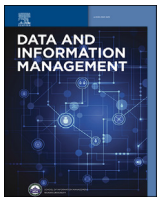
Data and Information Management 6 (2022) 100019

Contents lists available at ScienceDirect

Data and Information Management

journal homepage: www.journals.elsevier.com/data-and-information-management



PROVOKE: Toxicity trigger detection in conversations from the top 100 subreddits

Hind Almerekhi [a,*], Haewoon Kwak [b], Joni Salminen [c], Bernard J. Jansen [d]

[a] College of Science & Engineering, Hamad Bin Khalifa University, Doha, Qatar
[b] School of Computing and Information Systems, Singapore Management University, Singapore, Singapore
[c] School of Marketing and Communication, University of Vaasa, Vaasa, Finland
[d] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

ARTICLE INFO

Keywords:
Online toxicity
Conversation threads
Reddit
Toxicity triggers
Neural networks
Social media

ABSTRACT

Promoting healthy discourse on community-based online platforms like Reddit can be challenging, especially when conversations show ominous signs of toxicity. Therefore, in this study, we find the turning points (i.e., toxicity triggers) making conversations toxic. Before finding toxicity triggers, we built and evaluated various machine learning models to detect toxicity from Reddit comments.

Subsequently, we used our best-performing model, a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model that achieved an area under the receiver operating characteristic curve (AUC) score of 0.983 to detect toxicity. Next, we constructed conversation threads and used the toxicity prediction results to build a training set for detecting toxicity triggers. This procedure entailed using our large-scale dataset to refine toxicity triggers' definition and build a trigger detection dataset using 991,806 conversation threads from the top 100 communities on Reddit. Then, we extracted a set of sentiment shift, topical shift, and context-based features from the trigger detection dataset, using them to build a dual embedding biLSTM neural network that achieved an AUC score of 0.789. Our trigger detection dataset analysis showed that specific triggering keywords are common across all communities, like 'racist' and 'women'. In contrast, other triggering keywords are specific to certain communities, like 'overwatch' in r/Games. Implications are that toxicity trigger detection algorithms can leverage generic approaches but must also tailor detections to specific communities.

1. Introduction

Many communication channels that online platforms offer users facilitate information exchange effortlessly. One popular form of user interaction occurs in online communities, where participants converse about common interests (Mohan et al., 2017). To interact via online communities, some users rely on social media platforms like Facebook, Twitter, and Reddit because they offer various tools for establishing, joining, and discovering communities and like-minded community members (Mittos et al., 2020). This community-based interaction garnered much traffic for online platforms; for example, Reddit reported 30% growth from 330 million to 430 million[1] active monthly users.

Due to this rapid growth, more users engage in online communities by publishing content and colloquially responding to each other. However, the favorable increase in traffic within online communities poses a significant challenge to community moderators, especially with the increasing amounts of content that might include malice (Almerekhi et al., 2020a). To ensure that conversations remain civil in online communities, moderators regulate content to prevent the spread of toxicity (Obadimu et al., 2021): content that most users deem rude, hateful, offensive, or harassing (Risch & Krestel, 2020, pp. 85–109). Online communities discourage harmful content in conversation threads due to its adverse effects on the user experience and its role in aggravating users by fueling conflicts between community members (Kumar et al., 2018).

The intentions behind leaving toxic content in online communities vary; some users post harmful content to cyberbully or harass others, whereas others publish toxic content to troll or send hate toward various groups like racial or religious minorities (Dubois et al., 2022; Pronoza et al., 2021; Wulczyn et al., 2017; Zhao et al., 2016). Thus, online platforms have devised several strategies that rely on community members,

* Corresponding author.
E-mail address: hialmerekhi@hbku.edu.qa (H. Almerekhi).
[1] https://backlinko.com/reddit-users, retrieved on Apr. 27, 2022

https://doi.org/10.1016/j.dim.2022.100019
Received 16 May 2022; Received in revised form 27 July 2022; Accepted 15 August 2022

content moderators, and automatic moderation tools to hinder the spread of toxicity. The community moderators establish clear guidelines and ensure that community members adhere to the participation regulations (Riedl et al., 2021). Some moderators rely on counterspeech to respond to toxic content while others automatically remove or delete it and penalize the offending users. The latter approach is common in active communities due to its effectiveness in countering the rapid spread of toxicity (Nobata et al., 2016; Warner & Hirschberg, 2012; Watanabe et al., 2018). Therefore, the automatic detection of toxicity is favored by most moderators over manual moderation, primarily because it is exceptionally challenging to moderate thousands of posts and comments manually.

Although many studies investigate the problem of detecting toxicity in online content (Carton et al., 2020; Kwon & Gruzd, 2017), fewer studies focus on the causes of toxicity in online conversations (Almerekhi et al., 2019, 2020b). Toxicity in conversation threads can lead to a chain reaction in the form of toxic responses. Thus, it is crucial to identify (i.e., triggers) toxicity's causes to stop this toxic chain reaction, which can be an early mechanism for preventing the further spread of toxicity in online conversations (Kwon & Gruzd, 2017). Furthermore, the threaded nature of conversations enables detecting toxicity triggers by identifying consecutive toxic comments' starting points (Almerekhi et al., 2019).

Thus, in this study, we hypothesize that online conversations turn toxic due to particular starting points (i.e., triggers) making healthy conversations toxic. The challenge in identifying toxicity triggers is that they can differ according to context, community, norms, and language (Weninger et al., 2013). The earliest mention of the concept 'trigger' was in information behavior research, where 'trigger' represented a cause that made users start seeking information (Orton et al., 2000). Reddy and Jansen (Reddy & Jansen, 2008) view a 'trigger' as a juncture for moving from individual to cooperative information-seeking behavior that needs interaction, a sequence of occurrences requiring at least two actors and two actions (Wagner, 1994).

This work aims to detect toxicity triggers (i.e., toxicity causes) leading to toxicity in conversations. For this purpose, we describe toxicity triggers as *nontoxic starting points* coming before toxic engagements in conversation threads on social media platforms. To address this research goal, we propose two research questions to detect, characterize, and analyze toxicity triggers:

• RQ1: How can toxicity triggers in online conversations be detected across different communities?

• RQ2: What are the characteristics of toxicity triggers across different communities?

To address these research questions, we begin by analyzing a collection of 1,126,570,077 comments collected from the top 100 Reddit communities between March 2006 and April 2020. We chose Reddit to conduct this research due to its conversational nature and the availability of data for conducting research. We start with detecting toxic comments and rebuilding conversations based on the relationship between the comments and the replies. Then, we identify toxicity triggers, which we generally define as non-toxic parent comments with toxic children (i.e., replies).

We first try to understand the textual characteristics of toxicity triggers by comparing them with non-triggers. Then, an exhaustive literature review leads us to focus on the characteristics of conversation threads that can potentially link to what causes the initiation of toxic behavior. For example, prior research revealed that a *topic shift* could lead to a change in the sentiment tone of the conversation (Topal et al., 2016). Thus, it might also lead to toxicity triggers.

Moreover, we analyze the sentiment shifts in conversation posts and comments. The former captures the use of language within the conversation threads, and the latter captures the thread's toxic sentiment. Our premise is that, similar to a topical shift, these shifts in semantics and emotion may also be attributes of toxicity triggers (Oussalah et al., 2018). Another indicator of toxicity triggers is the conversation context (Tan et al., 2016), which captures the intents and interactions among users that trigger toxicity.

By focusing on these aspects, we analyze the impact of the topical and sentiment shifts and context relative to the occurrence of toxicity triggers across multiple communities on Reddit. Finally, we build a prediction model based on textual features, context, topic, and sentiment shift features. The contributions of this work are threefold:

- First, we extend the study (Almerekhi et al., 2020b) by formally redefining toxicity triggers and characterizing triggers using trigger-specific keywords.
- Second, we build a prediction model for detecting toxicity triggers. Furthermore, our study is one of the leading studies (Chong & Kwak, 2022) that focus on detecting toxicity triggers across many communities.
- Third, we make our toxicity-trigger detection dataset and our implementation scripts available[2] for the research community, including the gold standard labels and comments from 991,806 conversations. Reportedly, this is the only large-scale dataset and open source implementation for detecting toxicity triggers.

2. Related work

2.1. Toxicity detection

Scholars explored the problem of detecting hate and toxicity on various social media platforms, including YouTube, Facebook, and Twitter (Nobata et al., 2016; Warner & Hirschberg, 2012). These studies, with others, show hate and toxicity's infectious nature and the importance of preventing such content on online platforms. The impact of toxic content is prevalent among social media users. For instance, 41% of American adults report experiencing online harassment, 25% report experiencing severe forms of harassment, and 79% state that social media companies are fairly or poorly addressing online harassment or bullying on their platforms (Vogels, 2020). These percentages show that negative experiences on social media platforms, such as being exposed to toxicity or harassment, can discourage users from engaging in online interactions and social media, not to mention the lousy reputation social media platforms gain for not impeding the spread of toxic content, which also discourages users from online interactions. Similarly, cyberbullying (Bosque & Garza, 2016; Hosseinmardi et al., 2015) incidents show that toxicity decreases users' attempts at engaging in conversations with other users.

The general problem of toxic comment detection focuses on one main goal: classifying social media content as toxic or non-toxic to determine what to do with the detected content (e.g., delete or analyze toxic content). Scholars leveraged machine learning techniques to build models for detecting hate (Fortuna et al., 2021; Nobata et al., 2016), including developing deep neural networks (Badjatiya et al., 2017) with different features to detect toxicity. For example, Nobata et al. (Nobata et al., 2016) used syntactic (Sood et al., 2012), linguistic (Chu et al., 2021; Warner & Hirschberg, 2012), n-gram (Watanabe et al., 2018), and word embedding (Zhao et al., 2016) features to predict abusive and non-abusive comments. The study's outcomes showed that machine learning models with content-based features do not require retraining word embeddings as in deep learning models.

Likewise, Wulczyn et al. (Wulczyn et al., 2017) used machine learning techniques to predict targeted personal attacks at different levels from Wikipedia comments. The study extracted features from text comments without considering additional information about users or their communication network. Just like (Nobata et al., 2016), the study used n-gram and linguistic features to detect personal attacks. Findings showed that about 30% of the personal attacks were from users with registered accounts with at least 100 contributions. Moreover, the

[2] https://github.com/Hind-Almerekhi/toxicityTriggers.

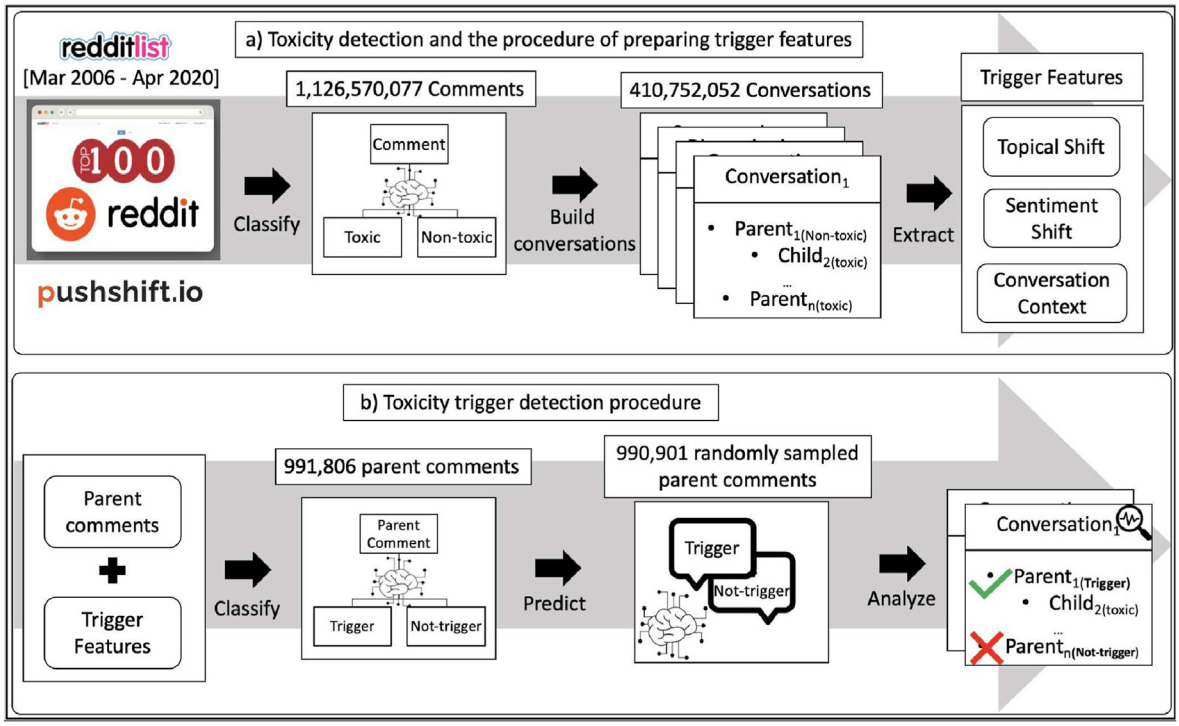


Fig. 1. Methodology pipeline to perform toxicity trigger detection from the top 100 subreddits on Reddit. The output from (a) is the trigger features input to (b).

personal attacks on Wikipedia users did not come from anonymous or toxic users (Wulczyn et al., 2017). Also, Spiros et al. (Georgakopoulos et al., 2018) leveraged the same Wikipedia dataset as (Wulczyn et al., 2017) to perform toxic comment classification with convolutional neural networks (CNNs). The study found that CNNs outperform text-based bag-of-words classification methods. On the other hand, Salminen et al. (Salminen et al., 2018a) discovered that in the comments on YouTube news, the targets of hate were mostly news media outlets and politicians.

Lastly, Jigsaw's Perspective API (Jain et al., 2018) is a tool that allows researchers and platforms to score comments for their toxicity. The backend models of the Perspective API used a vast collection of comments labeled by crowdsourcing efforts. Additionally, the API offers various experimental models focusing on specific toxicity detection problems. For example, earlier versions of the experimental models could classify obscenity, severity in toxicity, spam, and targeted attacks. The experimental models of the Perspective API are continuously updated to improve the performance of the scoring mechanism (Jain et al., 2018).

2.2. Contagion of toxicity

Besides investigating work on hate detection in online platforms, we found that previous studies confirmed that the contagion of toxicity holds serious infectious risks in online conversations (Mohan et al., 2017). A study by Del Vicario et al. (Del Vicario et al., 2016) dubbed this effect the emotional contagion. Similarly, Kwon and Gruzd (Kwon & Gruzd, 2017) studied the cascading effect that hatefulness forms in online conversations and found that when profanity appears in parent comments, the following child comments exhibit more profanity.

The notion of topic drift is essential to detecting toxicity triggers because, as we established earlier, conversation participants' emotions or sentiments strongly influence the tone and direction of conversations on social media platforms (Topal et al., 2016). For instance, Topal et al. (Topal et al., 2016) suggested that emotional shifts in online conversations could come from topical shifts. Hence, this type of shift can also detect changes in toxicity (Almerekhi et al., 2019), especially in a large dataset like this research.

Previous studies focused on multiple characteristics of toxic conversations from different standpoints, and the methods often focused on predicting and classifying toxic comments. However, few studies targeted toxicity triggers in conversation threads. Therefore, our method adds to the available and critical research question of detecting toxicity triggers by incorporating various methods and showing their importance in various fields.

3. Methodology

3.1. Research design

To identify toxicity triggers from conversation threads, we devise two main procedures, as illustrated by the top and bottom parts in Fig. 1. The procedures involve the following tasks: In the first procedure (denoted by a) in the top part of Fig. 1), we collect comments from different online communities and classify a sample of the obtained comments based on their toxicity. Next, we reconstruct conversations from the collected comments in a manner that helps detect toxicity triggers–if they appear in conversation threads. Then we derive features from the comments incorporating shifts in the topic and sentiment of conversations and the conversation context. In the second procedure (denoted by b) in the bottom part of Fig. 1), we train a classification model leveraging the obtained features from the first procedure to detect toxicity triggers in online conversations. Then we randomly sample a set of comments and use the trained trigger detection model to identify toxicity triggers. Lastly, we examine prediction results within conversation threads to identify misclassification errors' potential causes.

The following subsections describe the research context and explain the data collection and preprocessing, toxicity detection, and toxicity trigger detection methods.

3.2. Research context

Since Reddit's inception, users regard the website as an online

community including more than two million[3] communities, also known as 'subreddits'.One typically refers to a subreddit on Reddit by using 'r/' before its name, such as r/funny. Reddit is considered a 'congregation' of communities, where every subreddit is independent of others' content and rules, enriching the platform with diverse communities (Massanari, 2017). Registered Reddit users can engage in subreddits by posting and sharing content with articles, news, videos, and opinions on various topics. Users can create subreddits freely based on their topical interests, including memes, politics, music, video games, sports, and entertainment. Each subreddit is an independent community with its own rules and regulations. In most cases, the subreddit creator acts as a sole *moderator* or builds a moderation team to ensure that users follow community guidelines. Users can perform activities in each subreddit including a) creating posts (i.e., leaving posts in the community), b) leaving comments as responses to posts, and c) rating users' posts and comments (Choi et al., 2015).

To illustrate how Reddit works, Fig. 2 shows a post and a partial conversation thread from the subreddit 'r/AskReddit'. We omitted part of the conversation thread due to space limitations restricting viewing comments in a reply-chain manner. The conversation thread in Fig. 2 shows that users can respond (i.e., reply to comments) to any discussion participant, forming a conversation thread. Furthermore, other users can consider some comments toxic, as shown by the toxicity labels on every comment in the conversation thread from Fig. 2. In our analysis, we leverage the representation of a conversation in Reddit conversations to our advantage to detect toxicity triggers. By observing the types of comments in conversation trees within a post, we can see that the conversation tree includes (a) comments directly replying to the original post (first-level comments) and (b) comments replying to other comments or replies within the conversation-tree (child comments). To build the

parent–child relationships, we use the conversation's metadata, including the parent ID of each comment at every level in the conversation tree. The reconstructed conversations consist of a sequence of comments representing the post's topic (denoted by t). Each sequence of comments has a nesting level equalling the number of comments preceding the investigated comment in the conversation's reply chain (Topal et al., 2016).

3.3. Dataset

This extensive study focuses on the top 100 subreddits from August 2017 with the highest number of subscribers (i.e., the most extensive number of community members).[4] Then, for every subreddit, we collected all the comments between March 2006 and April 2020. In Table 1, we report the year, total number of subreddits (out of 100) from each year, and the number of comments in those subreddits. Only a handful of subreddits had comments in 2006. Moreover, when conducting the bulk of this study, 2020 only had data until April, which explains the lower number of comments in this particular year.

By performing toxicity detection on the comments from Table 1, we built a subset from the full dataset that contains conversations with potential toxicity triggers.

To build conversations, we relied on the parent ID, link ID, and comment ID. If the parent ID = link ID, the comment is a direct reply to the post, so it is a parent comment at the top level of the conversation thread. We used PostgresSQL for parent comments on other levels of the conversation thread[5]to look up all children of a given comment recursively. Comments that had no children were discarded (leaf comment). However, comments with children were considered a conversation thread consisting of a parent comment and sorted all its children chronologically. Using the conversations we built earlier, we defined a potential *toxicity trigger* as a non-toxic parent comment with at least one toxic child comment (Almerekhi et al., 2019). The last column from Table 1 shows the total number of conversations with potential toxicity triggers each year. The last row from Table 1 shows the total number of comments and conversations with potential toxicity triggers, accounting for 36.46% of the entire collection.

3.4. Data collection and preprocessing

To obtain the desired Reddit collection, we downloaded all the monthly files per year from Pushshift's public Reddit collection.[6] Next, we used MongoDB7[7] to query the files for comments that come from the top 100 subreddits only. When we obtained the comments, we guaranteed that the comments' JSON objects included the time stamp, parent comment's ID, comment's textual content, and comment's ID. After obtaining the comments, we discarded all the deleted comments because they were not helpful for our study of toxicity triggers. The total number of obtained comments from each subreddit is available in Tables A7 and A8, where we sorted subreddits in descending order based on the total number of comments in each subreddit. Table A7 shows that *r/AskReddit* has the highest number of comments (239,883,260), whereas Table A8 shows that *r/InternetIsBeautiful* has the least amount of comments (321,410). As for the collection statistics, the mean number of comments is 11,265,700.77, while the median is 3,926,523.

After performing toxicity detection on the comment collection, we constructed conversation threads from the comments using the IDs of parent and child comments. Finally, we excluded comments with no children since they were not useful for our study. This process resulted in an extensive collection of conversations from the top 100 subreddits, as in Tables B9 and B10. Just like the comment collection, Table B9 shows that *r/AskReddit* has the highest number of conversations (74,887,689), while Table B10 shows that *r/InternetIsBeautiful* has the lowest number of conversations (114,042). Further statistics from the conversations show that the mean is 94,107,520.52, and the median is 1,443,517.

3.5. Toxicity detection

While our ultimate goal is to find toxicity triggers, we first need to identify toxic comments; then, we can find their triggers. We first experimented with detecting toxic comments by scoring them with Google's Perspective API (Perspective, 2017). First, we scored 1000 random comments; then, we manually investigated the accuracy of the scores. However, results from our early version of the API were unreliable for toxicity detection–not to mention that, like any API, the Perspective API has a limited rate and maximum submission length for toxicity detection requests. Therefore, we could not use the Perspective API for largescale toxicity detection and instead built a toxicity detection model from a sample of the comments in our collection.

We had to obtain a labeled training set to build a machine learning model for supervised learning. Unfortunately, when conducting these experiments, we did not find any useful datasets from Reddit. Thus, we created a training set with crowdsourced labels from 10,000 sampled r/AskReddit comments. We chose this particular subreddit because it is one of the most prominent question–answer subreddits that deals with topics of varying toxicity levels (Lanius, 2019). The job required crowdsourcing workers to label given comments as either toxic or non-toxic based on the Perspective API toxicity definition–'a rude, disrespectful, or unreasonable comment that is likely to make you leave a conversation' (Perspective, 2017). Additionally, the task required that at least three annotators to agree on the label for a given comment. The results for the test collection comments showed that 81.57% were non-toxic while the remaining 18.43% were toxic. The observed agreement between

[3] https://frontpagemetrics.com/top; retrieved on Apr. 25, 2022
[4] http://redditlist.com; retrieved on Aug. 21, 2017
[5] https://www.postgresql.org; Retrieved on Jul. 24, 2022
[6] https://files.pushshift.io/reddit/, retrieved on May 22, 2019.
[7] https://www.mongodb.com, retrieved on May 30, 2019.

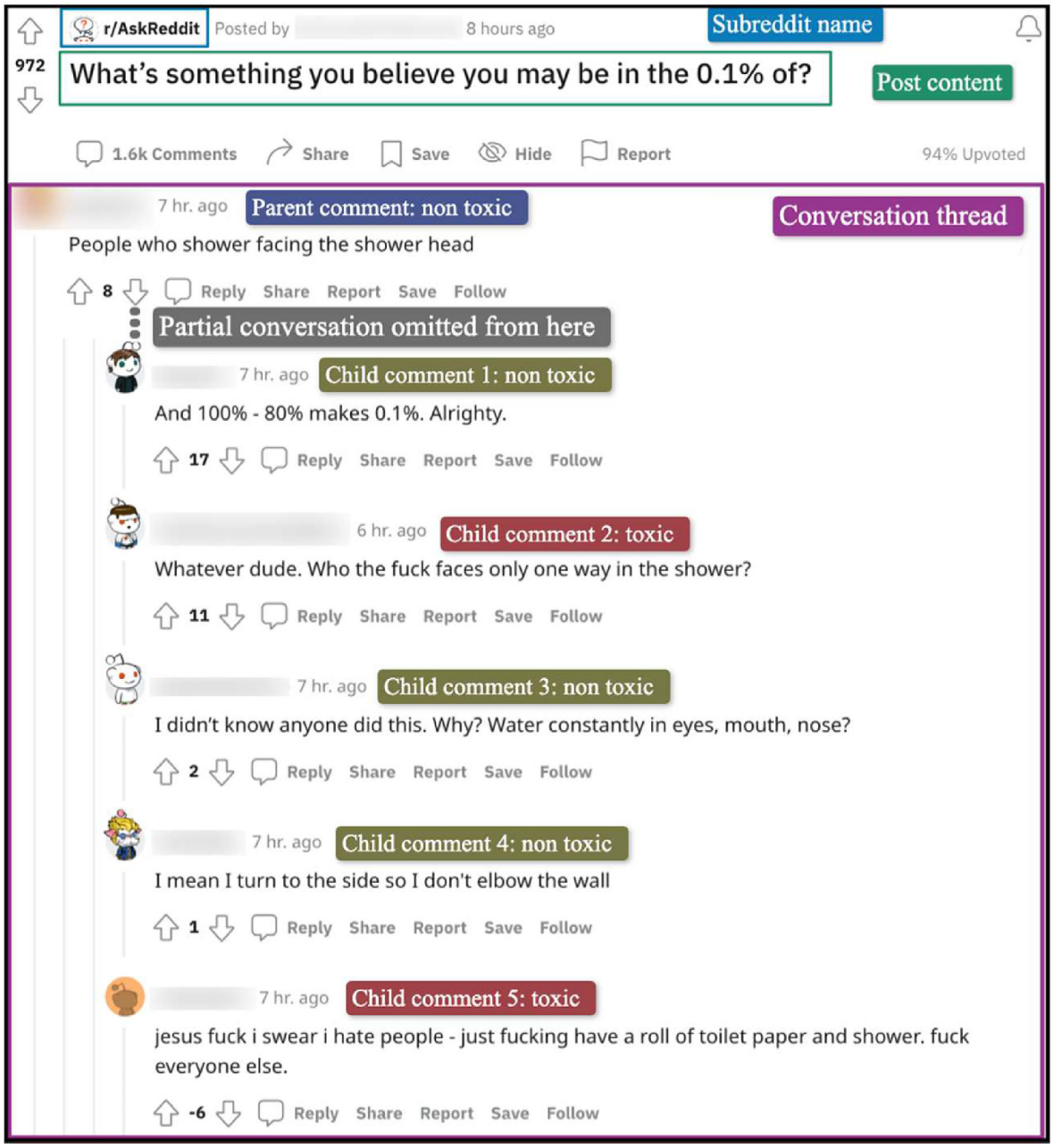


Fig. 2. A Reddit post (i.e., submission) in the subreddit "r/AskReddit" with non-toxic and toxic responses in the form of a conversation thread.

annotators was 0.85, the Gwet's gamma (Gwet, 2014) score was 0.70, and the Fleiss' kappa score was 0.48.

While the observed agreement is high, the Fleiss' Kappa score is relatively low; Feinstein et al. (Feinstein & Cicchetti, 1990) and Salminen et al. (Salminen et al., 2018b) heavily discussed this problem with the low Kappa score contributing to the class imbalance. On the other hand, Gwet's gamma (Gwet, 2014) and other average-distribution approaches handle such class imbalance differently and outperform the Kappa statistic measurements. Therefore, given the class imbalance in our dataset, representative of the online conversation community, we employ Gwet's measure in our research.

Then, we used machine learning techniques to create and evaluate models that predict comments as either toxic or non-toxic. We utilized various features and feature combinations that examine the comments' linguistic and stylistic properties. Foremost, we extracted n-gram features with varying configurations (Jansen et al., 2009); next, we refined our feature set by incorporating word-embedding features like word2vec (Kulkarni et al., 2016) and doc2vec (Laxmi et al., 2021). Lastly, we computed a set of 37 content-based features derived from the comment's text, similar to the work done by Salminen (Salminen et al., 2018a). The list of features is depicted in Table 2. To summarize, for machine learning techniques, we used combinations of n-gram features, embedding features, and content-based features.

We tested a variety of classifiers, including logistic regression, random forest, decision tree, multinomial Naïve Bayes, and XGBoost. Moreover, we ex-

perimented with neural networks like convolutional neural networks (CNN) and bidirectional long short-term memory (biLSTM) networks. Regarding features, we used pre-trained word embeddings from Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). The BERT embeddings are from the BERT-medium model, consisting of 8 encoder layers, 8 attention heads, and 512 hidden units. We opted for this configuration because the BERT-base model exceeded the available memory limits, even after adjusting the training epochs' batch size.

Table 1
The total number of subreddits, comments, and conversations with potential toxicity triggers per year from the full collection.

Year	Subreddits	Comments	Conversations
2006	2	20,973	8830
2007	6	322,986	144,265
2008	40	2,197,734	925,058
2009	53	6,788,950	2,748,872
2010	68	16,937,287	6,691,905
2011	77	38,044,859	14,677,080
2012	88	72,820,851	27,022,958
2013	94	96,142,855	33,884,728
2014	98	109,420,731	38,905,554
2015	99	126,560,964	45,361,043
2016	100	144,345,866	52,471,874
2017	100	151,211,058	55,007,125
2018	100	161,674,533	58,804,474
2019	100	160,989,402	59,541,621
2020	100	39,091,028	14,556,665
Total	100	1,126,570,077	410,752,052

We fine-tuned a BERT-base (uncased) model with 12 attention heads, 12 encoder layers, 768 hidden units, and 110 million parameters to overcome the memory issue. Furthermore, we used a custom binary focal loss (Lin et al., 2017) function and set gamma to 2. Moreover, we used the Adam optimizer and set the training embeddings' maximum length to 512. As for training the model, we chose a learning rate of $3e^{-5}$, batch size of 6, and 6 epochs with early stopping based on the ROC-AUC (Receiver Operating Characteristics-Area Under The Curve) score. Finally, we evaluated the neural network models' performance and chose the best-performing model to predict the comments' toxicity in the top 100 subreddits.

3.6. Toxicity trigger detection

Outcomes of toxicity detection showed widespread toxicity in online communities, motivating us to investigate its root causes (i.e., 'triggers'). Therefore, toxicity detection always ties in with toxicity trigger detection, as the latter depends on the previous procedure. Regarding online conversations, we define toxicity triggers as 'non-toxic parent comments with toxic child comments.' Based on this definition, we reconstructed 410,752,052 conversations with 1,126,570,077 comments into conversation threads with parent and child comments. To study if a trigger occurs in a conversation, we ensured that the conversations had at least one parent and one child comment. Then, we linked the parent and child comments in the conversations to their predicted toxicity scores. Previously, we described toxicity triggers as toxicity initiators in conversation threads. The intuition behind focusing on non-toxic comments is that we can expect that toxic comments are likely to have toxic child comments since toxicity is contagious (Watanabe et al., 2018). Therefore, we identify the point in the conversation where the conversation turns toxic. Regarding the association between parent and child comments, we can say that a non-toxic comment is a τ-toxicity trigger when the number of its toxic children is higher than or equal to τ.

To explore the toxicity distribution in conversation threads, we used the toxicity predictions we obtained previously to compute the proportions of toxic comments at different levels in conversation threads, as shown in Fig. 3. The level corresponds to the comment's position within a conversation thread. For example, a comment at level one directly replies to the submission (i.e., post), while a level two comment replies to first-level comments. Fig. 3 shows the first level has the highest proportion of toxic comments and the most posts triggering toxic comments in our collection. Moreover, the outliers in each box plot show that the subreddits with the highest toxicity proportions in the entire collection are *r/4chan* and *r/sex*. However, as the toxicity decreases in subsequent levels, it increases slightly as the conversation evolves. Henceforth, in upcoming experiments, we focus on finding toxicity triggers at all conversation levels to identify patterns across the investigated subreddits.

Next, we focused on non-toxic parent comments in conversation threads with toxic child comments, using our definition of toxicity triggers. In our collection, 164,679,311 non-toxic parent comments could be toxicity triggers, so we used these comments to build a model to predict whether a non-toxic parent comment is a toxicity trigger. For the prediction experiment, we used a dual embedding biLSTM neural network model with GloVe (Pennington et al., 2014) and FastText (Mikolov et al., 2018) embeddings.

To build the training set, we used the top 100 subreddits to randomly sample 991,806 parent comments (80% for training, 10% for validating, and 10% for testing). Then, we examined the number of toxic children in every conversation to determine whether parent comments are τ-toxicity triggers (i.e., they fall under the class trigger). Moreover, our subsequent experiments guaranteed the training set built was balanced, with an equal number of comments from the class trigger and not-trigger. Lastly, we assessed the performance of the toxicity trigger detection model and used it to label another random sample of comments to examine the spatiotemporal characteristics of toxicity triggers.

3.6.1. Characteristics of toxicity triggers

To understand the characteristics of toxicity triggers, we used the full triggers dataset to examine the key terms that represent the classes trigger and nottrigger. Before extracting the representative terms, we cleaned the training dataset by removing English stop words, URLs, and special characters from the comments text. Then, we computed the smoothed log-odds ratio of the term frequencies (Monroe et al., 2008) by using the following equation:

$$LOR(t_i, C_a, C_b) = log\frac{y_{ai} + \alpha}{n_a + \alpha.|y| - y_{ai} - \alpha} - log\frac{y_{bi} + \alpha}{n_b + \alpha.|y| - y_{bi} - \alpha} \quad (1)$$

where LOR stands for smoothed Log-Odds-Ratio, t_i is the ith term in the collection, C_a is class a, C_b is class b. The odd terms are computed for each class with the following equation:

$$Odds(t_i, C_a) = \frac{\#(t_i \in C_a)}{\sum_{t \neq t_i} \#(t, \in C)} = \frac{y_{ai}}{n_a - y_{ai}} \quad (2)$$

y_ai refers to the count of term i that belongs to class a, n_a–y_{ai} is the count of all terms (beside term i) that belong to the class. The log-odd ratios were smoothed with an α of $1e^{-8}$. Furthermore, we visualized

Table 2
The list of 37 content-based features split into two categories based on the type of computation.

Feature types	List of features
Counts 20 features	Characters (text length), words, capitals, nouns, verbs, adjectives, stop words, punctuations, periods, quotes, unknown words, discourse connectives, politeness words, rudeness words, single tokens, repeated punctuations, unique words, profane words, modal words, non alpha-numeric characters
Ratios (a:b) 17 features	a = counts of words, capitals, stop words, unique words, punctuations, nouns, verbs, adjectives b = text length a = counts of capitals, characters (without spaces), stop words, unique words, punctuations, profane words, nouns, verbs, adjectives b = count of words

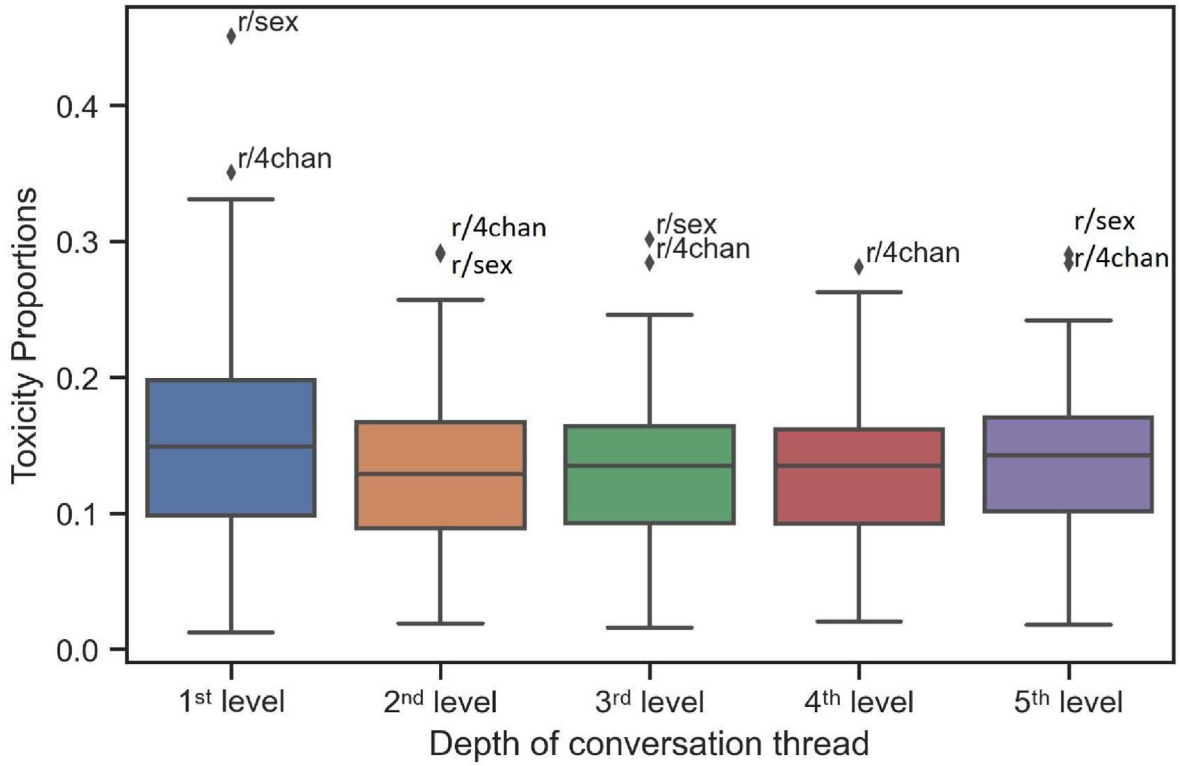


Fig. 3. The proportions of toxic comments at each level in conversation threads.

phrase associations between each category. We used PyTextRank (Nathan, 2016), which includes a Python implementation of the TextRank algorithm (Mihalcea & Tarau, 2004) that extracts the top-ranked phrases from data entries in each category. Lastly, we extracted unigram terms that characterize the class trigger and not-trigger from every subreddit. We used a ranking function that considers the frequency of terms in each subreddit relatively infrequently compared to general term frequencies in the full collection (Kessler, 2017). Then we compiled multiple lists of the top 10, 30, and 50 characteristic terms and found the most common terms across all subreddits.

3.6.2. *Trigger features*

The participant's state of mind highly influences the mood of conversations online. Therefore, changes in a conversation topic, context, or the general sentiment could be indicators that appear in comments that might potentially trigger toxicity (Topal et al., 2016). Motivated by this intuition, we present features that depend on shifts in the conversation topic, sentiment, and context. The features in the following subsections use information from conversation threads to find the conversation context along with shifts in topical and sentiment similarity.

3.6.2.1. Topical shift. To represent text, we used pre-trained word embeddings and extracted feature vectors from every comment. Then, we used the cosine similarity (Larson, 2010) distance measure to find the similarity between comments, as in Equation (3). This measure computes the unweighted semantic network (Kenter & de Rijke, 2015), which measures the similarity in topics between a potential toxicity trigger and all the predecessors from the conversation thread. For any given parent comment p and a prior comment c that came prior to p in a conversation, we compute the similarity between p and c as:

$$cos\left(\vec{v}_p, \vec{v}_c\right) = \frac{\vec{v}_p, \vec{v}_c}{\left|\vec{v}_p\right| \cdot \left|\vec{v}_c\right|} = \frac{\sum_{t=1}^{n} v_p(t) v_c(t)}{\sqrt{\sum_{t=1}^{n} v_p(t)^2} \sqrt{\sum_{t=1}^{n} v_c(t)^2}} \tag{3}$$

where v_p and v_c refer to the vector space representations of both p and c, while n refers to the total number of terms. To create the word embeddings of comments, we used the GloVe pre-trained model on five billion tokens to extract vectors of 100 dimensions per comment. Subsequently, we calculated the cosine similarity between the parent comment and all previous comments. Then, we used a threshold to decide if the comments were on-topic or off-topic. To identify this point, we used the k-means clustering algorithm (Hartigan & Wong, 1979) to create two clusters that characterize on-topic and off-topic comments. Then, we used the smallest clustering centroid to indicate comments that display topical shift (Topal et al., 2016). Thus, if the cosine similarity of a comment $\leq$ the threshold of the smallest centroid, then a topic shift ensued within the conversation.

3.6.2.2. Sentiment shift. To investigate the emotion of every comment in conversation threads, we used SenticNet 5 (Cambria et al., 2018) to find the sentiment of every comment based on the polarity of the text. We used the fifth installment of SenticNet, which consists of about 100,000 commonsense concepts obtained by deploying neural networks to infer and perform lexical substitutions. We used the Python API version of SenticNet and retrieved the polarity score of terms. To compute this feature, we mapped terms or keywords in the comment's text with phrases from SenticNet. Then, we combined the polarity of all the terms in the comment to a single score representing each comment's sentiment. Finally, we gauged the impact of changing sentiment on the toxicity of comments by using the same clustering technique we proposed in the topical shift analysis. First, we clustered the sentiment scores into two classes: one for comments that exhibit changes in sentiment, the other for comments with no changes in sentiment scores. Then, we used the lowest centroid as an indicator for comments that exhibit sentiment shift or not (Topal et al., 2016).

3.6.2.3. Conversation context. As users engage in conversations in online communities, the main topic of earlier entries in conversations typically governs the tone of the subsequent replies (Cunha et al., 2019). This

Data: A finite set $U_s = \{u_1, u_2, \ldots, u_n\}$ of utterances, where U_s has utterances from a subreddit s with labeled comments for toxicity
Result: Toxicity trigger label (L): trigger, not trigger
$L \leftarrow \emptyset$
$L_t \leftarrow \emptyset$
$F_t \leftarrow \emptyset$
$threshold \leftarrow 50\%$
for $i \leftarrow 1$ ***to*** n **do**
  while $u_i \neq \emptyset$ **do**
    $c_i \leftarrow$ read contents of $\{u_i\}$
    if *toxic children in* $c_i \geq threshold$ **then**
      training label L_t of parent p in $c_i \leftarrow$ toxicity trigger
    else
      training label L_t of parent p in $c_i \leftarrow$ *not* toxicity trigger
    end
    $F_t \leftarrow$ shift & context based training features from $\{c_i\}$
  end
end
PROVOKE model $\leftarrow$ build PROVOKE(*features*=F_t, *labels*=L_t)
$L \leftarrow$ inference labels with PROVOKE model on unseen instances
return L

Fig. 4. Pseudocode that shows how to identify toxicity triggers from a single utterance or a collection of utterances (conversations) in a subreddit.

conjecture inspired us to create a feature that uses the conversation context to find indicators of toxicity triggers. Before extracting the context of conversation threads, we extended our dataset by searching for all the submissions (i.e., posts) or comments that appear between the potential toxicity trigger and its immediate ancestor. This process yields a set of posts and comments that portray the conversation context. Next, we used GloVe pre-trained word embeddings to generate 100 feature vectors for every sentence in conversation contexts. Then, we aggregated the feature vectors by taking the average of the embeddings in every conversation thread to obtain a single context feature for every potential toxicity trigger.

3.7. Finding toxicity triggers with PROVOKE

To summarize our methodology, we propose PROVOKE: Prompting Responses Of Vocal Online Caustic Entities. This algorithm presents a systematic approach for identifying, detecting, and predicting toxicity triggers in online conversations. The pseudocode in Fig. 4 encapsulates the steps we followed to detect toxicity triggers from utterances (conversations).

4. Results

4.1. Toxicity detection

The first part of the toxic comment classification experiment addresses several problems with the dataset, such as the apparent class imbalance in the training set. To tackle this issue, we used a method that incorporates Synthetic Minority Over-sampling TEchnique (SMOTE) and Tomek Links (an undersampling technique) (Batista et al., 2004). This technique conducts over-sampling using SMOTE and cleaning with Tomek links. The Tomek method performs under-sampling by removing Tomek's links. In this case, a Tomek's link exists if the two samples are the nearest neighbors of each other. Then, we performed a min-max feature transformation to normalize each feature to a suitable range.

The following phase in the classification experiment involves feature selection, where we used Random Forest to classify, rank and choose the best features. The final step involves performing parameter tuning with grid search and running the classification with stratified five-fold cross-validation. The results of classifying the 10,000 labeled comments are in Table 3.

The results depicted in Table 3 show the classification performance of five different classical machine learning algorithms in terms of the F_1 score, ROC-AUC score, and accuracy. The results showed that the best performing classification model is XGBoost with unigram features, where it achieved an average F_1 score of 0.938, an accuracy of 92.2%, and an AUC-ROC score of 0.90.

Since several studies showed that neural networks (Georgakopoulos et al., 2018; Risch & Krestel, 2020, pp. 85–109) outperform classical machine learning methods in toxicity detection problems, we decided to evaluate the performance of several neural network models, as shown in Table 4.

First, we built a basic CNN model with one convolution layer, global max pooling, and batch normalization layers to normalize the layer dimensions and speed up the performance of the model (Zhou et al., 2016). Then, we set the learning rate to $2e^{-5}$, the optimizer to Adam, and the maximum sequence length for tokenizing the training set to 384. As for the training features, we used embedding features from a pre-trained BERT-medium model.[8] In these particular experiments, we used a BERT-medium because, unlike its larger counterpart, it did not cause any memory issues during training.

Second, we built a biLSTM model with one bidirectional LSTM layer and BERT-medium pre-trained embedding features. The model included average pooling layers with dense layers, and we used the same learning

[8] https://huggingface.co/google/bert_uncased_L-8_H-512_A-8; retrieved on Oct. 22.

Table 3
Classification performance of each feature category across five different classifiers using the 10,000 labeled comments.

Features	Logistic Regression			Random Forest			Decision Tree			MultinomialNB			XGBoost		
	F_1	*Auc*	*Acc.*(%)	F_1	*Auc*	*Acc.*(%)	F_1	*Auc*	*Acc.*(%)	F_1	*Auc*	*Acc.*(%)	F_1	*Auc*	*Acc.*(%)
Unigram	0.758	0.908	90.7	0.607	0.912	88.9	0.688	0.814	89.6	0.642	0.866	85.0	**0.768**	**0.938**	**92.2**
Bigram	0.359	0.649	66.7	0.206	0.625	82.8	0.205	0.570	81.9	0.367	0.654	71.4	0.358	0.660	84.5
N-gram (3–5)	0.306	0.545	50.2	0.299	0.551	42.5	0.076	0.518	81.9	0.253	0.552	67.7	0.216	0.552	76.9
TFIDF	0.737	0.921	89.9	0.699	0.902	90.0	0.682	0.815	88.8	0.658	0.893	85.4	0.733	0.923	90.9
Content	0.387	0.679	64.7	0.288	0.657	77.9	0.335	0.605	63.9	0.350	0.631	65.4	0.274	0.640	78.8
Word2vec	0.540	0.843	76.7	0.428	0.786	81.6	0.383	0.646	67.9	0.482	0.786	71.1	0.528	0.817	82.7
Doc2vec	0.571	0.846	82.3	0.551	0.832	84.7	0.430	0.719	73.0	0.500	0.803	72.6	0.529	0.821	83.3
All Features	0.664	0.892	87.5	0.574	0.908	87.8	0.657	0.820	87.3	0.919	0.718	88.4	0.700	0.936	90.5

Table 4
Performance of the toxicity detection neural network models in terms of macro F_1, ROC-AUC score, and accuracy.

Neural network models	Macro F1	ROC-AUC	Accuracy
CNN	0.847	0.823	91.5%
biLSTM	0.871	0.875	92.2%
CNN + LSTM	0.870	0.852	92.6%
LSTM + CNN	0.832	0.790	91.3%
fine-tuned BERT	**0.898**	**0.983**	**93.5%**

rate and sequence size as the previous CNN model. Third, we created a CNN + LSTM model with four channels, including convolution layers, global max pooling, and batch normalization. Every channel ended with an LSTM layer, while the final model had the combined channels with drop-out layers. As for the features and configurations, we used the exact BERT-medium pre-trained embeddings in the same configurations as the previous models. Fourth, we built an LSTM + CNN model that uses a bidirectional LSTM layer with convolution, global max pooling, and batch normalization layers. The model used the exact BERT-medium embeddings and configurations as prior models.

Lastly, we fine-tuned a transformer model and used BERT's uncased (i.e., lowercase) base model. Then, we split the 10,000 training data into 80% training, 10% testing, and 10% validation. As for fine-tuning the model, we used a custom loss function as we described in section 3.5 that accounted for class imbalance and used the class weight distributions during the training phase. The results from Table 4 showed that the average ROC-AUC was 0.983, the average model accuracy was 93.5%, and the average F_1 score was 0.898. The obtained scores from the fine-tuned BERT model outperformed the best-performing XGBoost classifier. So, we chose the fine-tuned BERT model to perform the prediction experiment. To evaluate the model's predictive accuracy, we computed a simple agreement score, and it yielded a result of 0.95, which further supports our choice of the neural network model to perform the prediction experiment.

We performed the prediction on all the comments from 2006 to 2020 across all the communities using our fine-tuned BERT prediction model. Next, we calculated the percentage of toxic and non-toxic comments from each subreddit, as in Table 5. This experiment shows that volume-wise, 2018 had the highest number of toxic comments, while 2012 had the highest percentage of toxic comments. Additionally, we showcased the total number of toxic comments and the toxicity percentage of each subreddit in Tables C11 and C.12. We ranked

subreddits in descending order based on the percentage of toxic comments. Our findings show that the subreddit *r/sex* (31.4%) was the most toxic in the entire collection, while the subreddit *r/4chan* (29.3%) was the second most toxic. Furthermore, the subreddit *r/askscience* (1.79%) showed the lowest percentage of toxicity across all subreddits. Our findings demonstrate that toxicity is pervasive in Reddit communities; thus, detecting toxicity is essential for such online communities, where subreddits exhibit varying toxicity amounts.

Table 5
Count and percentage of toxic and non-toxic comments in each year.

Year	Toxic	Toxic (100%)	Non-toxic	Non-toxic (100%)
2006	1049	5.002	19,924	94.998
2007	36,311	11.242	286,675	88.758
2008	334,076	15.201	1,863,658	84.799
2009	1,067,142	15.719	5,721,806	84.281
2010	2,719,148	16.054	14,218,133	83.946
2011	6,377,055	16.762	31,667,789	83.238
2012	12,343,586	16.951	60,477,231	83.049
2013	15,811,623	16.446	80,331,192	83.554
2014	17,533,156	16.024	91,887,534	83.976
2015	20,173,343	15.94	106,387,569	84.06
2016	22,059,698	15.313	121,995,751	84.687
2017	23,245,233	15.373	127,965,770	84.627
2018	24,639,531	15.32	136,197,684	84.68
2019	24,376,970	15.142	136,612,381	84.858
2020	5,774,622	14.772	33,316,392	85.228

4.2. Toxicity trigger detection

We address RQ1 by conducting experiments on the non-toxic parent comments we previously extracted from conversation threads to identify toxicity triggers. To determine the actual number of toxic children that follow a toxicity trigger, we tried changing the value of τ, which is the threshold that indicates the number of toxic child comments, beginning with a single toxic child ($\tau = 1$) and ending with more than 50% toxic children ($\tau \geq 50\%$). We tested the amount of τ by conducting smaller prediction trials for every threshold and manually assessed the prediction model's performance. Starting with $\tau = 1$, we randomly sampled 100,000 non-toxic parent comments and found that the model accuracy did not exceed 60%. Then, we gradually increased the value of τ and found that the model's performance degraded to 52% without any improvement.

Since conversation threads vary by length, we decided to make a custom τ that considers conversations with at least 50% toxic children as toxicity triggers.

Moreover, we included conversations with toxic parent comments in our nottrigger dataset because our original definition of a toxicity trigger is a nontoxic parent comment that causes conversations to be toxic. Thus, by default, anything that does not match this definition is not a trigger since we know that a toxic parent comment will probably lead to more toxicity in conversation threads (Almerekhi et al., 2019), so it is not a toxicity trigger. Our findings show that including more than 50% toxic children can detect toxicity triggers as long as the training set size is large. Thus, we built a dual embedding biLSTM model with FastText and GloVe embeddings as described in section 3.6. We randomly sampled 991,806 parent comments to train the model, where 496,648 were from the class trigger and the other 496,406 were from the class not-trigger. Moreover, we ensured that all non-toxic parents had at least 50% toxic children for the class trigger.

As for the class not-trigger, we ensured that 50% of the parent comments were toxic and the other 50% were non-toxic with no toxic children.

The outcomes of the classification experiment are in Table 6. As a baseline model, we experimented with detecting toxicity triggers with FastText and GloVe word embeddings only. Then, we introduced sentiment shift features to the model by concatenating the shift scores of parent comments with the embedding feature vector. The topical shift features did not show a significant improvement over the baseline. Similarly, we added topical shift features to the word-embedding feature vector. This time, the improvement over the baseline was slightly better than sentiment shift features. Then, we added the conversation context as a feature, providing another slight improvement over the topical shift features. Lastly, we combined topical shift, sentiment shift, and conversation context features with the embedding features. The model's achieved average accuracy was 78.81%. This result indicates that topical and sentiment shift features and the conversation context improve toxicity trigger detection. As for the predictive model's agreement score, we achieved 0.84–good enough for performing toxicity trigger detection.

Finally, we used the best-performing trigger detection model from Table 6 on a random sample of 990,900 parent comments (50% trigger and 50% nottrigger) from our 100 subreddits. The results showed that the model correctly predicted 339,688 (68.56%) comments as triggers with a precision of 0.86. As for the toxicity triggers across the top 100 subreddits, the prediction results in Tables D13 and D.14 indicate that the subreddits *r/4chan* and *r/sex* had the highest percentage of toxicity triggers, at 2.93 and 2.43, respectively. These results align with our previous experimental results showing that a high percentage of toxic comments appear at different conversation depths in *r/4chan* and *r/sex*.

4.3. Characteristics of toxicity triggers

With the scattertext visualization tool (Kessler, 2017), we answer RQ2 by computing the smoothed Log Odd Ratios (LOR) of trigger terms from a random sample of 14,000 comments split evenly between triggers and not-triggers. The term distribution depicted in Fig. 5 shows the sample's top trigger and not-trigger terms with their log frequency distribution along the y-axis.

Besides computing LOR, we ranked terms based on their frequency and showed the most frequent trigger and non-trigger terms in Fig. 6. The figure shows the top 20 terms from the class trigger and not-trigger. The characteristic terms refer to the most frequent terms in each class (i.e., unigram counts without applying the LOR equation). According to Fig. 6, terms from the class trigger include controversial or provocative keywords like *racist*, *dude*, and *gay* while the not-trigger terms contain fairly mild words like *recommend*, *app*, and *interesting*. Some of the top trigger terms in the figure come from the humor domain, where terms like *joke*, *funny*, and *lmao* are used the most for entertainment purposes. This outcome makes sense because popular Reddit communities like r/jokes and r/RoastMe use such terms to tell distasteful jokes or make fun of users in a spiteful manner (Ottoni et al., 2018).

As for the class not-trigger, many of the phrases shown in Fig. 6 come from the technology domain, such as the keywords *app* and *windows*. Overall, the figure shows that non-toxic trigger comments typically contain controversial terms that would most likely lead to subsequent child comments that are toxic. Moreover, some trigger terms had famous people, such as *Trump* and *Hillary*, which means that some named entities might trigger toxicity.

Table 6
Performance of the trigger detection neural network on different feature sets in terms of F1, ROC-AUC, and accuracy scores.

Features	Macro F1	ROC-AUC	Accuracy
FastText + GloVe (baseline)	0.7751	0.7783	77.79%
FastText + GloVe + sentiment	0.7754	0.7784	77.77%
FastText + GloVe + topic	0.7762	0.7791	77.84%
FastText + GloVe + context	0.7867	0.7889	78.64%
All features	0.7868	0.7891	78.81%

Additionally, we computed the term rank of comments from every subreddit in our collection. Then, we extracted the top k terms that appeared in each class (trigger and not-trigger), where k = 10, 30, and 50, respectively. The results depicted in Tables E15, E16, E17, and E18 show the top 10 words and characteristics (i.e., most frequent words) in each class ranked by the frequency of their appearance in every subreddit. Interestingly, in some subreddits like *r/worldnews*, the same keyword can be a trigger and not-trigger, like 'https.' Moreover, the keyword 'https' indicates that URLs with different types of content can trigger or not trigger toxicity. Moreover, we computed phrase associations between terms in our collection to understand which triggering or non-triggering terms coexist. We used scattertext's version of the TextRank (Mihalcea & Tarau, 2004) to find phrase associations between categories. The results from Fig. 7 show these phrases, which show some named entities like 'Bill Cosby,' are trigger phrases. Lastly, we tie in our observations from the prediction experiments

with some randomly sampled conversations where toxicity triggers might have appeared. The characteristics we investigate here are spatio-temporal because they occur at specific spaces (i.e., communities or subreddits) at certain times (i.e., year or exact timestamp). Firstly, we start with a conversation in Fig. 8, which shows a correctly classified toxicity trigger from the subreddit r/politics in 2016. The non-toxic parent comment in the figure includes some offensive jokes, which lead two of the three child comments (i.e., more than 50% of the child comments) to be toxic. Moreover, the conversation subject (i.e., the submission post) aims to evoke readers' responses by mentioning 'Trump'. Since the context of the parent comment was in agreement with the submission post, the child-comment responses were aggravated by the parent comment and were toxic. In other words, child comments were triggered by the parent's comment, which was not toxic.

As for incorrectly-classified toxicity triggers, the conversation in Fig. 9 shows an example of a toxicity trigger classified as not-trigger by the prediction model from the subreddit r/AskReddit in 2015. The preceding parent comment is a reply to a conversation that criticizes the post's topic. While the parent comment responds in a usual manner to the conversation, the subsequent child comments used crude language to respond to the parent comment, causing the model to predict the parent comment as a not-trigger, despite it being a toxicity trigger. Additionally, the user who posted the parent comment in Fig. 9 edited the comment and added more clarification and context to their stance on the conversation topic. Therefore, the child comments may have behaved unexpectedly to a prior version of the parent comment (like a version before clarification or edits (Yılmaz et al., 2021)), which can lead to incorrect classification.

On the other hand, one might argue that the parent comment does not trigger toxicity because we do not have more child comments that might prove that less than 50% of the child comments are toxic. In this case, the classification mistake was due to missing or deleted data from our collection. With this example, we demonstrate some of the challenges associated with the problem of detecting toxicity triggers in conversation threads, which mainly stem from a lack of conversational context (Zhang et al., 2018).

In other cases, the model might misclassify a parent comment that is not a toxicity trigger as a trigger, as in Fig. 10, where the conversation is from the subreddit r/Jokes, and it was in 2018. The context of the conversation is a lame joke about the age-old argument on movie adaptations of books and which version users typically prefer. In this case, the parent comment stated a controversial stance on the argument, but the subsequent child comment responded civilly.

In Fig. 10, the third child comment was a response from the user who wrote the potential toxicity trigger, and their response agreed with the preceding child comments. This behavior helped in preventing toxicity from starting in the conversation thread.

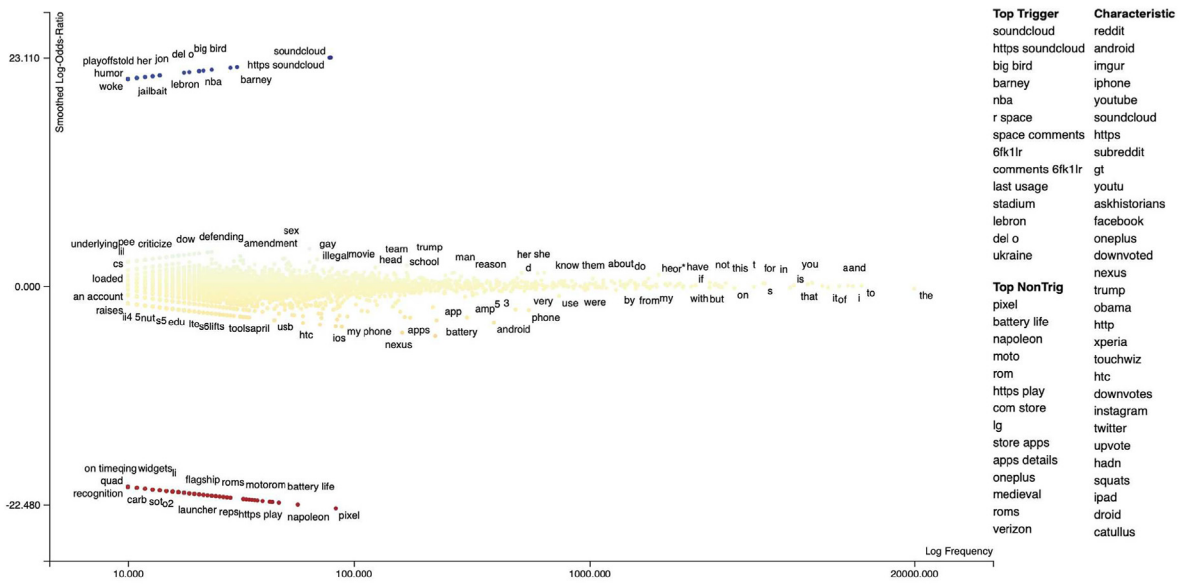


Fig. 5. The most common terms associated with triggers and non-triggers using smoothed LOR.

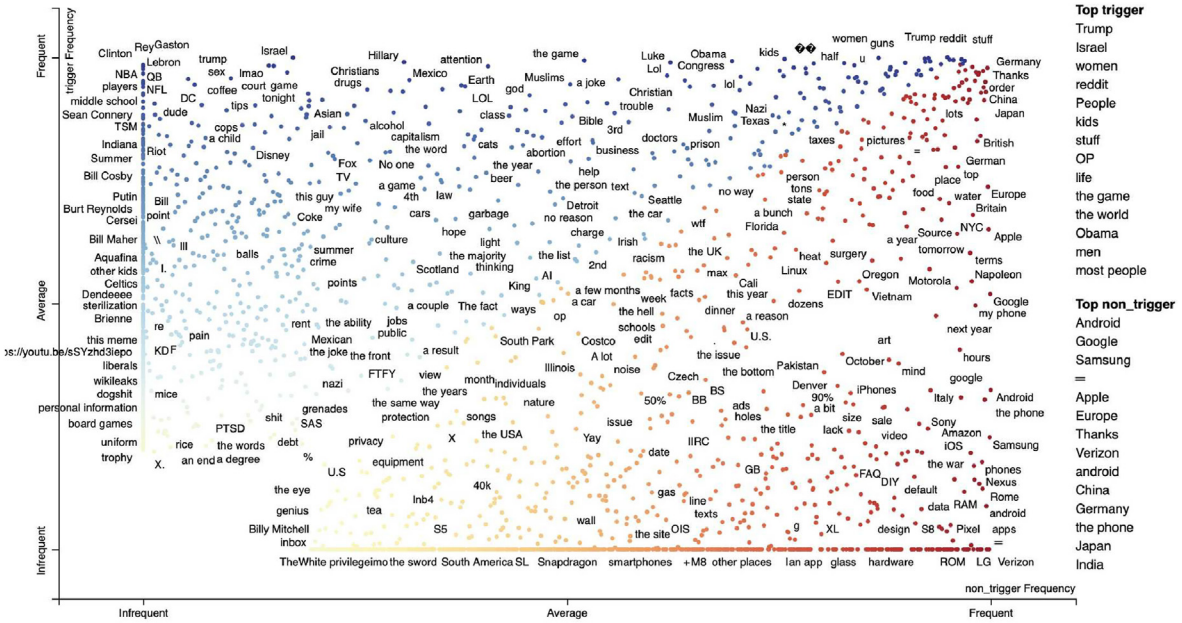


Fig. 6. The top phrases in each of the classes trigger and not-trigger, as computed by the smoothed log odds ratio.

5. Discussions and implications

This research predicts toxicity in online communities and builds conversation threads to predict toxicity triggers. We re-defined the toxicity trigger as a non-toxic parent comment within a conversation thread where at least half of the child comments are toxic. According to this definition, we built a dual embedding biLSTM neural network that combines topical shift, sentiment shift, and conversation context features to detect toxicity triggers. We found that adding trigger detection features increased the prediction model's performance by 1.02% over the baseline model. In terms of the distribution of toxicity and toxicity triggers across subreddits, we found that while the subreddit r/sex has the highest percentage of toxic comments (31.4%), the subreddit r/4chan has the highest amount of toxicity triggers (2.93%) in its conversation threads.

In terms of the evaluation of our prediction experiments, we found that when performing the toxicity detection experiment, despite the negligible performance improvement that neural networks offer, they are more efficient when performing prediction on a large volume of data. Notably, in our large-scale Reddit collection, the grid search and parameter tuning of XGBoost was less efficient when compared to the performance of the fine-tuned BERT model.

As for the toxicity trigger detection, the challenging part of our experiments was finding the appropriate number of toxic child comments (τ) that could help us create the training set to build the prediction model. Experimenting with exactly one toxic child comment yielded a decent prediction performance, with an average accuracy of about 61%. However, increasing the number of toxic child comments resulted in a degrading performance since there was not a large number of non-toxic parent comments with many toxic children. Thus, when the definition of toxicity triggers was restricted to include non-toxic parent comments with a specific number of toxic children, the resulting models performed

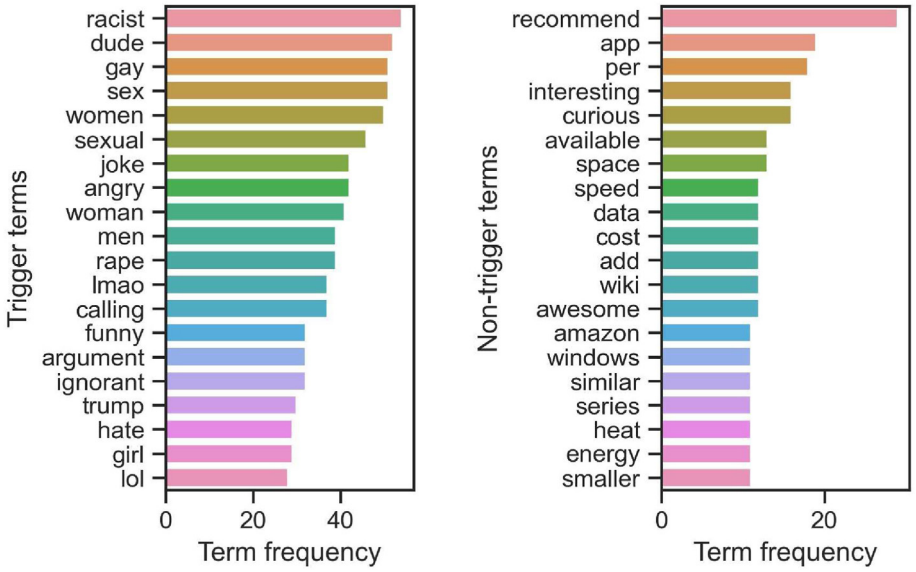


Fig. 7. The top terms in each of the classes trigger and not-trigger from the full collection, as ranked by term frequencies.

Post submission: Delta Bans Trump Supporter Who Ranted At 'Hillary Bitches' On Flight (VIDEO)

↳ **Parent comment:** I wonder if we can trade this guy (and all those like him) for a decent, hard working Mexican... – **Trigger**

↳ **Child comment$_1$:** Nah, Mexico doesn't want 'em. – **Toxic**

↳ **Child comment$_2$:** It's funny you say that. Growing up in Los Angeles I have a deep respect for the Mexicans' work ethic. If they're stealing jobs, they're stealing the ones no one else wants to do. The first time I saw Schrodinger's Immigrant I laughed my ass off. http://m.imgur.com/t3nvF7S?r I mean I never see white guys standing with the day laborers, but I do see them all over my town tweaking or drinking in the park or panhandling at the offramps or passed out on the sidewalk or pushing shopping carts and leaving a mess behind after going through our recycling bins. But never, ever, with the day laborers who are out there every morning. – **Non-toxic**

↳ **Child comment$_3$:** Honestly, 10/10 times I would much prefer a nice, fun and friendly Latino family move in across the street from me than the hate brimming brood of fuck-nuts that guys like this invariably have in tow... – **Toxic**

Fig. 8. A conversation from r/politics in 2016 where a trigger was correctly classified as a "trigger".

⋮

Preceding comment: There's like a dozen breakers in the box... do you shut off the main breakers and reset all the clocks or play "find the breaker" game? edit: speaking of which, breakers not using DIN rail, that's weird.

↳ **Parent comment:** What kind of electrician doesn't label the panel? – **Not-Trigger**

↳ **Child comment$_1$:** The garbage disposal is a special case when it's wired in. edit: plus it probably won't be hanging off GFCIs to avoid false-tripping them, unlike the rest of the kitchen. – **Non-toxic**

↳ **Child comment$_2$:** The breaker at my house is *fucked*. Oh, flip one of the breakers off? Well, that's the lights in the kitchen... except for the one above the sink, and it also shut off half the lights in the bathroom and one of the outlets in the basement. Fuck that thing. – **Toxic**

↳ **Child comment$_3$:** Every ass face in every place I've rented, I swear. – **Toxic**

Fig. 9. A conversation from r/AskReddit in 2015 where a trigger was incorrectly classified as a "not-trigger".

Post submission: Two mice are chewing on a film roll One says, "I liked the book better"

↳ **Parent comment:** The movie is usually better than the book – **Trigger**

↳ **Child comment$_1$:** Well the book probably tastes better than the movie – **Non-toxic**

↳ **Child comment$_2$:** Trust me, I've tried em – **Non-toxic**

↳ **Child comment$_2$:** Ok, I trust you – **Non-toxic**

Fig. 10. A conversation from r/Jokes in 2018 where a not-trigger was incorrectly classified as a "trigger".

poorly due to the small training set size. After several trials, we found that increasing the training set size and including all non-toxic parent comments with at least 50% toxic children improved the model's performance.

In the toxicity trigger training phase, we experimented with word-embedding (FastText + GloVe) features. We found that incorporating additional features such as topical shift, sentiment shift, and conversation context improves the model's performance. In particular, the combination of shift features and wordembedding features resulted in an average ROC-AUC score of 0.7891, which outperforms the baseline model (GloVe features model) by 1.02%.

Lastly, we examined the predicted toxicity triggers by looking at sample conversations that were correctly and incorrectly classified. The study's results showed that if the model correctly identifies triggers, it typically relies on precise trigger keywords that contain terms such as *argument* along with shift characteristics from the conversation thread. Nevertheless, when the model incorrectly classifies toxicity triggers, this is typically caused by missing data in the collection or a lack of information about the original editing status of the parent comment, which makes it challenging to identify if a parent comment triggers toxicity.

5.1. Implications

Many applications and domains could benefit from toxicity trigger detection. For instance, systems could use toxicity trigger detection to perform toxic topic discovery from various conversation threads. Additionally, toxicity trigger detection can shed light on user interactions in conversation settings. Toxicity trigger detection can aid content moderators with monitoring toxicity growth in conversations, thus protecting other users from harmful content and promoting healthy discourse in online spaces. In sum, PROVOKE can be used in various practical situations.

6. Conclusions and future work

By utilizing an extensive collection and numerous methods to predict toxic comments and toxicity triggers from conversation threads, this study contributes to the flow of research on online toxicity. Additionally, our technique demonstrates originality by being one of the earliest large-scale studies to examine online toxicity triggers and incorporates the conversation context with shift features to identify toxicity triggers based on their characteristics.

One of the limitations of this study is that it focuses on one platform only (i.e., Reddit). Although we used many communities, conversations, and comments over multiple years, future research can examine toxicity triggers on other platforms with conversation threads, such as Facebook and Twitter. Additionally, when we examined toxicity triggers, we did not consider URLs that appear in comments as indicators of toxicity triggers. In some subreddits like r/worldnews, keywords with 'https' trigger toxicity. Therefore, future research could examine the relationship between URLs and toxicity triggers in specific communities that enable the exchange of URLs.

Furthermore, misclassification examples showcased some challenges and limitations with detecting toxicity triggers and open areas for future work, such as incorporating additional features into the toxicity trigger detection model that detect trolling (Zhang et al., 2018) and introducing NLP-based shift features (Topal et al., 2016) to track the stylistic writing characteristics of the investigated comments.

The results from this research lead to many exciting directions. For instance, we show evidence that some subreddits are more vigorous against toxicity than others, leading to a vital question of isolating 'opposition tactics' that healthy subreddits utilize to curb the spread of toxicity. Moreover, using our largescale dataset, researchers can generate personas with some toxicity-trigger traits to simulate how they interact with social media stakeholders (Jansen et al., 2020). Moreover, researchers can conduct temporal studies on the evolution of toxicity in online communities and thus produce research that can positively impact the health of those communities.

Acknowledgements

This publication was funded by Qatar Research Leadership program grant, ID# QRLP9-G-3330102, from the Qatar National Research Fund (a member of Qatar Foundation).

Appendix A. Total comments

This appendix shows the 100 subreddits ranked by the total number of comments in a descending order.

Table A.7
The total number of comments in each subreddit from the full collection (1 of 2).

Rank	Subreddit	Comments	Rank	Subreddit	Comments
1	r/AskReddit	239,883,260	26	r/Overwatch	9,501,857
2	r/politics	78,228,929	27	r/aww	9,334,713
3	r/funny	44,709,237	28	r/europe	9,171,039
4	r/worldnews	42,503,595	29	r/relationships	9,019,790
5	r/pics	40,302,054	30	r/Android	8,129,269
6	r/nfl	40,028,933	31	r/dankmemes	7,920,818
7	r/leagueoflegends	39,461,036	32	r/mildlyinteresting	7,361,344
8	r/nba	38,581,328	33	r/Fitness	7,274,911
9	r/gaming	33,921,662	34	r/television	6,852,698
10	r/soccer	31,138,530	35	r/pokemon	6,835,332
11	r/news	30,914,270	36	r/explainlikeimfive	6,607,933
12	r/todayilearned	29,393,679	37	r/personalfinance	6,307,316
13	r/videos	26,826,394	38	r/BlackPeopleTwitter	6,154,037
14	r/WTF	21,458,545	39	r/gameofthrones	5,610,807
15	r/movies	20,748,338	40	r/Music	5,342,828
16	r/AdviceAnimals	20,175,324	41	r/StarWars	4,992,920
17	r/pcmasterrace	13,809,389	42	r/science	4,762,216
18	r/IAmA	13,701,108	43	r/TwoXChromosomes	4,720,541
19	r/atheism	12,381,164	44	r/programming	4,389,961
20	r/gifs	12,117,534	45	r/nottheonion	4,387,043
21	r/Games	11,862,689	46	r/sex	4,301,802

(*continued on next column*)

Table A.7 (*continued*)

Rank	Subreddit	Comments	Rank	Subreddit	Comments
22	r/Showerthoughts	10,999,516	47	r/tifu	4,187,780
23	r/technology	10,817,392	48	r/malefashionadvice	4,178,010
24	r/trees	9,863,810	49	r/Futurology	4,120,524
25	r/buildapc	9,643,472	50	r/interestingasfuck	4,054,820

Table A.8
The total number of comments in each subreddit from the full collection (2 of 2).

Rank	Subreddit	Comments	Rank	Subreddit	Comments
51	r/LifeProTips	3,798,226	76	r/CrappyDesign	1,592,596
52	r/fffffffuuuuuuuuuuu	3,759,719	77	r/WritingPrompts	1,581,073
53	r/books	3,616,906	78	r/gadgets	1,501,505
54	r/Jokes	3,507,508	79	r/woahdude	1,421,227
55	r/pokemongo	3,293,185	80	r/OutOfTheLoop	1,417,417
56	r/OldSchoolCool	3,078,475	81	r/RoastMe	1,351,058
57	r/me irl	2,873,991	82	r/history	1,288,481
58	r/4chan	2,843,609	83	r/EarthPorn	1,210,438
59	r/food	2,666,934	84	r/comics	1,205,484
60	r/sports	2,655,783	85	r/philosophy	1,198,194
61	r/cringepics	2,655,719	86	r/creepy	1,133,538
62	r/facepalm	2,581,024	87	r/nosleep	1,108,463
63	r/space	2,535,511	88	r/GetMotivated	1,028,421
64	r/dataisbeautiful	2,529,011	89	r/Art	1,021,702
65	r/bestof	2,199,292	90	r/rickandmorty	1,009,748
66	r/loseit	2,095,824	91	r/AskHistorians	890,251
67	r/Frugal	2,048,887	92	r/YouShouldKnow	751,565
68	r/askscience	1,981,221	93	r/HistoryPorn	721,604
69	r/oddlysatisfying	1,841,753	94	r/photoshopbattles	721,003
70	r/Documentaries	1,779,634	95	r/FoodPorn	666,759
71	r/UpliftingNews	1,694,871	96	r/tattoos	601,485
72	r/DIY	1,640,506	97	r/lifehacks	532,976
73	r/reactiongifs	1,630,167	98	r/AnimalsBeingJerks	434,943
74	r/Unexpected	1,603,833	99	r/listentothis	357,462
75	r/wholesomememes	1,598,188	100	r/InternetIsBeautiful	321,410

Appendix B. Total discussions

This appendix shows the 100 subreddits ranked by the total number of discussions in a descending order.

Table B.9
The total number of discussions with potential toxicity triggers in each subreddit from the full collection (1 of 2).

Rank	Subreddit	Discussions	Rank	Subreddit	Discussions
1	r/AskReddit	74,887,689	26	r/gifs	3,972,412
2	r/politics	30,389,467	27	r/Overwatch	3,834,139
3	r/worldnews	16,497,223	28	r/relationships	3,824,454
4	r/funny	15,540,023	29	r/dankmemes	3,319,322
5	r/leagueoflegends	15,329,153	30	r/Android	3,260,424
6	r/nfl	14,937,894	31	r/aww	3,255,958
7	r/nba	14,911,600	32	r/Fitness	2,889,184
8	r/pics	14,020,452	33	r/pokemon	2,859,970
9	r/gaming	12,648,001	34	r/personalfinance	2,671,970
10	r/soccer	12,281,753	35	r/mildlyinteresting	2,585,339
11	r/news	11,915,128	36	r/explainlikeimfive	2,420,104
12	r/todayilearned	10,552,071	37	r/television	2,400,697
13	r/videos	9,072,803	38	r/BlackPeopleTwitter	2,277,184
14	r/AdviceAnimals	7,407,213	39	r/gameofthrones	2,028,016
15	r/WTF	7,284,520	40	r/StarWars	2,018,912
16	r/movies	7,121,848	41	r/TwoXChromosomes	1,994,585
17	r/pcmasterrace	6,072,066	42	r/sex	1,870,730
18	r/atheism	5,413,665	43	r/programming	1,819,203
19	r/IAmA	5,218,293	44	r/Music	1,795,533
20	r/buildapc	4,788,643	45	r/science	1,762,016
21	r/Games	4,548,203	46	r/Futurology	1,610,783
22	r/trees	4,117,116	47	r/malefashionadvice	1,609,456
23	r/Showerthoughts	4,107,457	48	r/nottheonion	1,577,091
24	r/europe	4,038,498	49	r/fffffffuuuuuuuuuuu	1,523,746
25	r/technology	3,988,462	50	r/tifu	1,493,205

Table B.10
The total number of discussions with potential toxicity triggers in each subreddit from the full collection (2 of 2).

Rank	Subreddit	Discussions	Rank	Subreddit	Discussions
51	r/interestingasfuck	1,393,829	76	r/philosophy	592,794
52	r/Jokes	1,341,972	77	r/Unexpected	575,286
53	r/pokemongo	1,324,437	78	r/gadgets	548,944
54	r/LifeProTips	1,302,721	79	r/OutOfTheLoop	545,264
55	r/books	1,276,881	80	r/WritingPrompts	517,929
56	r/me irl	1,162,579	81	r/woahdude	503,453
57	r/4chan	1,118,082	82	r/history	480,678
58	r/OldSchoolCool	1,086,077	83	r/RoastMe	474,402
59	r/facepalm	1,020,147	84	r/nosleep	466,336
60	r/food	995,262	85	r/comics	463,744
61	r/cringepics	992,363	86	r/EarthPorn	445,641
62	r/sports	936,959	87	r/rickandmorty	402,218
63	r/space	920,118	88	r/creepy	399,439
64	r/dataisbeautiful	895,257	89	r/Art	395,201
65	r/bestof	869,350	90	r/GetMotivated	363,470
66	r/loseit	846,754	91	r/AskHistorians	362,435
67	r/Frugal	789,403	92	r/YouShouldKnow	290,363
68	r/askscience	769,033	93	r/HistoryPorn	274,951
69	r/Documentaries	722,196	94	r/FoodPorn	265,133
70	r/DIY	657,399	95	r/tattoos	255,775
71	r/UpliftingNews	650,457	96	r/photoshopbattles	220,877
72	r/oddlysatisfying	645,388	97	r/lifehacks	196,860
73	r/reactiongifs	618,790	98	r/AnimalsBeingJerks	158,025
74	r/CrappyDesign	608,218	99	r/listentothis	132,416
75	r/wholesomememes	599,030	100	r/InternetIsBeautiful	114,042

Appendix C. Total toxic comments

This appendix shows the 100 subreddits ranked by the percentage of toxicity in a descending order.

Table C.11
The total number and percentage of toxic comments in each subreddit from the full collection (1 of 2).

Rank	Subreddit	Toxic (%)	Rank	Subreddit	Toxic (%)
1	r/sex	1,350,871 (31.4)	26	r/Documentaries	302,945 (17.02)
2	r/4chan	833,291 (29.3)	27	r/atheism	2,093,820 (16.91)
3	r/BlackPeopleTwitter	1,541,190 (25.04)	28	r/trees	1,655,019 (16.78)
4	r/RoastMe	335,581 (24.84)	29	r/bestof	368,302 (16.75)
5	r/tifu	957,697 (22.87)	30	r/relationships	1,503,741 (16.67)
6	r/cringepics	575,974 (21.69)	31	r/worldnews	6,966,618 (16.39)
7	r/WTF	4,528,237 (21.1)	32	r/politics	12,530,629 (16.02)
8	r/AnimalsBeingJerks	89,201 (20.51)	33	r/todayilearned	4,665,415 (15.87)
9	r/AdviceAnimals	4,049,532 (20.07)	34	r/gameofthrones	884,271 (15.76)
10	r/fffffffuuuuuuuuuuuu	752,869 (20.02)	35	r/UpliftingNews	264,636 (15.61)
11	r/rickandmorty	202,195 (20.02)	36	r/sports	414,212 (15.6)
12	r/videos	5,106,504 (19.04)	37	r/OutOfTheLoop	219,550 (15.49)
13	r/news	5,832,274 (18.87)	38	r/Music	826,622 (15.47)
14	r/funny	8,294,412 (18.55)	39	r/OldSchoolCool	475,889 (15.46)
15	r/nottheonion	808,490 (18.43)	40	r/nfl	6,159,601 (15.39)
16	r/reactiongifs	299,838 (18.39)	41	r/Showerthoughts	1,684,647 (15.32)
17	r/facepalm	474,659 (18.39)	42	r/nba	5,818,276 (15.08)
18	r/Jokes	635,054 (18.11)	43	r/television	1,030,318 (15.04)
19	r/Unexpected	286,316 (17.85)	44	r/gaming	4,908,871 (14.47)
20	r/AskReddit	42,653,978 (17.78)	45	r/nosleep	160,061 (14.44)
21	r/dankmemes	1,386,859 (17.51)	46	r/comics	174,008 (14.43)
22	r/gifs	2,115,808 (17.46)	47	r/movies	2,931,681 (14.13)
23	r/TwoXChromosomes	809,564 (17.15)	48	r/CrappyDesign	222,291 (13.96)
24	r/creepy	194,260 (17.14)	49	r/GetMotivated	142,591 (13.87)
25	r/pics	6,880,864 (17.07)	50	r/aww	1,285,666 (13.77)

Table C.12
The total number and percentage of toxic comments in each subreddit from the full collection (2 of 2).

Rank	Subreddit	Toxic (%)	Rank	Subreddit	Toxic (%)
51	r/interestingasfuck	558,079 (13.76)	76	r/FoodPorn	65,547 (9.83)
52	r/IAmA	1,879,727 (13.72)	77	r/gadgets	140,356 (9.35)
53	r/soccer	4,169,837 (13.39)	78	r/Futurology	383,209 (9.3)
54	r/mildlyinteresting	981,886 (13.34)	79	r/Games	1,096,032 (9.24)
55	r/me_irl	379,906 (13.22)	80	r/pcmasterrace	1,234,401 (8.94)
56	r/leagueoflegends	5,159,155 (13.07)	81	r/food	238,004 (8.92)
57	r/LifeProTips	496,548 (13.07)	82	r/explainlikeimfive	578,529 (8.76)

(*continued on next column*)

Table C.12 (*continued*)

Rank	Subreddit	Toxic (%)	Rank	Subreddit	Toxic (%)
58	r/woahdude	181,883 (12.8)	83	r/pokemon	577,060 (8.44)
59	r/Overwatch	1,202,269 (12.65)	84	r/books	302,190 (8.35)
60	r/photoshopbattles	90,867 (12.6)	85	r/EarthPorn	99,612 (8.23)
61	r/technology	1,351,517 (12.49)	86	r/pokemongo	268,511 (8.15)
62	r/lifehacks	66,526 (12.48)	87	r/science	357,496 (7.51)
63	r/wholesomememes	198,558 (12.42)	88	r/malefashionadvice	295,595 (7.08)
64	r/YouShouldKnow	91,368 (12.16)	89	r/DIY	115,508 (7.04)
65	r/oddlysatisfying	214,087 (11.62)	90	r/programming	295,865 (6.74)
66	r/InternetIsBeautiful	36,973 (11.5)	91	r/history	85,551 (6.64)
67	r/europe	1,040,622 (11.35)	92	r/Android	535,505 (6.59)
68	r/listentothis	38,713 (10.83)	93	r/philosophy	77,502 (6.47)
69	r/tattoos	64,773 (10.77)	94	r/space	161,320 (6.36)
70	r/WritingPrompts	166,732 (10.55)	95	r/Frugal	119,024 (5.81)
71	r/Fitness	747,830 (10.28)	96	r/loseit	120,919 (5.77)
72	r/dataisbeautiful	259,432 (10.26)	97	r/personalfinance	257,816 (4.09)
73	r/Art	103,732 (10.15)	98	r/buildapc	295,477 (3.06)
74	r/HistoryPorn	71,850 (9.96)	99	r/AskHistorians	21,872 (2.46)
75	r/StarWars	496,064 (9.94)	100	r/askscience	35,510 (1.79)

Appendix D. Total toxicity triggers

This appendix shows the 100 subreddits ranked by the percentage of toxicity triggers in a descending order.

Table D.13
The total number and percentage of toxicity triggers in each subreddit from the full collection (1 of 2).

Rank	Subreddit	Triggers (%)	Rank	Subreddit	Triggers (%)
1	r/4chan	32,809 (2.93)	26	r/UpliftingNews	9166 (1.41)
2	r/sex	45,426 (2.43)	27	r/OldSchoolCool	15,289 (1.41)
3	r/RoastMe	10,946 (2.31)	28	r/atheism	76,205 (1.41)
4	r/BlackPeopleTwitter	51,867 (2.28)	29	r/AskReddit	1,047,316 (1.4)
5	r/cringepics	18,732 (1.89)	30	r/bestof	12,023 (1.38)
6	r/rickandmorty	7567 (1.88)	31	r/worldnews	226,914 (1.38)
7	r/tifu	27,238 (1.82)	32	r/todayilearned	143,978 (1.36)
8	r/WTF	129,498 (1.78)	33	r/TwoXChromosomes	27,057 (1.36)
9	r/AdviceAnimals	125,552 (1.69)	34	r/OutOfTheLoop	7298 (1.34)
10	r/AnimalsBeingJerks	2619 (1.66)	35	r/sports	12,419 (1.33)
11	r/fffffffuuuuuuuuuuu	25,241 (1.66)	36	r/Music	23,708 (1.32)
12	r/news	196,549 (1.65)	37	r/relationships	50,396 (1.32)
13	r/videos	148,631 (1.64)	38	r/Showerthoughts	53,133 (1.29)
14	r/Jokes	21,554 (1.61)	39	r/television	30,564 (1.27)
15	r/facepalm	16,383 (1.61)	40	r/nfl	188,432 (1.26)
16	r/reactiongifs	9855 (1.59)	41	r/nba	187,450 (1.26)
17	r/Documentaries	11,472 (1.59)	42	r/politics	381,996 (1.26)
18	r/funny	246,617 (1.59)	43	r/photoshopbattles	2753 (1.25)
19	r/dankmemes	51,943 (1.56)	44	r/gameofthrones	24,454 (1.21)
20	r/Unexpected	8976 (1.56)	45	r/comics	5536 (1.19)
21	r/creepy	6069 (1.52)	46	r/me irl	13,733 (1.18)
22	r/nottheonion	23,902 (1.52)	47	r/gaming	148,394 (1.17)
23	r/trees	62,155 (1.51)	48	r/movies	83,458 (1.17)
24	r/pics	208,611 (1.49)	49	r/interestingasfuck	16,290 (1.17)
25	r/gifs	58,536 (1.47)	50	r/mildlyinteresting	29,525 (1.14)

Table D.14
The total number and percentage of toxicity triggers in each subreddit from the full collection (2 of 2).

Rank	Subreddit	Triggers (%)	Rank	Subreddit	Triggers (%)
51	r/IAmA	58,996 (1.13)	76	r/explainlikeimfive	16,532 (0.68)
52	r/CrappyDesign	6810 (1.12)	77	r/Futurology	10,899 (0.68)
53	r/soccer	135,957 (1.11)	78	r/gadgets	3645 (0.66)
54	r/nosleep	5116 (1.1)	79	r/Games	28,607 (0.63)
55	r/aww	35,520 (1.09)	80	r/EarthPorn	2795 (0.63)
56	r/GetMotivated	3854 (1.06)	81	r/food	6135 (0.62)
57	r/leagueoflegends	160,451 (1.05)	82	r/pcmasterrace	36,828 (0.61)
58	r/LifeProTips	13,633 (1.05)	83	r/pokemon	16,574 (0.58)
59	r/woahdude	5216 (1.04)	84	r/books	7344 (0.58)
60	r/lifehacks	2008 (1.02)	85	r/pokemongo	7546 (0.57)
61	r/Overwatch	38,744 (1.01)	86	r/science	9818 (0.56)
62	r/YouShouldKnow	2902 (1.0)	87	r/WritingPrompts	2875 (0.56)
63	r/wholesomememes	5955 (0.99)	88	r/malefashionadvice	7952 (0.49)
64	r/technology	38,181 (0.96)	89	r/DIY	3161 (0.48)
65	r/oddlysatisfying	5945 (0.92)	90	r/programming	8277 (0.45)

(*continued on next column*)

Table D.14 (*continued*)

Rank	Subreddit	Triggers (%)	Rank	Subreddit	Triggers (%)
66	r/tattoos	2342 (0.92)	91	r/philosophy	2695 (0.45)
67	r/listentothis	1196 (0.9)	92	r/history	2071 (0.43)
68	r/europe	35,501 (0.88)	93	r/Frugal	3179 (0.4)
69	r/Art	3293 (0.83)	94	r/Android	12,645 (0.39)
70	r/InternetIsBeautiful	947 (0.83)	95	r/loseit	3185 (0.38)
71	r/dataisbeautiful	7210 (0.81)	96	r/space	3397 (0.37)
72	r/Fitness	21,911 (0.76)	97	r/personalfinance	5649 (0.21)
73	r/FoodPorn	1918 (0.72)	98	r/buildapc	7650 (0.16)
74	r/StarWars	14,447 (0.72)	99	r/AskHistorians	452 (0.12)
75	r/HistoryPorn	1898 (0.69)	100	r/askscience	690 (0.09)

Appendix E. Toxicity terms

This appendix shows the 10-most frequent words in the classes trigger and not-trigger from the top 100 subreddits.

Table E.15
The 10-most frequent words in toxicity triggers and non-triggers using term frequency. T:Trigger, NT:Not-trigger. (1–4).

Subreddit	Top 10 trigger and non-trigger terms computed using term frequency
r/4chan	T: faggot, gay, tell, women, kanjiklub, chan, www, com, funny, v
	NT: source, also, whole, possible, though, anyone, american, far, point, app
r/AdviceAnimals	T: http, women, guy, girl, someone, men, imgur, op, know, reddit
	NT: https, trump, government, pay, company, money, tax, companies, cost, democrats
r/Android	T: people, lol, r, literally, someone, argument, said, company, post, saying
	NT: app, screen, use, battery, apps, though, google, nexus, play, rom
r/AnimalsBeingJerks	T: bad, v, watch, trying, youtube, aggressive, maybe, yeah, animals, except
	NT: much, get, find, two, always, keep, could, sounds, stuff, looks
r/Art	T: people, trump, us, said, bad, say, wrong, hate, lol, everyone
	NT: love, work, thank, use, amazing, looks, painting, thanks, piece, paint
r/AskHistorians	T: man, sex, get, thing, women, know, like, sexual, r, reddit
	NT: use, military, long, however, also, th, states, western, far, several
r/AskReddit	T: people, like, never, said, women, guy, get, yeah, made, lol
	NT: http, com, www, reddit, need, org, comments, page, system, also
r/BlackPeopleTwitter	T: women, people, men, racist, girl, want, trying, lol, man, mean
	NT: http, bernie, pretty, title, com, work, school, imgur, subreddit, name
r/CrappyDesign	T: someone, r, said, post, joke, reddit, funny, people, gay, oh
	NT: probably, used, two, one, different, number, might, could, quite, looks
r/DIY	T: people, guy, op, someone, reddit, ever, say, nothing, man, even
	NT: would, use, wood, wall, thanks, using, looks, add, could, great
r/Documentaries	T: trump, white, women, lol, people, men, comment, guy, said, reddit
	NT: http, interesting, documentary, com, watch, documentaries, lot, much, amount, wikipedia
r/EarthPorn	T: people, r, reddit, nothing, everyone, think, joke, yeah, comment, better
	NT: shot, beautiful, hike, amazing, trail, camera, lake, taken, looks, lot
r/Fitness	T: people, guy, gym, someone, man, dude, never, lol, said, girl
	NT: week, calories, weight, muscle, program, protein, day, cut, days, routine
r/FoodPorn	T: people, food, even, person, reddit, sub, say, eat, burger, lol
	NT: recipe, looks, use, add, top, minutes, place, make, mix, delicious
r/Frugal	T: people, someone, guy, man, yeah, yes, comment, said, sorry, know
	NT: use, month, lot, year, card, cost, price, cheaper, bought, new
r/Futurology	T: people, trump, lol, man, want, literally, oh, nothing, said, bad
	NT: energy, could, cars, ai, would, power, solar, still, cost, technology
r/Games	T: people, literally, https, even, bad, saying, care, guy, wrong, comment
	NT: http, steam, pc, graphics, version, bit, new, games, available, rpg
r/GetMotivated	T: life, yeah, reddit, guy, world, everyone, said, man, s, oh
	NT: work, start, new, goals, feel, job, without, day, help, use
r/HistoryPorn	T: people, man, said, white, say, want, every, person, bad, dude
	NT: could, would, taken, org, picture, two, jpg, http, https, wiki
r/IAmA	T: http, guy, man, reddit, com, sex, guys, ever, girl, women
	NT: https, use, data, hi, new, also, many, different, work, help
r/InternetIsBeautiful	T: http, com, people, imgur, say, person, lol, jpg, others, leave
	NT: one, like, much, find, probably, also, add, site, great, similar
r/Jokes	T: racist, women, trump, said, funny, gay, people, feel, hate, black
	NT: ago, heard, number, r, many, english, repost, line, seen, would
r/LifeProTips	T: people, man, someone, guy, life, said, person, comment, tell, say
	NT: use, water, using, google, usually, also, add, app, works, lot
r/Music	T: people, someone, said, hate, saying, opinion, rap, kanye, reddit, lol
	NT: album, song, songs, great, new, spotify, thanks, check, love, sound
r/OldSchoolCool	T: people, reddit, racist, trump, comment, nazi, racism, nazis, women, say
	NT: looks, photo, dad, could, always, hair, picture, great, like, look

Table E.16
The 10-most frequent words in toxicity triggers and non-triggers using term frequency. T:Trigger, NT:Not-trigger. (2–4).

Subreddit	Top 10 trigger and non-trigger terms computed using term frequency
r/OutOfTheLoop	T: people, racist, women, woman, hate, person, white, bad, saying, lol
	NT: new, game, answer, google, specific, states, set, play, wikipedia, data
r/Overwatch	T: people, mercy, someone, genji, roadhog, saying, junkrat, hanzo, guy, person
	NT: beta, tank, new, role, dps, would, queue, sigma, ow, moira
r/RoastMe	T: try, better, lol, sorry, funny, attention, get, even, gay, people
	NT: bot, leaderboard, replying, ignore, look, ownagepwnage, reply, nices, gillysdaddy, u
r/Showerthoughts	T: women, https, r, men, want, know, lol, reddit, gay, someone
	NT: would, http, different, universe, could, infinite, light, number, space, possible
r/StarWars	T: people, tlj, movie, snoke, bad, kylo, character, literally, nothing, fans
	NT: clone, canon, sith, republic, new, rebels, episode, series, would, could
r/TwoXChromosomes	T: rape, sex, sexual, guy, men, women, guys, people, consent, man
	NT: doctor, help, love, years, hair, work, health, family, great, new
r/Unexpected	T: people, got, bad, nothing, anything, hate, animals, wrong, believe, person
	NT: expected, game, vredditdownloader, u, actually, bot, unexpected, use, link, water
r/UpliftingNews	T: people, trump, white, racist, argument, trying, right, man, black, nothing
	NT: years, cost, use, would, many, used, expensive, plastic, energy, companies
r/WTF	T: r, imgur, wtf, https, like, com, gif, op, jpg, guy
	NT: system, article, pay, org, however, government, money, use, wikipedia, using
r/WritingPrompts	T: com, exclamation, topic, icon, announcements, user, detail, responses, https, please
	NT: read, interesting, write, thank, prompt, short, many, hope, might, written
r/YouShouldKnow	T: people, everyone, internet, guy, care, real, feel, yeah, bad, life
	NT: also, something, much, data, app, year, used, though, using, phone
r/askscience	T: people, scientific, sex, science, saying, said, person, food, know, ask
	NT: light, energy, earth, mass, since, https, possible, gravity, space, speed
r/atheism	T: r, http, gay, people, post, reddit, hate, posts, atheism, subreddit
	NT: https, universe, exist, evidence, existence, gods, god, exists, question, belief
r/aww	T: people, https, person, reddit, comment, comments, man, r, nt, mean
	NT: breed, looks, name, mix, http, vet, cat, rescue, shelter, pup
r/bestof	T: women, trump, people, guy, racist, men, woman, oh, rape, said
	NT: system, new, pay, work, would, money, could, cost, tax, companies
r/books	T: people, bad, sex, sexual, saying, rape, someone, trying, women, scene
	NT: read, books, reading, book, recommend, love, really, series, looking, time
r/buildapc	T: people, even, https, said, lol, like, back, never, got, time
	NT: cpu, pcpartpicker, w, motherboard, item, supply, processor, breakdown, cooler, include
r/comics	T: women, people, said, men, someone, think, make, nothing, care, right
	NT: comics, comic, one, though, new, page, style, good, love, panel
r/creepy	T: trying, got, life, comment, com, guy, https, v, say, nothing
	NT: movie, since, far, many, photo, sure, could, stuff, times, remember
r/cringepics	T: guy, women, sex, gay, guys, woman, men, people, sexual, saying
	NT: use, name, english, looks, last, seen, number, though, pretty, language
r/dankmemes	T: questions, message, dankmemes, gay, compose, contact, https, let, r, moderators
	NT: jpeg, needs, fags, repostsleuthbot, image, e, balance, ayy, view, cuck
r/dataisbeautiful	T: people, said, trump, someone, nothing, guy, man, men, hate, go
	NT: data, graph, interesting, number, used, map, per, com, would, https
r/europe	T: https, people, trump, literally, nt, brexit, lol, hate, yeah, someone
	NT: http, eu, european, euro, countries, uk, money, italy, economy, english
r/explainlikeimfive	T: people, women, guy, man, said, someone, men, got, everyone, police
	NT: energy, light, speed, different, example, number, faster, water, air, heat
r/facepalm	T: trump, people, guy, hate, white, women, said, dude, really, woman
	NT: used, different, water, would, part, many, us, earth, correct, british

Table E.17
The 10-most frequent words in toxicity triggers and non-triggers using term frequency. T:Trigger, NT:Not-trigger. (3–4).

Subreddit	Top 10 trigger and non-trigger terms computed using term frequency
r/fffffffuuuuuuuuuuuu	T: people, women, girls, say, girl, person, want, woman, sex, someone
	NT: troll, rtroll, work, megusta, melvin, fy, wat, text, perfect, two
r/food	T: people, food, eat, r, http, know, sorry, saying, everyone, even
	NT: recipe, looks, salt, use, sauce, oil, pepper, pan, oven, sugar
r/funny	T: imgur, r, people, gif, women, jpg, guy, want, op, everyone
	NT: link, video, wikipedia, org, would, work, google, wiki, paper, use
r/gadgets	T: people, know, apple, comment, person, man, wrong, company, someone, everyone
	NT: battery, screen, price, android, much, two, gb, life, power, less
r/gameofthrones	T: cersei, people, arya, bad, literally, kill, killing, nk, mad, lol
	NT: books, book, http, spoiler, adwd, asos, wall, b, spoilers, com
r/gaming	T: people, com, r, imgur, reddit, http, post, said, guy, jpg
	NT: games, play, pc, ps, switch, game, played, fps, xbox, version
r/gifs	T: https, people, trump, get, dog, lol, man, everyone, literally, life
	NT: http, gif, com, imgur, movie, gifs, source, youtube, www, title
r/history	T: something, http, man, go, women, children, killed, saying, r, yeah
	NT: roman, empire, also, long, germany, new, rome, much, example, due
r/interestingasfuck	T: people, r, said, lol, man, comment, trump, dude, oh, got
	NT: water, http, long, used, light, air, would, looking, different, wikipedia

(*continued on next column*)

Table E.17 (*continued*)

Subreddit	Top 10 trigger and non-trigger terms computed using term frequency
r/leagueoflegends	T: people, na, tsm, nt, said, eu, toxic, reddit, worlds, even
	NT: damage, ap, ad, build, runes, jungle, champs, items, early, armor
r/lifehacks	T: people, r, funny, right, like, op, enough, say, toilet, kid
	NT: use, work, always, works, water, made, even, little, oil, new
r/listentothis	T: even, punk, try, going, saying, yeah, http, dude, everyone, read
	NT: song, love, album, great, spotify, thanks, songs, sound, well, like
r/loseit	T: people, someone, person, say, guy, look, lol, friend, tell, yeah
	NT: calories, day, week, eat, calorie, food, exercise, protein, eating, scale
r/malefashionadvice	T: people, fashion, someone, saying, lol, mfa, dude, thing, man, everyone
	NT: fit, shirt, color, casual, pair, jacket, looking, navy, black, shoes
r/me irl	T: skeltal, mr, thank, mention, haha, yeah, thanks, pupper, https, gon
	NT: good, bot, could, would, remindme, jpeg, link, needs, image, number
r/mildlyinteresting	T: people, https, reddit, comments, comment, lol, bot, r, person, care
	NT: http, wikipedia, imgur, org, used, wiki, looks, water, jpg, use
r/movies	T: people, https, literally, white, comment, lol, said, bad, guy, women
	NT: http, film, imdb, www, seen, com, films, title, tt, watch
r/nba	T: lol, https, warriors, people, lmao, fans, kd, literally, kawhi, curry
	NT: heat, wade, http, kobe, nash, howard, dwight, lin, bynum, bulls
r/news	T: trump, https, people, literally, white, gun, black, racist, women, cops
	NT: http, government, money, news, may, however, market, org, article, com
r/nfl	T: https, lol, people, fans, com, literally, got, mean, dude, na
	NT: tebow, packers, season, teams, bears, vick, jets, defense, play, see
r/nosleep	T: com, http, www, watch, guy, let, police, everyone, real, v
	NT: since, work, could, something, perhaps, many, place, great, find, hope
r/nottheonion	T: guy, rape, said, people, sexual, trump, man, racist, sex, women
	NT: new, less, states, data, government, pay, though, state, would, wiki
r/oddlysatisfying	T: people, r, lol, life, anything, said, read, guy, joke, post
	NT: looks, actually, like, paint, use, side, gifreversingbot, beautiful, though, seems
r/pcmasterrace	T: people, game, someone, saying, bad, every, said, watch, fact, reddit
	NT: cpu, gpu, ram, ryzen, motherboard, gb, monitor, pc, case, card
r/personalfinance	T: people, got, someone, op, never, said, ever, job, work, bad
	NT: tax, interest, income, fund, ira, roth, k, account, savings, taxes

Table E.18
The 10-most frequent words in toxicity triggers and non-triggers using term frequency. T:Trigger, NT:Not-trigger. (4 - 4).

Subreddit	Top 10 trigger and non-trigger terms computed using term frequency
r/philosophy	T: people, sex, sexual, get, nt, animals, bad, good, killing, wrong
	NT: consciousness, true, seems, universe, different, brain, theory, might, science, physical
r/photoshopbattles	T: nsfw, http, jpg, imgur, like, know, com, maybe, game, png
	NT: make, image, submission, background, removed, nice, gif, link, original, edit
r/pics	T: people, trump, man, hate, imgur, someone, person, post, racist, nothing
	NT: photo, virus, city, work, looks, pic, wikipedia, org, water, wiki
r/pokemon	T: people, com, hate, http, even, bad, stop, r, ever, joke
	NT: would, ds, trade, get, thanks, iv, type, gen, ivs, breeding
r/pokemongo	T: game, someone, people, http, saying, com, post, life, playing, fun
	NT: raid, raids, shiny, event, mon, pok, day, evolve, research, iv
r/politics	T: trump, https, bernie, hillary, literally, sanders, lol, white, donald, democrats
	NT: paul, http, health, government, mccain, bush, ron, insurance, obama, system
r/programming	T: people, someone, reddit, said, comment, guy, person, nothing, job, everyone
	NT: c, language, languages, code, data, type, use, compiler, using, memory
r/reactiongifs	T: trump, people, say, person, girl, said, talking, thing, hate, president
	NT: gif, imgur, http, movie, com, new, gifv, lot, use, good
r/relationships	T: sex, op, sexual, cheating, women, guy, men, told, yeah, porn
	NT: together, help, work, time, need, long, may, talk, luck, therapy
r/rickandmorty	T: subreddit, bot, doot, contact, moderators, performed, concerns, automatically, questions, action
	NT: time, episode, season, c, infinite, might, first, also, new, back
r/science	T: sex, women, man, people, wrong, guy, want, http, got, men
	NT: https, energy, could, study, solar, research, system, effects, low, field
r/sex	T: sex, orgasm, guy, porn, women, men, pleasure, girl, guys, like
	NT: doctor, pill, birth, iud, years, com, control, test, www, herpes
r/soccer	T: https, people, lol, fans, mate, literally, comment, lmao, said, comments
	NT: http, torres, play, mls, epl, think, spain, teams, game, see
r/space	T: people, reddit, everyone, guy, comment, r, com, feel, man, said
	NT: orbit, would, earth, launch, atmosphere, light, could, speed, mass, far
r/sports	T: people, said, com, nothing, lol, saying, man, comment, https, want
	NT: ball, teams, play, games, team, football, baseball, sport, lot, much
r/tattoos	T: people, bad, opinion, someone, say, even, worst, r, sub, lol
	NT: artist, love, style, design, awesome, would, amazing, color, going, piece
r/technology	T: people, trump, https, literally, government, said, someone, lol, right, person
	NT: http, use, windows, apple, using, os, software, iphone, microsoft, mac
r/television	T: people, trump, women, said, sexual, someone, lol, white, racist, saying
	NT: show, shows, season, episodes, episode, netflix, series, watch, tv, watching
r/tifu	

(*continued on next column*)

Table E.18 (*continued*)

Subreddit	Top 10 trigger and non-trigger terms computed using term frequency
	T: sex, girl, guy, man, happened, op, tifu, said, yeah, women
	NT: use, water, though, us, different, system, works, work, p, car
r/todayilearned	T: people, women, said, https, men, guy, bad, want, man, person
	NT: http, wikipedia, wiki, org, interesting, used, use, water, til, number
r/trees	T: people, guy, man, said, dude, someone, trees, saying, police, comment
	NT: use, thc, vape, water, glass, recommend, cbd, work, oil, using
r/videos	T: https, people, someone, women, person, said, white, literally, youtu, stop
	NT: http, www, com, movie, awesome, song, org, watch, video, v
r/wholesomememes	T: people, gay, guy, reddit, joke, said, saying, u, must, us
	NT: day, happy, thank, though, really, always, makes, friend, little, amazing
r/woahdude	T: r, http, people, com, reddit, www, guy, watch, youtube, v
	NT: would, two, though, much, could, side, make, speed, imagine, work
r/worldnews	T: trump, https, literally, people, nt, lol, russia, like, isis, president
	NT: israel, http, israeli, gaza, palestinian, org, palestinians, wikipedia, jews, jewish

References

Almerekhi, H., Jansen, S.b. B. J., & Kwak, c.-s. b. H. (2020a). Investigating toxicity across multiple reddit communities, users, and moderators. In *Companion proceedings of the web conference 2020, ACM, New York, NY, USA* (pp. 294–298).

Almerekhi, H., Kwak, H., Jansen, B. J., & Salminen, J. (2019). Detecting toxicity triggers in online discussions. In *Proceedings of the 30th ACM conference on hypertext and social media, HT '19* (pp. 291–292). New York, NY, USA: ACM.

Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2020b). Are these comments triggering? Predicting triggers of toxicity in online discussions. In *Proceedings of the web conference 2020* (pp. 3033–3040). New York, NY, USA: ACM.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on world wide web companion* (pp. 759–760). Switzerland: WWW '17 Companion, ACM, Republic and Canton of Geneva.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter, 6*(1), 20–29.

Bosque, L. P. D., & Garza, S. E. (2016). Prediction of aggressive comments in social media: An exploratory study. *IEEE Latin America Transactions, 14*(7), 3474–3480.

Cambria, E., Poria, S., Hazarika, D., & Kwok, K. (2018). SenticNet 5: Discovering Conceptual Primitives for Sentiment Analysis by Means of Context Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1), 1795–1802.

Carton, S., Mei, Q., & Resnick, P. (2020). Feature-based explanations don't help people detect misclassifications of online toxicity. *Proceedings of the International AAAI Conference on Web and Social Media, 14*(1), 95–106.

Choi, D., Han, J., Chung, T., Ahn, Y.-Y., Chun, B.-G., & Kwon, T. T. (2015). Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 ACM on conference on online social networks, COSN '15* (pp. 233–243). New York, NY, USA: ACM.

Chong, Y. Y., & Kwak, H. (2022). Understanding toxicity triggers on reddit in the context of Singapore. *Proceedings of the International AAAI Conference on Web and Social Media, 16*(1), 1383–1387.

Chu, S. K. W., Xie, R., & Wang, Y. (2021). Cross-language fake news detection. *Data and Information Management, 5*(1), 100–109.

Cunha, T., Jurgens, D., Tan, C., & Romero, D. (2019). Are all successful communities alike? Characterizing and predicting the success of online communities. In *The world wide web conference* (pp. 318–328). New York, NY, USA: WWW '19, ACM.

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on facebook. *Scientific Reports, 6*, Article 37825.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers), ACM, Minneapolis, Minnesota* (pp. 4171–4186).

Dubois, E., Yuan, X., Bennett, D., Khurana, P., Knight, T., Laforce, S., Turetsky, D., & Wild, D. (2022). Socially vulnerable populations adoption of technology to address lifestyle changes amid covid-19 in the us. *Data and Information Management*, Article 100001.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549.

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management, 58*(3), Article 102524.

Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., & Plagianakos, V. P. (2018). Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th hellenic conference on artificial intelligence, SETN '18, ACM, New York, NY, USA* (p. 35), 1–35:6.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100–108.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the instagram social network. In *Social informatics, lecture notes in computer science* (pp. 49–66). Cham: Springer.

Jain, E., Brown, S., Chen, J., Neaton, E., Baidas, M., Dong, Z., Gu, H., & Artan, N. S. (2018). Adversarial text generation for google's perspective api. In *2018 international conference on computational science and computational intelligence (CSCI)* (pp. 1136–1141).

Jansen, B. J., Booth, D. L., & Spink, A. (2009). Patterns of query reformulation during web searching. *Journal of the American Society for Information Science and Technology, 60*(7), 1358–1371.

Jansen, B. J., Salminen, J. O., & Jung, S.-G. (2020). Data-driven personas for enhanced user understanding: Combining empathy with rationality for better insights to analytics. *Data and Information Management, 4*(1), 1–17.

Kenter, T., & de Rijke, M. (2015). Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge Management* (pp. 1411–1420). New York, NY, USA: CIKM '15, ACM.

Kessler, J. (2017). Scattertext: A browser-based tool for visualizing how corpora differ. In *Proceedings of ACL 2017, system demonstrations* (pp. 85–90). Vancouver, Canada: Association for Computational Linguistics.

Kulkarni, V., Perozzi, B., & Skiena, S. (2016). Freshman or fresher? Quantifying the geographic variation of language in online social media. In *Proceeding of the international AAAI conference on web and social media (ICWSM 2016)* (pp. 615–618).

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference, WWW '18, international world wide web conferences steering committee* (pp. 933–943). Switzerland: Republic and Canton of Geneva.

Kwon, K. H., & Gruzd, A. (2017). Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to donald trump's youtube campaign videos. *Internet Research, 27*(4), 991–1010.

Lanius, C. (2019). Torment porn or feminist witch hunt: Apprehensions about the #metoo movement on/r/askreddit. *Journal of Communication Inquiry, 43*(4), 415–436.

Larson, R. R. (2010). Introduction to information retrieval. *Journal of the American Society for Information Science and Technology, 61*(4), 852–853.

Laxmi, S. T., Rismala, R., & Nurrahmi, H. (2021). Cyberbullying detection on Indonesian twitter using doc2vec and convolutional neural network. In *2021 9th international conference on information and communication technology (ICoICT)* (pp. 82–86).

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988).

Massanari, A. (2017). # gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society, 19*(3), 329–346.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404–411). Barcelona, Spain: Association for Computational Linguistics.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), European language resources association (ELRA), Miyazaki, Japan* (pp. 52–55).

Mittos, A., Zannettou, S., Blackburn, J., & De Cristofaro, E. (2020). And we will fight for our race!" a measurement study of genetic testing conversations on reddit and 4chan. *Proceedings of the International AAAI Conference on Web and Social Media, 14*(1), 452–463.

Mohan, S., Guha, A., Harris, M., Popowich, F., Schuster, A., & Priebe, C. (2017). The impact of toxic language on the health of reddit communities. In M. Mouhoub, & P. Langlais (Eds.), *Advances in artificial intelligence* (pp. 51–56). Cham: Springer International Publishing.

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis, 16*(4), 372–403.

Nathan, P. (2016). *PyTextRank, a Python implementation of TextRank for phrase extraction and summarization of text documents*. https://doi.org/10.5281/zenodo.4637885. URL https://github.com/DerwenAI/pytextrank.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on*

world wide web (pp. 145–153). Switzerland: WWW '16, ACM, Republic and Canton of Geneva.

Obadimu, A., Khaund, T., Mead, E., Marcoux, T., & Agarwal, N. (2021). Developing a socio-computational approach to examine toxicity propagation and regulation in covid-19 discourse on youtube. *Information Processing & Management, 58*(5), Article 102660.

Orton, R., Marcella, R., & Baxter, G. (2000). An observational study of the information seeking behaviour of members of parliament in the United Kingdom. *ASLIB Proceedings, 52*(6), 207–217.

Ottoni, R., Cunha, E., Magno, G., Bernardina, P., Meira, W., Jr., & Almeida, V. (2018). Analyzing right-wing youtube channels: Hate, violence and discrimination. In *Proceedings of the 10th ACM conference on web science* (pp. 323–332). New York, NY, USA: WebSci '18, ACM.

Oussalah, M., Faroughian, F., & Kostakos, P. (2018). On detecting online radicalization using natural language processing. In *International conference on intelligent data engineering and automated learning* (pp. 21–27). Springer.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Perspective. (2017). Using machine learning to reduce toxicity online. available at: https://www.perspectiveapi.com. (Accessed 30 November 2017).

Pronoza, E., Panicheva, P., Koltsova, O., & Rosso, P. (2021). Detecting ethnicitytargeted hate speech in Russian social media texts. *Information Processing & Management, 58*(6), Article 102674.

Reddy, M. C., & Jansen, B. J. (2008). A model for understanding collaborative information behavior in context: A study of two healthcare teams. *Information Processing & Management, 44*(1), 256–273.

Riedl, M. J., Whipple, K. N., & Wallace, R. (2021). Antecedents of support for social media content moderation and platform regulation: The role of presumed effects on self and others, information. *Communications Society*, 1–18, 0 (0).

Risch, J., & Krestel, R. (2020). *Deep learning-based approaches for sentiment analysis*. Singapore: Springer Singapore. Ch. Toxic Comment Detection in Online Discussions.

Salminen, J. O., Al-Merekhi, H. A., Dey, P., & Jansen, B. J. (2018b). Inter-rater agreement for social computing studies. In *2018 fifth international conference on social networks analysis, Management and security (SNAMS)* (pp. 80–87).

Salminen, J., Almerekhi, H., Milenkovic, M., Jung, S.-g., An, J., Kwak, H., & Jansen, B. J. (2018a). Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceeding of the international AAAI conference on web and social media* (pp. 330–339). ICWSM 2018).

Sood, S., Antin, J., & Churchill, E. (2012). Profanity use in online communities. In *Proceedings of the SIGCHI conference on human factors in computing systems, CHI '12, ACM, New York, NY, USA* (pp. 1481–1490).

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web* (pp. 613–624). Switzerland: Republic and Canton of Geneva. WWW '16, International World Wide Web Conferences Steering Committee.

Topal, K., Koyuturk, M., & Ozsoyoglu, G. (2016). Emotion -and area-driven topic shift analysis in social media discussions. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 510–518).

Vogels, E. (2020). The state of online harassment. available at: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/. (Accessed 13 January 2022).

Wagner, E. D. (1994). In support of a functional definition of interaction. *American Journal of Distance Education, 8*(2), 6–29.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media* (pp. 19–26). Stroudsburg, PA, USA: LSM '12, Association for Computational Linguistics.

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access, 6*, 13825–13835.

Weninger, T., Zhu, X. A., & Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the reddit community. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 579–583). New York, NY, USA: ASONAM '13, ACM.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web* (pp. 1391–1399). Switzerland: WWW '17, ACM, Republic and Canton of Geneva.

Yılmaz, G. S., Gasaway, F., Ur, B., & Mondal, M. (2021). Perceptions of retrospective edits, changes, and deletion on social media. *Proceedings of the International AAAI Conference on Web and Social Media, 15*(1), 841–852.

Zhang, J., Chang, J., Danescu-Niculescu-Mizil, C., Dixon, L., Hua, Y., Taraborelli, D., & Thain, N. (2018). Conversations gone awry: Detecting early signs of conversational failure. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ume 1*, 1350–1361. Long Papers).

R. Zhao, A. Zhou, K. Mao, Automatic detection of cyberbullying on social networks based on bullying features, in: Proceedings of the 17th international conference on distributed computing and networking, ICDCN '16, ACM, New York, NY, USA, 2016, pp. 43:1–43:6.

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., & Xu, B. (2016). Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. In *Proceedings of the 26th international conference on computational linguistics* (pp. 3485–3495). Osaka, Japan: ACL.