Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

Are Proactive Interventions for Reddit Communities Feasible?

Hussam Habib[1], **Maaz Bin Musa**[1], **Fareed Zaffar**[2], **Rishab Nithyanand**[1]

[1] University of Iowa,
[2] Lahore University of Management Sciences
[1] {hussam-habib, maazbin-musa, rishab-nithyanand}@uiowa.edu, [2] fareed.zaffar@lums.edu.pk

Abstract

Reddit has found its communities playing a prominent role in originating and propagating problematic socio-political discourse. Reddit administrators have generally struggled to prevent or contain such discourse for several reasons including: (1) the inability for a handful of human administrators to track and react to millions of posts and comments per day and (2) fear of backlash as a consequence of administrative decisions to ban or quarantine hateful communities. Consequently, administrative actions (community bans and quarantines) are often taken only when problematic discourse within a community spills over into the real world with serious consequences. In this paper, we investigate the feasibility of deploying tools to proactively identify problematic communities on Reddit. Proactive identification strategies show promise for three reasons: (1) they have potential to reduce the manual efforts required to track communities for problematic content, (2) they give administrators a scientific rationale to back their decisions and interventions, and (3) they facilitate early and more nuanced interventions (than banning or quarantining) to mitigate problematic discourse.

1 Introduction

Reddit has over 138K active communities, called *subreddits*, with over 330M active users making it the 6th most popular website in the USA.In recent years, the site has been mired in controversies around the role that its communities played in originating and propagating sexist, racist, and generally hateful online socio-political discourse. A few of the recent controversies have involved communities such as: *r/Physical_Removal (banned in 8/2017)* which advocated for the physical removal of 'liberals' in the United States prior to and even after the murder of Heather Heyer in Charlottesville (Collins 2017), *r/incels (banned in 11/2017)* which endorsed and celebrated the murder of and violence against sexually active women (Collins and Zadronzny 2018), and *r/greatawakening (banned in 3/2018)* and *r/pizzagate (banned in 11/2016)* which falsely alleged the existence of child trafficking rings by the US Democratic Party and left-wing corporations resulting in real-life attacks, threats, and harrassment (Ohlheiser 2016).

In reaction to many of these controversies, Reddit has resorted to banning or quarantining subreddits citing violations of the Reddit content policy which prohibits specific types of content including content which "*encourages or incites violence*". However, the effectiveness and timeliness of such bans and quarantines are frequently debated. While previous research (Chandrasekharan et al. 2017) concluded that such bans "*worked for Reddit*", others have pointed out that they are too reactionary and occur only after a significant amount of damage has already been observed (Morse 2019; Romano 2017). Along another dimension, Reddit has also faced criticism for inconsistent and seemingly ad-hoc applications of the content policy by those claiming that the platform provides a safe-haven for extremist ideologies and others claiming that the platform leverages the content policy as a mechanism to censor "non-mainstream" opinions and ideologies. Furthermore, Reddit admins and moderators have claimed that the non-static nature of communities requires them to perform constant monitoring and community guidelines updates (Seering et al. 2019) – a task which makes administration more challenging. Despite these criticisms and known challenges surrounding Reddit administration, little is actually known about the evolutionary characteristics of subreddits and the predictors of problematic subreddits (i.e., *those deemed to have violated the content policy*). Further, there are no publicly available tools to help Reddit administrators make timely and sound intervention decisions. We seek to fill these gaps by studying the evolutionary characteristics of subreddits and developing an administrative tool to help with early identification of potentially problematic subreddits. This report describes our analysis and methods related to the following two hypotheses.

H1. Subreddits do not converge to stability. This hypothesis demonstrates the need for automated tools to monitor subreddit evolution. If valid, it shows that constant monitoring is required for subreddits due to the evolving nature of discourse and participation – a prohibitively expensive proposition for administrators without automated monitoring tools. In order to test this hypothesis, we develop techniques to track subreddit evolution in terms of vocabulary and participating users. Our work validates this hypothesis.

H2. Evolution in problematic subreddits can be predicted. This hypothesis demonstrates the promise of automated tools to help administrators make timely and sound intervention decisions. If valid, it shows that pre-emptive identification of subreddits likely to violate Reddit's content policy is possible. In order to test this hypothesis, we develop explainable ML models which rely on a variety of features including structural-, linguistic-, community-, and user-related features to predict the evolutionary outcome of a subreddit. Our models show that that problematic communities can be identified by their evolutionary behavior early in their lifetime. Further, the explainability of our models provides administrators with an understanding of the causes for classification decisions and the ability to use their expertise to overrule and augment them. Finally, we deploy our classifier in a real world continuous learning scenario, identical to its use-case for Reddit, and study its predictions.

Taken together, our study demonstrates the feasibility of proactive and explainable machine-aided strategies to help, but not replace, human administration of Reddit.

2 Reddit: The Platform and Dataset

In this section, we provide a high-level overview of Reddit with a focus on its content and administrative policies (§2.1) and the datasets we rely on (§2.2).

2.1 An Overview of the Reddit Platform

Reddit allows its users to create and moderate subreddits. Subreddit moderators typically choose their own fellow moderators from within the community, with a few exceptions for newly created communities and cases where there are no volunteers within the community. Subreddit moderators are tasked with setting and enforcing the rules of engagement within a subreddit. Moderators may enforce rules via the use of user bans and content deletion. However, the actions of subreddit moderators do not impact redditor experiences outside of that subreddit (e.g., a subreddit moderator cannot enforce site-wide bans). In addition to relying on volunteers, Reddit also employs administrators to set and enforce site-wide policies for content and user engagement. These content policies are mandatory and applied *in addition to* a subreddit's own policies. In the event of content policy violations, administrators have the ability to: (1) ban users from making posts or comments visible to the rest of the platform (i.e., *shadow ban*), (2) prevent subreddits from appearing on the Reddit front-page and in search results (i.e., *quarantine*), and (3) ban subreddits from the platform. Currently, this administration process is largely manual requiring a team of administrators to manually study reports of content violations submitted by Redditors with little support provided in the way of automated decision-making aids. Beyond poor scalability, these manual efforts have also impacted the mental health of content-policy administrators on Reddit (Lagorio-Chafkin 2018).

2.2 Datasets

In this paper, we focus on specific subsets of the entire platform. These were gathered using the Pushshift API (Baumgartner et al. 2020). We note that due to computational limitations, the vocabulary vectors used in our analysis (described in §3.1) leverages random 10% samples of each of the datasets below. However, the remainder of our analysis has no such limitations.

Most active subreddits ($\mathfrak{D}_A$). This dataset contains the 424M posts and 4.5B comments made by 42M unique users to the 3K most active subreddits (i.e., highest average number of comments per month) which did not receive any administrative interventions (i.e., bans or quarantines) during the period from 01/2015 to 04/2020.

Subreddits with administrative interventions ($\mathfrak{D}_I$). This dataset contains 38M posts and 353M comments made by 2.3M unique users from 264 subreddits which, between 2015 to 2020, were the subject of either administrative bans, quarantines, or both. Since most bans or quarantines are not announced by the Reddit administrative team, the list of banned subreddits was obtained by visiting the webpages of subreddits in which user activity had ceased and confirming the presence of a ban notification from Reddit [1]. The date of the last post made on the subreddit was used as the ban date for each banned subreddit. To identify quarantined subreddits, we leveraged data from *r/reclassified* which lists a crowd-sourced subset of all quarantine events on the Reddit platform. Quarantine actions were confirmed by visiting the webpages of the subreddits and confirming the presence of a quarantine notification from Reddit. The date of the post on *r/reclassified* was used as the quarantine date for each quarantined subreddit. Since our goal is solely to capture the characteristics of problematic communities, we consider bans and quarantines as equivalent.

Control subreddits without administrative interventions ($\mathfrak{D}_C$). Since $\mathfrak{D}_A$ and $\mathfrak{D}_I$ have vastly different sizes and contain subreddits with different amount of activity, we create a control dataset to facilitate comparisons with $\mathfrak{D}_I$. This dataset contains subreddits most similar to those in $\mathfrak{D}_I$ along two parameters: vocabulary and activity. For each subreddit in $\mathfrak{D}_I$ and $\mathfrak{D}_A$, we create two vectors: a vocabulary vector (using the techniques outlined in §3.1) and an activity vector which denotes the number of comments in the subreddit during each month. For each subreddit in $\mathfrak{D}_I$, we then find the subreddit in $\mathfrak{D}_A$ which has the most similar topic and activity vector. Similarity is computed by cosine similarity and weights are equally assigned for the topic and activity vector. We manually verified the similarities of each matched pair of subreddits. Examples of $\mathfrak{D}_C$ subreddits and their corresponding $\mathfrak{D}_I$ subreddits are: (*r/conspiracy*, *r/911truth*), (*r/Conservative*, *r/The_Donald*), (*r/PurplePillDebate*, *r/MGTOW*), and (*r/niceguys*, *r/Incels*). This dataset contains 44M posts and 489M comments made by 44M unique users to all $\mathfrak{D}_I$-matched subreddits.

3 Subreddit Evolution and Convergence

In this section, we focus on testing the following hypothesis: **H1. Subreddits may not converge to stability**. We measure stability by the vocabulary of the community and the 'backgrounds' of the users participating them them. If valid, this hypothesis demonstrates the need: (1) for computational techniques to monitor subreddit evolution (in terms

[1] see https://www.reddit.com/r/incels/ for an example

of vocabulary and user bases) and (2) to frequently evaluate the suitability of making administrative interventions. Put another way, the validity of this hypothesis would suggest vocabulary and user base in communities are always evolving and therefore require consistent and frequent monitoring — a task which is known to be non-scalable, expensive, and extremely laborious for human moderators (Lagorio-Chafkin 2018; Roberts 2014; Wohn 2019). In order to test this hypothesis, we conduct analysis to check if the vocabulary used and users participating in subreddits in $\mathfrak{D}_A$, $\mathfrak{D}_I$, and $\mathfrak{D}_C$ stabilize over time.

3.1 Methods

Representing subreddit vocabulary and users with fixed-length vectors. A *subreddit state* is all the activity associated with a subreddit during a given month – e.g., the subreddit state associated with *r/politics* on 09/2019 contains all user posts and comments made on *r/politics* during the month of 09/2019. Our goal is to create two fixed-length vectors which capture: (1) the vocabulary associated with each subreddit state and (2) the active user base associated with each subreddit state.

Creating fixed-length vocabulary vectors for each subreddit state. At a high-level, our vocabulary vector for each subreddit state is the vector of TF-IDF weights associated with each unique token in our dataset.

- *Random sampling and document corpus creation.* We begin by randomly sampling 10% of all comments in our dataset (a necessity owing to the large dataset and computational limitations). The sampled comments are then used to create documents associated with each subreddit state – e.g., the document associated with the (*r/news*, 09/2019) subreddit state will contain all the sampled comments which were made by users on *r/news* during the month of 09/2019. At the end of this step, we have a corpus of documents (D) containing one document for each subreddit state in our dataset.
- *Text pre-processing and token corpus creation.* We perform standard text pre-processing operations including removing English stop words, tokenization, and lemmatization for each document. The unique tokens, across D, at the end of this pre-processing stage form the corpus of words and determine the length of the vectors associated with each subreddit state. At the end of this step, we have a corpus of all unique lemmatized tokens (T) observed in D. The length of the topic vector associated with each subreddit state is $|T|$.
- *Computing vocabulary vectors for each subreddit state.* For each document $d \in D$, we compute the TF-IDF weight of each token $t \in T$. Therefore, the vocabulary associated with each subreddit state are represented by a vector denoting the importance of each token with respect to all sampled comments made on the subreddit during the corresponding month.

Creating fixed-length active user vectors for each subreddit state. At a high-level, our active user vector for each subreddit state is the vector of fractions of active user co-occurrences with other subreddit states. This is a standard technique used in collaborative filtering and recommender systems research (Liang et al. 2016; Dunning and Friedman 2014) to identify shared interests between sets of users.

- *Active user identification.* We begin by creating sets of active users for each subreddit state in our dataset. Active users associated with a subreddit state are identified as all users with any posting or commenting activity on the corresponding subreddit during the corresponding month. We refer to the set of active users of subreddit state s as A_s. Active user vectors were created using the complete dataset without sampling.
- *Computing active user vectors for each subreddit state.* For each subreddit state, we compute the fraction of active users who overlap with every other subreddit state. The j^{th} entry in the active user vector for subreddit state i corresponds to the fractional overlap with subreddit state j – i.e., it has the value $\frac{|A_i \cap A_j|}{|A_i|}$. A higher overlap ratio between two subreddit states indicates that the two subreddits shared a large cohort of users during the specific month. The length of this vector is equal to the total number of subreddit states in our dataset.

Measuring the distance between two subreddit states. There are two types of distance measures that are available for use: absolute (e.g., cosine similarity between the vectors of two subreddit states) and relative (e.g., similarity of ranked lists of nearest subreddit state neighbors for two subreddit states). We chose to leverage the latter. Note, however, that our analysis with absolute distance measures yielded similar results to those presented in §3.2. Relative distance measures, in the context of subreddit states, allow us to quantify how a subreddit's vocabulary and user bases have evolved as a function of the vocabularies and user bases of other subreddits. For example, let us once again consider the states associated with (*r/news*, 11/2019) and (*r/news*, 04/2020). The relative distance between the vocabulary vectors associated with these subreddits will account for the fact that although the absolute change between the vectors is large due to the change in discussion topics from American primary elections to COVID-19, the changes with relative to other subreddits is smaller – i.e., subreddits (e.g., *r/usnews*) which shared similar topics with *r/news* in 11/2019 still shared similar topics in 04/2020. Analysis with a relative distance measure therefore identifies how much subreddits have changed with respect to their neighbors and how their role on the platform has changed. We say that convergence has occurred if the relative distance computed over consecutive months converges to the minimum.

Quantifying relative distance using Rank-Biased Overlap (RBO) (Webber, Moffat, and Zobel 2010). Given two vectors of identical length which represent either vocabulary or active user base vectors of two subreddit states, we perform the following operations to obtain their relative distance.

- *Generate list of neighbors ordered by euclidean distances.* Let v_1 and v_2 be the two input (topic or user base) vectors, associated with subreddit states s_1 and s_2 and belonging to months m_1 and m_2, whose relative distance we wish to compute. Let S_{m_1} be the set of all subreddit states from month m_1 and S_{m_2} be the list of all subreddit states from m_2. We begin by computing the euclidean distances between (1) v_1 and every subreddit state in S_{m_1} and (2) v_2 and every subreddit state in S_{m_2}. Finally, we sort the elements of S_{m_1} and S_{m_2} in ascending order of their euclidean distance to v_1 and v_2 and store the sorted list in X_1 and X_2, respectively. These lists are effectively the subreddit neighbors of s_1 and s_2 during m_1 and m_2, respectively.

- *Computing RBO scores.* Given the lists of neighbors, X_1 and X_2, we then compute the RBO similarity score between them. We use RBO since it automatically imposes higher penalties for disagreements at top ranks and works for non-conjoint and arbitrarily long ranked lists. These properties are not available using methods such as Kendall's Tau rank similarity metrics. A high similarity score (≈ 1) indicates a low relative distance between the two subreddit states – i.e., the two states have nearly identical sets of nearest neighbors.

3.2 Results

Figure 1 shows the results of our measurement of monthly subreddit vocabulary evolution using a relative distance measure (RBO). The plot shows the distribution of the measured RBO distances between any two consecutive months for each subreddit in $\mathfrak{D}_A$, $\mathfrak{D}_C$, and $\mathfrak{D}_I$. We can make several observations from these results. First, it appears that, on average, there is a consistent change in vocabulary from month to month – regardless of the subreddit category. We see smaller changes in the active user cohorts from one month to the next, on average, however. Second, there is a statistically significant difference (KS test, $p < .01$) between the monthly (vocabulary and active user cohort) changes seen by subreddits in $\mathfrak{D}_I$ and all other subreddits in our dataset. This is evident by the observed bimodal RBO distance distribution seen in $\mathfrak{D}_I$. To ensure the consistency of our results related to the evolution of vocabulary vectors (which were gathered on a 10% sample of our dataset), we repeated our analysis on three independent 10% samples and confirmed the statistically significant differences between the RBO distance distributions between consecutive months for subreddits in $\mathfrak{D}_I$ and all other subreddits in our dataset. Further, manual validation confirmed that the months showing higher RBO distance to the prior months were the result of abnormal activities. For example, the *r/The_Donald* subreddit observed anomalous evolution of active user cohort in late 2015 when a migration of active users from *r/european*, an extremist subreddit which was eventually quarantined and banned by Reddit, was observed. Other large migrations appear to occur on $\mathfrak{D}_I$ subreddits throughout their lifespan. One hypothesis is that, similar to the above outlined case of *r/european*, eventually problematic subreddits see large migration events when currently problematic subreddits are banned. This hypothesis is supported by the results from a previous study (Ribeiro et al. 2020). Put another way, active users of a banned community migrate to a new community which eventually sees the same administrative action imposed on it due to the eventual occurrence of the same problematic behaviors. Studying the largest changes in the $\mathfrak{D}_A$ subreddits, we see that *r/feminism* had an RBO vocabulary distance of .65 when comparing 2016/10 and 2016/11. Closer inspection shows that the vocabulary change is largely driven by Hillary Clinton's loss in the 2016 US Presidential elections. Other prominent examples of $\mathfrak{D}_A$ subreddits with large changes were: the active user cohort for *r/Australia* during 01/2020 which corresponds to the outbreak of the wildfires and the active user/vocabulary of *r/newzealand* in 02/2019 following the Christchurch Mosque shootings. As a consequence of these real-world events, both subreddits saw increased activity from redditors not usually active on the subreddits. These observations suggest that the large changes in $\mathfrak{D}_A$ subreddits are often driven by external events while changes in $\mathfrak{D}_I$ subreddits are largely driven by on-platform administrative actions and community raids.

Takeaways. Our analysis confirms our hypothesis: *Subreddits may not converge to a stable vocabulary or user base.* On average, subreddit vocabulary evolves at a higher rate than subreddit active user cohorts. Interestingly, we see that the monthly changes observed by $\mathfrak{D}_I$ subreddits are, on average, statistically significantly higher than when compared to all other subreddits. This suggests that the differences in evolutionary patterns as well as an understanding of the causes for large changes (e.g., were they due to on-platform or real world events?) might allow for early detection of potentially problematic subreddits. For example, tools which identify the subreddits which are the targets of mass migrations from recently banned subreddits might facilitate early interventions to prevent degradation of the target community. We operationalize these insights in §4.

4 Identifying Predictors

In this section, we test the following hypothesis: **H2. Evolution into problematic subreddits can be predicted**. If valid, this hypothesis will show that tools may be built to help moderators pre-emptively identify subreddits likely to devolve into problematic subreddits. In order to test this hypothesis, we extract a variety of features from different points in a subreddit's lifespan and utilize explainable ML to understand the predictive capabilities of each of these features and perform early identification of problematic communities. We then evaluate the performance of our explainable models in a real world continuous learning setting, similar to how Reddit administrators may leverage them.

4.1 Methods

Extracting subreddit features. For each subreddit in our datasets, we break their lifespan into four quarters and extract features from each. With this approach: (1) we are able to get an identical number of features from all subreddits – even if they have vastly different lifespan values, and (2) we are able to capture features from different phases in the evolution of a subreddit. Our extracted features fall in six categories (listed in Table 1): community-, moderator-, user-, structure-, mentions, language-, and vocabulary-related features. These features were largely influenced by the insights from our analysis in §3 and existing literature seeking to predict community dynamics (§5).

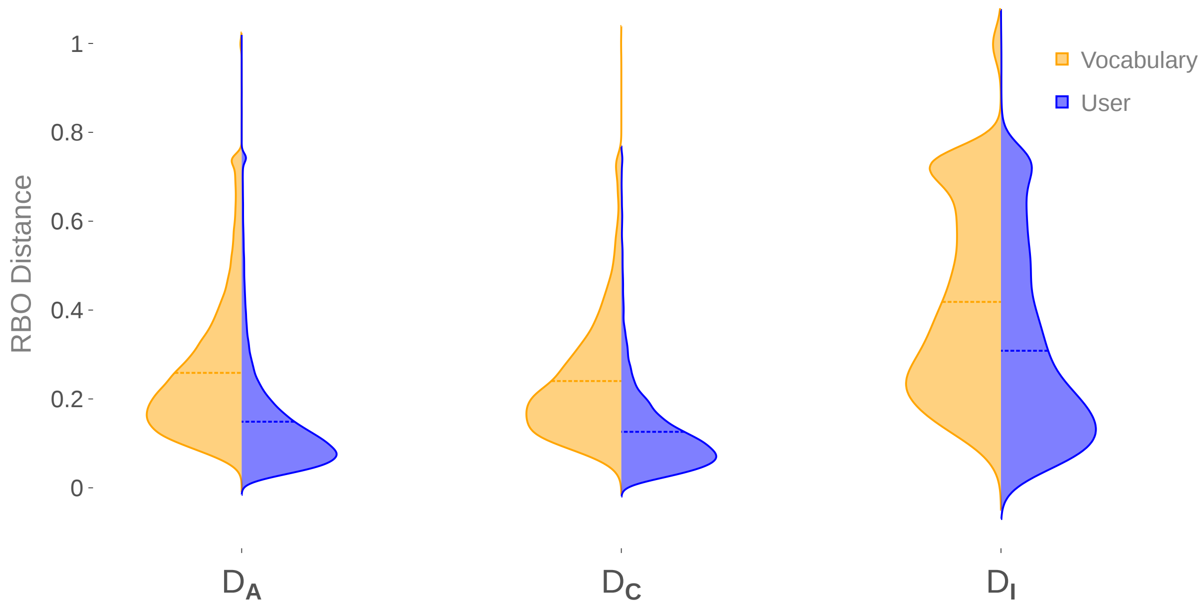

Figure 1: Distribution of RBO distances of vocabulary/user vectors between consecutive months for subreddits in $\mathfrak{D}_A$, $\mathfrak{D}_C$, and $\mathfrak{D}_I$.

Community features. This category of features captures the dynamics of the interactions occurring within the community – e.g., how large is the active community, how highly do community members rate each others posts, etc.

Moderator features. Moderators play a large role in directing the growth and policies within each community. This category of features captures how the moderator team interacts with the community – e.g., how many moderators does the community has, how many comments are removed by moderators, etc.

User features. This category of features captures characteristics of the average user within a community – e.g., how active are users, how frequently do they delete their comments, etc.

Structural features. We introduce a category of features to capture how a subreddit is connected (in terms of shared user base) to other communities – e.g., how isolated is the subreddit, what fraction of its connections are to other communities which were previously classified as problematic, etc.

Mention features. This category of features represent the mentions of a subreddit on other subreddits and in popular media. To obtain these features, we identified the number of: (1) news articles written about the specific subreddit prior to the end of the quarter being studied. The dates and article counts were obtained using MIT and Harvard's Media Cloud [2], an open source platform that gathers and tracks content of online news, with the search restricted to their U.S. Top Online News collection, and (2) references to the specific subreddit on comments made on other subreddits. Finally, for both types of mentions we compute the sentiment towards the community and categorize the mention as either negative or not.

Language features. This category represents the language style of the users in these communities. We use LIWC 2015 (Pennebaker et al. 2015) to extract language style features. These features help understand the psychometrics of the language within the community. In addition, we use the Perspective API [3], a toxic speech classifier developed by Google, to identify toxic comments. We also include the similarity of the community's vocabulary vector with the vectors of previously known problematic communities.

Preventing feature "leakage". Extracting features without careful considerations can result in leakage that impacts the quality of the classification task. To avoid this problem we need to make sure that features used in our task are actually available for use, by administrators, at the time of the classification task. For example in Q1 of a subreddit's lifetime which may be between months m_1 and m_2, we cannot extract features which might rely on data from after m_2 – from the community or from external communities. This is particularly important when considering the features in the structural, mentions, and vocabulary categories. In our feature extraction process for each subreddit, we take care to only consider information available from each quarter. For example, when extracting the '% of users with connections to previously banned communities' structural fea-

[2]https://mediacloud.org/

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

Category	Features
Community	# active unique commenters, # posts, # comments, dist. of comments & posts, dist. of score & comments, % of active user growth, dist. of controversial score per post, # controversial comments, # gilded comments
Moderators	# moderators, # incoming moderators, # outgoing moderators, dist. of moderators comment score, # AutoModerator comments, dist. of moderators comment score, # removed posts, # removed posts score
Users	# active months, # comments by deleted accounts
Structural	# uniquely connected communities, # total connections, # users with connections to previously banned communities
Mentions	# of mentions in other communities # of negative mentions in other communities, # of mentions in popular news outlets, # of negative mentions in popular news outlets
Language	93 standard LIWC 2015 features, # toxic comments, mean cosine similarity between vocabulary vectors of previously banned communities

Table 1: Features extracted from each quarter of a subreddit's life.

ture we only consider connections with communities which were banned before the end of the corresponding quarter, i.e., m_2, for the corresponding subreddit. Similarly, when creating vocabulary vectors, the vocabulary is limited to tokens observed only prior to m_2. This is maintained for every subreddit active during m_2. *Therefore, our features are all obtained from data that is available to administrators at the time of classification and are leakage-free*.

Accounting for class label imbalances in training and testing. Our dataset has a severe imbalance of class labels with only 264 $\mathfrak{D}_I$ subreddits and 3K $\mathfrak{D}_A$ subreddits. We take care to address the model-building and performance-reporting challenges that arise from this imbalance. We used and evaluated two standard approaches for model training in the presence of imbalanced class labels – oversampling from the minority class ($\mathfrak{D}_I$) using ADASYN (He et al. 2008) and undersampling from the majority class ($\mathfrak{D}_A$) using ensemble learning (Zhu, Xu, and Wu 2013; Liu, Wu, and Zhou 2009).

Approach 1: Oversampling $\mathfrak{D}_I$. ADASYN oversampling works by creating synthetic samples from the minority class. ADASYN is very similar to other synthetic oversampling techniques such as the SMOTE algorithm (Chawla et al. 2002). However, unlike SMOTE, ADASYN adaptively creates synthetic points while considering the neighborhoods of the class borders. Consequently, it generates synthetic points near the class borders and mitigates the challenges associated with overfitting (e.g., seen in SMOTE oversampling). For the purposes of training our model, we used ADASYN to create a perfectly balanced dataset with equal numbers of $\mathfrak{D}_I$ and $\mathfrak{D}_A$. Validation and testing were carried out only on samples not used or generated by ADASYN.

Approach 2: Undersampling $\mathfrak{D}_A$. Our undersampling approach works by: (1) splitting the training samples of the majority class ($\mathfrak{D}_A$) into equal sized datasets ($\mathfrak{D}_{A_1}$, $\mathfrak{D}_{A_2}$, ..., $\mathfrak{D}_{A_n}$) with the number of samples in each dataset equal to the total number of minority class ($\mathfrak{D}_I$) training samples, (2) training a set of classifiers, $c_1 \dots c_n$, with each using one of newly split majority class datasets and the entire minority class training samples as training input – i.e., classifier c_i trains on samples from $\mathfrak{D}_{A_i}$ and $\mathfrak{D}_I$, and (3) assigning the label output by the majority of the n classifiers when given a feature set for classification into $\mathfrak{D}_I$ or $\mathfrak{D}_A$.

Results for both sampling approaches are comparable and reported in §4.2. In addition to the aforementioned sampling techniques we repeat our experiments using our control dataset ($\mathfrak{D}_C$), SMOTE (Chawla et al. 2002), BorderlineSMOTE (Han, Wang, and Mao 2005), and random oversampling. Using these additional techniques yielded similar results. We note that the sampling approaches were utilized *only for expanding the training sets used and did not impact the testing and holdout datasets* in our classification experiments. To avoid the pitfalls with reporting accuracy in imbalanced class settings, we report F1 and AUC metrics.

Building interpretable models and extracting the predictive value of features. Given labels for each subreddit and a set of features associated with each stage in its lifetime, we now seek to understand the predictive values of these features. We achieve this in two steps: First, we build a machine learning classifier model which uses these features to predict the labels associated with each subreddit. Next, for high-performing classifiers, we analyze the weights associated with each feature by the classifier. Our argument is that *if a classifier is able to achieve a reasonably good performance, then the features it weighs heavily must have some predictive value*. Due to the need for transparency in such models and administrative tasks, *we focus solely on interpretable models (logistic regressions, decision trees, and random forests)*.

Classifier model training, validation, and testing. To evaluate the performance of each classifier model we first split the samples in our dataset with 80% of each class randomly allocated for training and validation and the remaining 20% reserved for holdout testing. We then used five-fold cross-validation to evaluate the classifier performance on the training and validation dataset. We apply our oversampling and undersampling strategies *only on the training samples in each fold*. Finally, we evaluate the classifier performance on the holdout set. In our results we report the mean F1-score and AUC in the holdout samples.

Interpreting models. Logistic Regression models a relationship between an outcome variable y and a group of predictor variables in terms of log odds. In order to interpret the model, we compute the estimated weights for each feature and their corresponding odds ratio (Molnar 2019). If the odds ratio for a feature (f) is x, it means that a unit increase in f changes the odds of our outcome variable y by a factor

Model	Classifier Score (%AUC, %F1-Negative, %F1-Positive)			
	Q1	Q1+Q2	Q1+Q2+Q3	Total
RF-AS	$(93, 97, 68)$	$(95, 98, 78)$	$(96, 99, 84)$	$(97, 99, 88)$
RF-US	$(91, 93, 65)$	$(91, 92, 68)$	$(92, 94, 74)$	$(94, 95, 80)$
LR-AS	$(79, 79, 59)$	$(80, 81, 60)$	$(79, 83, 65)$	$(85, 92, 70)$
LR-US	$(77, 76, 57)$	$(80, 82, 59)$	$(81, 83, 63)$	$(84, 86, 68)$

RF-AS = Random Forest with ADASYN sampling.
RF-US = Random Forest with random undersampling.
LR-AS = Logistic Regression with ADASYN sampling.
LR-US = Logistic Regression with random undersampling.

Table 2: Performance of our classifiers in predicting $\mathfrak{D}_I$ and $\mathfrak{D}_A$ labels at the end of each quarter of a subreddit's lifespan. Values denote area under the ROC, F1-score for non-problematic subreddits (F1-Negative), F1-scores for problematic subreddits (F1-Positive) on the holdout set.

of x when all other features remain the same. By calculating the features with the highest odds ratios for different labels, we are able to identify which features are the best predictors of problematic subreddits, as later decided by Reddit administrators. For our decision tree and random forest models, we find the importance of each feature using Gini Importance (Breiman et al. 2017). At a high-level the Gini importance counts the number of times a feature is used as a splitting variable, in proportion with the fraction of samples it splits. For random forests, the Gini importance is averaged over all the constructed trees. We expect more important features to have higher Gini importance scores. Unlike logistic regression interpretation, a limitation here is that this metric only allows us to rank feature importance, but not quantify the relative difference of their importance.

4.2 Results

Can we identify problematic subreddits by their evolutionary features? Column "Total" of Table 2 shows how our different explainable classifier models performed at classifying subreddits into $\mathfrak{D}_I$ and $\mathfrak{D}_A$ when given access to all evolutionary features of the subreddit, as would be available at the end of the quarter Q4 of a subreddit's lifespan (which for $\mathfrak{D}_A$ subreddits is the last month of data used in this study – 04/2020). As we can see all our models perform reasonably well, achieving F1-scores as high as 95% on our holdout set and a mean F1-score of up to 88% in our five-fold cross-validation experiments, regardless of whether models were built using ADASYN oversampling or majority class undersampling. More interestingly, we notice that our classifiers are able to achieve high F1-scores even as early as after Q1 (between 91% to 95% AUC) with performance only increasing with longer observations of a subreddit's evolution. These results indicate that, by observing the evolutionary features described in §4.1, problematic subreddits can be identified much earlier than they currently are. Further, the performance of our interpretable classifiers are reasonable enough to warrant their use to understand feature importance and the predictors of problematic subreddits.

What features are most important for predicting the evolution into problematic subreddits? Our random forest and logistic regression model were in agreement for the top-5 most important predictive features with slightly different ordering. We found that the most predictive feature in both models is the 'number of users who were once active on banned communities'. This feature had the highest log-odds ratio of .32 in our LR models while simultaneously ranking as the most important in our RF models. This lends additional validity to our findings in §3. Similarly, other features ranked in top-5 important features by the random forest include average percentage of toxic comments (log odds ratio: 0.20), negative mentions in other communities (log odds ratio: 0.24), percentage of comments removed (log odds ratio: 0.18), and negative mentions in media outlets (log odds ratio: 0.20). These results suggest that communities which entertain users with interactions in previously banned communities have a significantly higher likelihood of becoming problematic as well. Therefore, administrator interventions on problematic subreddits, rather than the users of problematic subreddits, may not be the most effective strategy for preventing the re-occurrence of problematic subreddits.

How well do explainable classifiers do as a real-world administrative tool? Our previous results reflect our classifier performance over the entire dataset of $\mathfrak{D}_I$ and $\mathfrak{D}_A$. We now seek to understand how well our classifier would perform in a real-world deployment as a tool to aid Reddit administration. We design a continuous learning (Chen and Liu 2018) experiment which emulates how Reddit administration would use our classifier – with input from human administrators. A continuous learning framework is important because content policies of Reddit have changed significantly over time and leveraging a single snapshot of problematic subreddits does not allow for our models to learn new patterns associated with subreddits which violate newly added content guidelines – e.g., guidelines regarding the incitement of violence were only added in 10/2017 therefore a model trained largely on prior data would have little ability to identify problematic communities in this category.

Experiment setup. First, we begin by training our RF-ADASYN model on data from 01/2018 to 06/2018. Next, we obtain the subreddits identified as problematic by this model based on features obtained from subreddits in 07/2018 only. From this list, we pay attention to three cases: (1) the identified subreddit was eventually banned or quarantined by Reddit some time after 07/2018 (true positives), (2) the identified subreddit was not banned or quarantined by Reddit (false positives), and (3) ban or quarantine decisions that were made in 07/2018 that were not identified as problematic by our model (false negatives). We then use the false negatives as new training samples for the classifier so it may learn from the human administrator's decisions while performing classifications for subsequent months – therefore mitigating the challenges associated with an evolving content policy. This process is repeated for each month between 07/2018 and 04/2020. The false positives identified in each month are indicative of problematic subreddits that have not yet been identified as such by administrators and are potentially yet unknown to them while the true positives serve as validation for our model's performance and allow us to quantify the time advantage that proactive strategies yield – i.e., how much earlier they are able to identify problematic communities. We note that the results obtained in this setup are not comparable to our previous results for several reasons including: (1) different training periods – our current setup leverages only data from 01/2018 as opposed to our previous experiment which included data from 2015/01, (2) different testing duration – our test features are obtained from just one month of the subreddit's lifespan as opposed to an entire quarter, and (3) our model is retrained each month with new administrator-identified problematic subreddits. It is precisely these differences, which appear in a real world deployment, that warrant this experiment.

Results. Our model reported a total of 106 true positives and 26 false negatives. As one might expect from the continuous learning setup, the false negative rate decreased and the true positive rate increased over time. The 106 true positives included subreddits banned for toxicity (e.g., *r/TheRedPill* and *r/The_Donald*) and piracy (e.g., *r/soccerstreams*), amongst others. Our model identified them as problematic 9.3 months (mean) prior to their ban date by Reddit. Across the entire continuous learning experiment (until 04/2020), our model identified 43 subreddits as problematic that have not yet received administrative actions. These include *r/KotakuInAction*, *r/TumblrInAction*, *r/metacanada*, and *r/MensRights*. We note that there have been several controversies and many reports of toxic behavior (e.g., overt misoginy and racism) in these communities which support our model's decision. For example, *r/KotakuInAction* was Reddit's primary pro-GamerGate community. In fact, in an effort to prevent the spread of toxicity, the subreddit was made private for a brief period during the peak of the movement. More recently, the community has expressed strong anti-transgender sentiment in the form of slurs and hate speech. Our classifier identified it as problematic based on features associated with vocabulary, and exceptionally high toxicity and negative media mentions (>2 standard deviations from the mean). It remains unclear if the subreddits in our false-positives are receiving administrative attention from Reddit. In total our model suffered 26 false negatives. A large fraction of these were subreddits associated with eating disorders (e.g., *r/proED*, *r/EDFood*, and *r/thinspo*) which were simultaneously banned due to their violation of a content policy regarding 'encouraging self-harm'. We found that this was the first case of administrative action against such subreddits. As evidence of success in the continuous learning framework, we note that the subsequent quarantining of *r/thinspocommunity* was correctly predicted by our model. Taken together, our qualitative analysis suggests that our models are effective and deployable in the real world as an administrative aid.

Takeaway: Can evolution into problematic subreddits be predicted? Our results show that evolutionary features can be used to identify subreddits likely to be problematic in the future. This finding validates hypothesis H2. Our feature analysis which identified 'number of users who were once active on banned communities' as the most predictive suggests that interventions aimed at users of banned communities might be an effective strategy to mitigate problematic behavior. The explainable models also perform well in a continuous learning real world deployment – suggesting that they make effective administrative aids.

5 Related Work

We make contributions in two dimensions: we perform measurements to understand how vocabulary and user bases of online communities change over time (§3) and then identify the predictors of problematic communities (§4). We break down the related work in each of these dimensions.

Evolution of online communities. Studying behavioral patterns and evolution in online communities has been the subject of several research efforts. These efforts can be taxonomized by whether the goal is to understand evolution of interaction quantity or quality. Research in characterizing interaction has generally focused on understanding how the amount of interaction occurring in a community changes over time and under different conditions. A general approach is to model community interactions as a network graph where edges denote interactions (e.g., messages sent between two users) between nodes (i.e., community members) and track their evolution under different conditions. Especially relevant to our work is research from Crandall et al. (Crandall et al. 2008) which among other results showed that *interaction network related features are predictive of future user behavior in topic-centered communities*. Researchers have also tried to distinguish communities using interactive and linguistic features. Mensah et al. (Mensah, Xiao, and Soundarajan 2020) observe growing and failing subreddits in an attempt to distinguish their evolution using user interaction and language patterns. Although their results show that there are no significant differences in these features for growing and failing communities, their results suggest the possibility of using interaction and linguistic features as classifiers of other classes of communities. Several studies have also investigated how specific user interactions are influenced by the age of a community. Kiene et al. (Kiene, Monroy-Hernández, and Hill 2016) showed that after a certain point in the life-cycle of a community, large influxes of users had no impact on the quality of discourse within the subreddit. *These studies highlight the need to consider age and stability of a community when predicting its evolution*. Danescu et al. (Danescu-Niculescu-Mizil et al. 2013) found that linguistic features in a community were constantly evolving and found that its newest members were most likely to adapt their own linguistic features to those of the community. Gazan (Gazan 2009) found that, when communities stabilized, topics tended to move away from topical and factual to personal and social. This generally resulted in increased participation, often at the cost of conflict and factionalism. The importance of external events is highlighted by Zannettou et al. (Zannettou et al. 2018, 2017) who focused on the evolution of memes and news sources within communities and uncovered their influence on external communities. Focusing exclusively on Reddit, Mills et al. (Mills 2018; Mills and Fish 2015) showed that, for *r/The_Donald* and *r/Sanders4President*, external events and their community participation guidelines were largely responsible for their rise in popularity and large influx of users. *These studies highlight the need to consider cross-community interactions and external events when considering evolution of communities*. In terms of methods, we find most similarity between our approach and the work of Matias (Matias 2016) which used a logistic regression model to attribute weights to survey-derived features to uncover the factors associated with moderators and subreddits participating in the Reddit-wide blackout of 2015 – in protest of Reddit's administrative actions. They uncovered a strong correlation between moderator participation in meta-reddit subreddits and community participation in the protest. *These findings further highlight the important role played by a few key members (elites and moderators) in a community*.

Predicting future community behavior. To maintain civil behavior in online communities timely identification and removal and violators is necessary as observed by Scrivens et al. (Scrivens, Davies, and Frank 2020), they measure the evolution of radical posts against particular vulnerable groups over-time. Their results show approval (upvotes) shown towards hate speech increases gradually as users consistently and frequently keep posting hate speech.These results corresponds with previous works which show extremist communities polarizes opinion over time (Caiani and Kröll 2015; Wojcieszak 2010; Simi and Futrell 2015). Furthermore, Seering et al. (Seering et al. 2019) conduct a semi-structured interview with moderators to find, among many other things, that inconsistent moderation of communities lead to communities evolving chaotically then requiring constant moderation and community policy updates. Massanari (Massanari 2017) conducted a qualitative analysis of the Reddit communities at the center of the Fappening and Gamergate controversies. The study highlights how the inaction of Reddit administrators and community moderators resulted in the emergence of toxic technocultures and argues for the exploration of alternative designs and moderation tools to combat the spread of such toxicity. *These findings highlight the importance and need of timely moderation and intervention to maintain civil behavior*. Research in automated moderation for online communities have been mainly focused on content moderation at a 'content' level. Our work aims to aid administrators perform moderation at a community level. Chandrasekharan et al. (Chandrasekharan et al. 2019) created CrossMod, a tool to aid Reddit moderators by detecting and moderating comments. Similarly, Pavlopoulos et al. (Pavlopoulos, Malakasiotis, and Androutsopoulos 2017) and Santos et al. (Santos, Osman, and Schorlemmer 2021) developed and trained machine learning models to detect violations by users on Wikipedia edits.

6 Conclusion

Implications for other social platforms. Due to the similarities in community structure and platform designs, our methodologies to measure evolution of communities and detect problematic communities has the potential for application on other social media platforms such as Facebook. For example, the concepts of communities, posts, comments, user migrations, and administrator interventions have direct parallels with Facebook groups. However, further investigation is needed to understand whether similar evolutionary patterns can be exploited to develop moderation-aids on such platforms where the online disinhibition effect (Suler 2004) might be weaker since users are not anonymous to their communities and user accounts are required to reveal their real-world identities (Facebook 2021).

The challenge of human-only moderation. Currently, Reddit employs a small number of human administrators (Lagorio-Chafkin 2018) to identify communities in violation of the Reddit content policy and intervene to prevent future violations by those communities. Due to the growing size of the platform, rather than seeking consistency in policy enforcement, administrators are often only able to act in response to user generated and media reports of egregious violations of the content policy. Compounding their challenges, in §3, we showed that Reddit communities, on average, are constantly evolving in both vocabulary and active user bases. This implies the need for constant monitoring and attention, from moderators and administrators, to proactively identify problematic communities – a prohibitively expensive proposition for human-only administration given the large size of the platform. As regulations surrounding online social media companies liability in content moderation (e.g., §230 of the US Penal Code) are being re-evaluated world over, there is an urgency to develop tools to aid human administrators perform such proactive identification and interventions at scale.

Proactive moderation using predictive strategies. We observe that the evolutionary characteristics of problematic subreddits is different from other subreddits. This yields opportunities for providing machine-assisted human moderation. We exploit these differences in evolutionary characteristics to build simple, explainable, and accurate machine learning models to characterize the current and predict the future behavior of different communities. The accuracy of our predictions suggest that tools based on our approach and features can be used to identify communities that are likely to exhibit behavior similar to known problematic subreddits in the future. Therefore, the output of these tools can be used by administrators to proactively focus moderation efforts on a smaller set of communities. It is important to keep in mind that such proactive approaches, when used autonomously, may have negative consequences. For example, there have been reports of discrimination of LGBTQ content creators by YouTube's automated content moderation system (Farokhmanesh 2018) and, more critically, strong racial profiling by autonomous predictive policing systems (Angwin 2016). Therefore, we only recommend using such tools to assist human moderators by emphasizing which communities may require special attention.

Acknowledgements

This work is dedicated to the memory of Dr. Emma Mumper. Her support and creativity were instrumental to this project. The authors are grateful to the anonymous reviewers for their feedback. This material is based upon work supported by the Air Force Office of Scientific Research under award # FA9550-20-1-0346. The opinions, interpretations, conclusions, and recommendations are those of the authors and are not endorsed by the funding bodies.

References

Angwin, J. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2019-04-23.

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 & Social Media*.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 2017. *Classification and regression trees*. Routledge.

Caiani, M.; and Kröll, P. 2015. The transnationalization of the extreme right and the use of the Internet. *International Journal of Comparative & Applied Criminal Justice*.

Chandrasekharan, E.; Gandhi, C.; Mustelier, M.; and Gilbert, E. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the Conference on Computer Supported Cooperative Work*.

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. *Proceedings of the 2017 Conference on Computer Supported Cooperative Work*.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*.

Chen, Z.; and Liu, B. 2018. Lifelong machine learning. *Synthesis Lectures on AI & Machine Learning*.

Collins, B. 2017. Reddit Bans Forum Inciting 'Physical Removal' of Democrats From Society. https://www.thedailybeast.com/reddit-bans-forum-inciting-physical-removal-of-democrats-from-society. Accessed: 2019-04-23.

Collins, B.; and Zadronzny, B. 2018. After Toronto attack, online misogynists praise suspect as 'new saint'. https://www.nbcnews.com/news/us-news/after-toronto-attack-online-misogynists-praise-suspect-new-saint-n868821. Accessed: 2020-04-21.

Crandall, D.; Cosley, D.; Huttenlocher, D.; Kleinberg, J.; and Suri, S. 2008. Feedback Effects Between Similarity and Social Influence in Online Communities. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *Proceedings of the 22nd International Conference on World Wide Web*.

Dunning, T.; and Friedman, E. 2014. *Practical Machine Learning: Innovations in Recommendation*. " O'Reilly Media, Inc.".

Facebook. 2021. What names are allowed on Facebook? www.facebook.com/help/112146705538576. Accessed: 2021-12-09.

Farokhmanesh, M. 2018. YouTube is still restricting and demonetizing LGBT videos — and adding anti-LGBT ads to some. https://www.theverge.com/2018/6/4/17424472/youtube-lgbt-demonetization-ads-algorithm. Accessed: 2019-04-22.

Gazan, R. 2009. When Online Communities Become Self-Aware. In *2009 42nd Hawaii International Conference on System Sciences*.

Han, H.; Wang, W.-Y.; and Mao, B.-H. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*.

He, H.; Bai, Y.; Garcia, E. A.; and Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks*.

Kiene, C.; Monroy-Hernández, A.; and Hill, B. M. 2016. Surviving an "Eternal September": How an Online Community Managed a Surge of Newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

Lagorio-Chafkin, C. 2018. *We Are the Nerds: The Birth and Tumultuous Life of Reddit*.

Liang, D.; Altosaar, J.; Charlin, L.; and Blei, D. M. 2016. Factorization Meets the Item Embedding: Regularizing Matrix Factorization with Item Co-Occurrence. In *Proceedings of the 10th ACM Conference on Recommender Systems*.

Liu, X.; Wu, J.; and Zhou, Z. 2009. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*.

Massanari, A. 2017. # Gamergate & The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*.

Matias, J. N. 2016. Going dark: Social factors in collective action against platform operators in the Reddit blackout. In *Proceedings of the 2016 CHI conference on human factors in computing systems*.

Mensah, H.; Xiao, L.; and Soundarajan, S. 2020. Characterizing the Evolution of Communities on Reddit. In *International Conference on Social Media & Society*.

Mills, R.; and Fish, A. 2015. A computational study of how and why reddit. com was an effective platform in the campaign against sopa. In *International Conference on Social Computing & Social Media*.

Mills, R. A. 2018. Pop-up political advocacy communities on Reddit. com: SandersForPresident and The Donald. *AI & Society*.

Molnar, C. 2019. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.

Morse, J. 2019. Reddit waits until it's too late to ban violence-glorifying subreddits. https://mashable.com/article/reddit-bans-violent-subreddits-christchurch. Accessed: 2019-04-21.

Ohlheiser, A. 2016. Fearing yet another witch hunt, Reddit bans ‘Pizzagate’. https://www.washingtonpost.com/news/the-intersect/wp/2016/11/23/fearing-yet-another-witch-hunt-reddit-bans-pizzagate/. Accessed: 2020-04-21.

Pavlopoulos, J.; Malakasiotis, P.; and Androutsopoulos, I. 2017. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993*.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of LIWC2015. Technical report.

Ribeiro, M. H.; Jhaver, S.; Zannettou, S.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and West, R. 2020. Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald & r/Incels. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work & Social Computing*.

Roberts, S. T. 2014. *Behind the screen: The hidden digital labor of commercial content moderation*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Romano, A. 2017. Why Reddit won’t ban The_Donald. https://www.vox.com/culture/2017/11/13/16624688/reddit-bans-incels-the-donald-controversy. Accessed: 2019-04-23.

Santos, T. F. d.; Osman, N.; and Schorlemmer, M. 2021. Learning for Detecting Norm Violation in Online Communities. *arXiv preprint arXiv:2104.14911*.

Scrivens, R.; Davies, G.; and Frank, R. 2020. Measuring the evolution of radical right-wing posting behaviors online. *Deviant Behavior*.

Seering, J.; Wang, T.; Yoon, J.; and Kaufman, G. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society*.

Simi, P.; and Futrell, R. 2015. *American Swastika: Inside the white power movement’s hidden spaces of hate*.

Suler, J. 2004. The online disinhibition effect. *Cyberpsychology & behavior*.

Webber, W.; Moffat, A.; and Zobel, J. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.*

Wohn, D. Y. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference*.

Wojcieszak, M. 2010. ‘Don’t talk to me’: Effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media & Society*.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *Proceedings of the 2018 Internet Measurement Conference*.

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Kourtelris, N.; Leontiadis, I.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream & Alternative News Sources. In *Proceedings of the 2017 Internet Measurement Conference*.

Zhu, M.; Xu, C.; and Wu, Y.-F. B. 2013. IFME: Information Filtering by Multiple Examples with under-Sampling in a Digital Library Environment. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*.