply to adverse comments. We found that r/books is an outlier to this trend with a positive correlation between precursor politeness and the conversation proceeding. This variable within r/books also positively correlates with future prosocial comments, thus it seems that precursor politeness encourages continued prosocial conversation in r/books.

Author Characteristics. We found three key associations when examining characteristics of the first adverse event's author. First, adverse comments posted by authors who are highly active in the subreddit are less likely to end conversations. Second, we found that adverse comments posted by authors with higher rates of removal within the subreddit are more likely to end a conversation. This may indicate that norm violations by repeat offenders lead to less engagement from other members. Finally, adverse behavior by authors who have a history of non-toxic behavior is more likely to be followed by comments. All associations generalize across study subreddits, indicating how the author of an adverse comment matters in the context of resilience.

Subreddit Characteristics. We observed that conversations with adverse events in smaller subreddits (based on number of active members and subscribers) are more likely to continue after the first removal than similar conversations in larger subreddits. Additionally, subreddits with higher rates of activity and removals tend to have branches that continue after the first adverse event. Even though conversations may proceed within subreddits with higher rates of removal, it is likely that they will have more future removals (RQ2).

Factors that Lead to Future Removals (RQ2). Out of all event branches in $\mathcal{D}$ that proceed after the first adverse event, 64.31% contain future removals (see Table 2).

Event Characteristics. Upon examining the adverse events within branches that proceed, we observed that toxic events (measured by the two toxicity flags) are negatively associated with future removals. This suggests that toxicity in comments does not necessarily encourage further adverse behavior within communities. Alternatively, toxicity may not serve as an accurate measure of adversity within a community, thus demonstrating the need for context-sensitive measures for adversity like comment removals. High macro norm violation scores, however, are positively associated with future removals, indicating that comments which violate Reddit macro norms may lead to more adverse behavior. We also found a positive correlation between adversity initially occurring later in the branch and future removals, but this did not generalize across subreddits.

Precursor Characteristics. We found that politeness in precursor comments is negatively correlated with future removals, however we noticed surprising relationships between the other prosocial metrics and future removals. Agreement, donation, and laughter are all positively associated with future removals. Laughter and donation are lexicon-based counts and these associations could indicate their unreliability as signals of prosocial behavior in some contexts. Laughter is frequently sarcastic and donation requests can be scams. Also, users agreeing with one another in the precursor might engage in similar adverse behavior.

We noticed some interesting correlations with respect to the precursor participation metrics. If the author of the first adverse event posts more comments in the precursor, it is less likely that there will be future removals. This may indicate that the author has already established themselves in the conversation as a non-adverse presence capable of posting comments that do not violate norms. Looking back, we found that conversations where the first adverse event's author has posted more comments in the precursor are not only more likely to proceed after the first removal (RQ1), but also less likely to contain future removals (RQ2). Only exceptions to this observation are r/books and r/PoliticalDiscussion, revealing positive correlations between event author precursor comments and future removals within these two subreddits. Given that these subreddits' express purpose is discussion, it is possible that back-and-forth exchanges between an author and other posters may get heated more quickly than conversations with more unique participants. Branches with more unique authors in the precursor are less likely to have future removals. When many authors contribute to a conversation, one user's adverse comment may not cause the conversation to devolve. On the flip side, a conversation with few users contributing prior to an adverse event is more susceptible to future removals.

Author Characteristics. We observed that users with higher removal rates in the past are more likely to incite future removals in response to their adverse comments, while users with higher monthly activity and proportion of non-toxic comments have a negative association with future removals. The first observation uncovers the presence of repeat offenders, and the second shows that more frequent participants of a community may be less likely to cause further problems. Closer examination revealed that r/science is the only subreddit with a significant negative association between author monthly activity and future removals. This may indicate that long-time participants of r/science are more familiar with the norms and thus will not commit additional norm violations. Subreddits like r/AskWomen and r/relationships see the reverse trend, thus there may not be the same adoption of norms from their users.

There is an unexpected positive correlation between authors with a large number of prosocial posts and future removals. However, there is no significant correlation with author's proportion of prosocial comments. This result may be skewed by sparsity of prosocialilty in branches, and these correlations do not generalize across all study subreddits.

Subreddit Characteristics. We found negative associations between larger subreddits (i.e., more active authors and subscribers) and future removals. Looking ahead to the analysis of prosocial outcomes (RQ3), we will see that such subreddits also tend to have more future prosocial outcomes.

Features	RQ 1		RQ 2		RQ 3	
	AUC-ROC	F1	AUC-ROC	F1	AUC-ROC	F1
E+A+Pre+Sub	0.825	0.747	0.874	0.845	0.726	0.270
E+A+Pre	0.799	0.715	0.839	0.820	0.701	0.251
E+A+Sub	0.799	0.729	0.855	0.831	0.716	0.260
E+Pre+Sub	0.767	0.687	0.835	0.811	0.708	0.222
A+Pre+Sub	0.743	0.659	0.838	0.828	0.700	0.241
Text	0.895	0.843	0.896	0.855	0.744	0.278
E+Text	0.909	0.858	0.899	0.860	0.748	0.289
E+Text+A+Pre+Sub	**0.918**	**0.865**	**0.911**	**0.869**	**0.762**	**0.290**

Table 4: AUC-ROC and F1 scores for each prediction task described in the methods section. Results are shown only for the top performing models from 5-fold cross-validation, by repeating Stratified K-Fold repeated twice over all event branches in $\mathcal{D}$.

Factors that Lead to Future Prosocial Outcomes (RQ3). We found that out of all event branches in $\mathcal{D}$ that proceed, only 5.47% contain future prosocial outcomes (see Table 2).

Event Characteristics. First, we observed a positive association between adverse events appearing early in the conversation and future prosocial comments. Looking back at RQ1's analysis, we saw that adverse events appearing at the start of a conversation do not necessarily cause the discussion to halt. Combining both observations, we found that conversations that proceed may actually contain prosocial outcomes, thereby indicating conversational resilience.

Precursor Characteristics. Referring back to RQ2, we previously observed that more author comments in the precursor negatively correlate with future removals, and now we see a positive correlation with future prosocial comments. Thus, authors already established in a conversation may be less disruptive to a conversation should they violate a community norm. Also, gratitude and laughter before an adverse event are more likely to lead to prosocial outcomes.

Author Characteristics. Three main observations were made about the associations between author features and prosocial outcomes. We saw that an author's monthly prosocial count and non-toxic proportion are positively correlated with prosocial outcomes while their monthly removal proportion is negatively correlated. This suggests that if an adverse comment is from an author who violates fewer norms and contributes more non-toxic, prosocial content, the outcome of the conversation is likely to be more prosocial.

Subreddit Characteristics. We found that conversation threads posted in subreddits with more active users and subscribers are positively associated with prosocial outcomes after the first adverse event. Conversely, subreddits with higher rates of comments and removals have fewer prosocial outcomes. This may indicate that subreddits where a smaller number of users post frequently are less capable of achieving prosocial outcomes after adverse events.

Predictive Modeling

Table 4 shows results for the three prediction tasks described in the methods section. For each prediction task, we identified the best performing model as the one that had the highest area under the receiver operating characteristics curve (AUC-ROC). AUC-ROC is a measure of a model's ability to distinguish between two binary classes. A model with an AUC-ROC of 1 can distinguish between two classes perfectly. All three tasks performed best when using the Random Forest model over Logistic Regression, thus all models described in Table 4 used Random Forest. Models using text features were trained using bag-of-words text representations, which had the highest average AUC-ROC (0.81) and F1 (0.62) scores across all tasks. We used the representations of the 5000 most frequent words as features. BERT sentence embeddings, GLoVe-T and GLoVe-W all had roughly the same performance (avg. AUC-ROC: 0.71 ; avg. F1 0.56).

We also measured the Matthews correlation coefficient (MCC) for each model to further evaluate the quality of our binary classification task. This measure ranges from -1 to 1 where a value of 1 indicates perfect predictions. MCC is especially helpful to evaluate models on imbalanced data, because it considers all four categories—True Positives, False Positives, True Negatives, False Negatives—in a confusion matrix, taking into account the size of each binary class (Chicco and Jurman 2020).

Predicting Whether a Conversation Proceeds (RQ1). Our best-performing model for RQ1 achieved AUC-ROC of 0.918, F1 of 0.865, and MCC value of 0.74, and this model used all five sets of available features—event (E), text (Text), author (A), precursor (Pre) and subreddit (Sub) features.

Informative Non-text Features. We evaluated model performance using different combinations of non-text features. The top portion of Table 4 presents the best performance of models trained only using *non-text* features. The first row reports the results of running a model on all non-text features and subsequent rows exclude one category of feature at a time to reveal which are the most informative. For RQ1, the features that contribute the most to AUC-ROC and F1 scores, and are thus the most informative features, are event features. Author features are the next most important for this prediction task. Precursor features and subreddit features contribute the same boost in AUC-ROC scores for the models, however, excluding subreddit features has a more detrimental effect on the F1 score than we observe when excluding precursor features.

Predictive Power of Textual Features. Next, we evaluated model performance when using textual features, alongside

the others described earlier. The bottom portion of Table 4 presents the best performance of models trained using *text* features. We see that training a model using the text representation as the only feature surpasses the best non-text model in terms of AUC-ROC (0.895 vs. 0.825) and F1 scores (0.843 vs. 0.747). Augmenting the text features with the remaining features of the first adverse event increases model performance even more (AUC-ROC of 0.909, F1 score of 0.858). Finally, using all non-text and text features results in the best model performance across both metrics (AUC-ROC of 0.918, F1 score of 0.865).

Predicting Future Removals (RQ2). Our best-performing model for RQ2 achieved AUC-ROC of 0.911, F1 of 0.869, and MCC value of 0.65. Like the best-performing model for RQ1, all five types of features were used in RQ2's best model.

Informative Non-Text Features. Unlike RQ1, the most informative feature for predicting future removals in a conversation are author-level features. The model excluding author features faces the largest drop in both AUC-ROC and F1 scores. This indicates a strong association between certain types of authors and outcomes containing more removals. With respect to AUC-ROC scores, event features are the second most informative followed by subreddit features, however this trend is reversed when considering F1 scores. Finally, similar to RQ1's model, we find that precursor features are the least informative for predicting future removals.

Predictive Power of Textual Features. Similar to the prediction task for RQ1, utilizing text features alone is a performance improvement over the model using all non-text features when examining AUC-ROC (0.896 vs. 0.874) and F1 scores (0.855 vs. 0.845). Again, we augment this basic text model with more event features, which shows a small boost in performance (AUC-ROC of 0.899, F1 of 0.86), and finally our model trained on all available features out-performs all other models (AUC-ROC of 0.911, F1 of 0.869).

Predicting Future Prosocial Behavior (RQ3). Our best-performing model for RQ3 achieved AUC-ROC of 0.762, F1 of 0.29, and MCC value of 0.29 using all available features.

Informative Non-text Features. The most informative non-text features for predicting whether an outcome will contain prosocial behavior is somewhat more complicated than for the other two prediction tasks. If we examine the AUC-ROC scores, event features appear to be most informative, closely followed by subreddit features. However, focusing on F1 scores we notice that the most informative types of features are author features, followed by event features. Across both metrics, we find again that precursor features are the least informative. This observation holds across all three prediction tasks, indicating that precursor features are generally the least informative types of features across the board.

Predictive Power of Textual Features. Again, looking at the bottom portion of Table 4, we report results for models trained using text for RQ3. Using text features alone improves the best non-text model's AUC-ROC (0.744 vs. 0.726) and F1 scores (0.278 vs. 0.27). Including event features in the model slightly improves its performance (AUC-ROC of 0.748, F1 of 0.289). As with the other prediction tasks, models trained on all features, both text and non-text, perform the best (AUC-ROC of 0.762, F1 of 0.29).

Discussion

In this paper, we have uncovered factors contributing to conversational resilience in the face of adversity. By examining the characteristics of resilient online conversations we can learn what qualities online community members and moderators should value and try to promote. Encouraging conversation-level characteristics that promote resilience is a proactive method for minimizing the capacity of antisocial behavior to propagate. Additionally, this research provides insights into the types of conversations that are typically not resilient to adverse behavior. The methods we developed can inform moderators on where they should focus their attention in order to prevent further detrimental outcomes within conversations. Our findings have implications for the development of moderation tools to predict which conversations are more likely to devolve and which are more likely to persist, sometimes even leading to prosocial outcomes.

Factors Affecting Conversational Outcomes. Through our statistical analyses, we present valuable insights at the user, comment, and subreddit levels. At the user-level, we observed that first adverse events in conversation threads posted by authors with a history of norm violations are especially dangerous to the flow of conversation, associated not only with the conversation not proceeding afterwards, but with future removals and no future prosocial comments when there is an outcome. On the other hand, users with a history of non-toxic comments may help a conversation continue without further violations, and even encourage prosocial outcomes. Looking at the comment-level, events that are macro norm violations are prone to either ending conversations or resulting in future removals. However, we found some features of a conversation discourage future norm violations and, in some cases, encourage prosocial outcomes. In particular, conversations in which the author of the first adverse event contributes more comments in the precursor to that event tend to not only proceed after the violation, but are less likely to have future removals and more likely to have prosocial outcomes. On the subreddit-level, we observed that conversation threads from subreddits containing more active authors and subscribers are more likely to have conversations with no removals after the first violation, and are actually more likely to have prosocial outcomes, pointing to the resilience of larger communities.

Forecasting Conversational Resilience. The strong associations from our analyses indicate the predictive power of our independent variables at determining whether a conversation will end after the first adverse event, continue to have future removals, or continue to have prosocial outcomes. Our motivation was to train forecasting models to identify conversations that demonstrate resilience after an adverse event. In other words, we intended to predict the subsequent damage that the first removal in a conversation can potentially lead to. Findings from our predictive models trained for each RQ indicate that we were able to achieve AUC-ROC scores between 0.76 and 0.91. This predictive power

demonstrates how our approach can allow moderators to identify adverse events that can lead to further adverse behavior, thereby needing their attention first over other events.

While the AUC-ROC scores are reasonably high for RQ3's model, the F1 score indicates very low precision and/or recall values. This is a result of the data imbalance caused by the sparsity of prosocial content. The amount of prosocial behavior observed after an adverse event was low in comparison to the other variables being measured. This is an indication that we may need more, high-quality metrics for capturing prosocial behavior. While the training data was appropriately weighted during training, the test data used to evaluate the models was not balanced in order to retain class distributions. Since the data used for RQ3's task had many more branch outcomes without prosocial behavior than those with, our models suffered from a high frequency of false positives (i.e., falsely predicting the presence of prosocial behavior) which drastically affected MCC and F1 scores.

Limitations & Future Work

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

Focusing on More Communities. At this time, only eight subreddits were considered based on their availability of removed comments in the dataset publicly released by Chandrasekharan et al. (2018). Even though these eight subreddits represent unique communities with varying topics and norms, they are not entirely representative of all Reddit communities. In future research, a wider range of subreddits should be included to best capture the range of community characteristics on Reddit. Nevertheless, we examined the generalizability of our findings across all study subreddits, by making sure to run all regressions and classifications tasks on each individual subreddit, in addition to all event branches in $\mathcal{D}$, and carefully highlighting subreddit-wide patterns and differences in associations should they arise.

Expanding Feature Sets. In future research, we hope to incorporate additional subreddit data and expand the features and variables in the analysis. There are many other conversational attributes that are worth considering (e.g., counter speech). For example, looking into the first adverse event's author's karma and gender are two more features that could be worth exploring. Additionally, some findings in the previous section raise questions about who is contributing to the outcomes being measured. Specifically, if an outcome contains removals, who is responsible for those violations? Conversely, are some users more capable of inciting prosocial behavior? We intend to explore whether particular users have more power to incite prosocial imitation effects.

Examining Causal Factors. Future work should explore causal relationships between the different features we highlighted in this work and pro/anti-social outcomes in online conversations. We note that our current approach examined associations between conversational outcomes and various other factors surrounding adverse events, with an eye towards forecasting conversational resilience. Our results are limited in their explanatory power since they do not allow us to make causal claims about the observed associations.

Incorporating Context-Sensitive Metrics. In order to evaluate the prosocial metrics used in this paper, we performed an instrument validation through manual review of comments containing prosocial outcomes. Upon manual review of 100 randomly-selected comments determined to contain prosocial outcomes, we found that 69 of them were confirmed as prosocial. This may impact any observations made about prosocial outcomes in conversations. Also, the correlations associated with the toxicity-related features reveal that toxicity scores alone may not be reliable as indicators of adverse behavior within communities. As a result, we turned to more context-sensitive indicators of adversity like comment removals and macro norm violation scores.

Accounting for Cumulative Conversational Impacts. Aside from additional variables, our future work will explore ranking methods for taking in two adverse events and predicting which will have the worse outcome and/or which will have a more prosocial outcome. Another approach to measure adversity would be to calculate the cumulative impact of an adverse event by considering not only one branch that follows at a time, but aggregating across all branches stemming from each adverse event.

Conclusion

In this work, we examined the factors contributing to conversation-level resilience in Reddit conversations. In exploring eight subreddits with sufficient norm violations, we employed statistical methods to identify characteristics of comment threads, users, and subreddits associated with conversational outcomes. Specifically, we identified features of a comment thread that may afford resilience to a conversation, and features that may weaken a conversation. We can use this understanding of resilience to expand the ways we moderate online communities towards more proactive approaches. These methods can be applied to many more subreddits and used to provide specialized guidance and assistance to overworked moderators, hopefully affording a path towards fostering resilience in our online communities.

Broader Perspective, Ethics, and Competing Interests. This research has implications for improving community health beyond Reddit—our methods can be extended to other online spaces that allow for threaded conversations. But using our findings at face-value to guide moderation strategies may have negative effects on users. For example, we observe that users with a history of removed comments may incite future removals. Should moderators decide to keep an eye on users with a history of norm violations or newcomers still learning community norms, it may become difficult for them to learn how to participate meaningfully.

With respect to data collection and analysis, we recognize that using "removed data" (i.e., moderated comments released by prior work) can give rise to ethical concerns. Before performing this research, we discussed these issues with

our IRB and colleagues. Based on our discussions, we concluded that examining moderated comments provided valuable insights about online governance, and as long as we mitigated potential risks, the benefits outweighed the risks. In an effort to mitigate risks, we purposefully discard comments deleted by the author, which we believe would violate the author's privacy. Finally, we used public data collected via Pushshift API to protect Reddit itself from harm.

Acknowledgments

We thank the members of the Social Computing Lab (SCUBA)—specifically Jackie Chan, Fred Choi, and Tanvi Bajpai—at the University of Illinois at Urbana-Champaign for their valuable input that improved this work.

References

Bao, J.; Wu, J.; Zhang, Y.; Chandrasekharan, E.; and Jurgens, D. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of the Web Conference 2021*, WWW '21.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 830–839.

Bernstein, M.; Monroy-Hernández, A.; Harry, D.; André, P.; Panovich, K.; and Vargas, G. 2011. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1).

Chancellor, S.; Lin, Z.; and Choudhury, M. D. 2016. "This Post Will Just Get Taken Down": Characterizing Removed Pro-Eating Disorder Social Media Content. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

Chandrasekharan, E.; and Gilbert, E. 2019. Hybrid Approaches to Detect Comments Violating Macro Norms on Reddit. *arXiv preprint arXiv:1904.03596*.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Chandrasekharan, E.; Samory, M.; Jhaver, S.; Charvat, H.; Bruckman, A.; Lampe, C.; Eisenstein, J.; and Gilbert, E. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Chang, J. P.; and Danescu-Niculescu-Mizil, C. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *Proceedings of WWW*.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial Behavior in Online Discussion Communities. In *Proceedings of ICWSM*.

Chicco, D.; and Jurman, G. 2020. The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, 21.

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. A Computational Approach to Politeness with Application to Social Factors. arXiv:1306.6078.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Hessel, J.; and Lee, L. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. arXiv:1904.07372.

Jhaver, S.; Appling, D. S.; Gilbert, E.; and Bruckman, A. 2019a. "Did You Suspect the Post Would be Removed?" Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–33.

Jhaver, S.; Birman, I.; Gilbert, E.; and Bruckman, A. 2019b. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5): 1–35.

Jurgens, D.; Chandrasekharan, E.; and Hemphill, L. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. *arXiv preprint arXiv:1906.01738*.

Mitra, T.; and Gilbert, E. 2014. The Language That Gets People to Give: Phrases That Predict Success on Kickstarter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, 49–61.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct): 2825–2830.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Seering, J.; Kraut, R.; and Dabbish, L. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, 111–125. New York, NY, USA: Association for Computing Machinery.

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.