Investigating Toxicity changes of cross-community Redditors from 2 Billion Posts and Comments

Hind Almerekhi [Corresp., 1] , Haewoon Kwak [2] , Bernard J Jansen [3]

[1] Hamad Bin Khalifa University, Doha, Qatar

[2] Singapore Management University, Singapore, Singapore

[3] Qatar Computing Research Institute, HBKU, Doha, Qatar

Corresponding Author: Hind Almerekhi
Email address: h.almerekhi@gmail.com

This research investigates changes in online behavior of users who publish in multiple communities on Reddit by measuring their toxicity at two levels. With the aid of crowdsourcing, we built a labeled dataset of 10,083 Reddit comments, then used the dataset to train and fine-tune a Bidirectional Encoder Representations from Transformers (BERT) neural network model. The model predicted the toxicity levels of 87,376,912 posts from 577,835 users and 2,205,581,786 comments from 890,913 users on Reddit over 16 years, from 2005 to 2020. This study utilized the toxicity levels of user content to identify toxicity changes by the user within the same community, across multiple communities, and over time. As for the toxicity detection performance, the BERT model achieved a 91.27% classification accuracy and an Area Under the Receiver Operating Characteristic Curve (AUC) score of 0.963 and outperformed several baseline machine learning and neural network models. The user behavior toxicity analysis showed that 16.11% of users publish toxic posts, and 13.28% of users publish toxic comments. However, results showed that 30.68% of users publishing posts and 81.67% of users publishing comments exhibit changes in their toxicity across different communities, indicating that users adapt their behavior to the communities' norms. Furthermore, time series analysis with the Granger causality test of the volume of links and toxicity in user content showed that toxic comments are Granger caused by links in comments.

Investigating Toxicity Changes of Cross-Community Redditors from 2 Billion Posts and Comments

Hind Almerekhi[1]**, Haewoon Kwak**[2]**, and Bernard J. Jansen**[3]

[1]**Hamad Bin Khalifa University, Doha, Qatar**
[2]**Singapore Management University, Singapore, Singapore**
[3]**Qatar Computing Research Institute, HBKU, Doha, Qatar**

Corresponding author:
Hind Almerekhi[1]

Email address: hialmerekhi@hbku.edu.qa

ABSTRACT

This research investigates changes in online behavior of users who publish in multiple communities on Reddit by measuring their toxicity at two levels. With the aid of crowdsourcing, we built a labeled dataset of 10,083 Reddit comments, then used the dataset to train and fine-tune a Bidirectional Encoder Representations from Transformers (BERT) neural network model. The model predicted the toxicity levels of 87,376,912 posts from 577,835 users and 2,205,581,786 comments from 890,913 users on Reddit over 16 years, from 2005 to 2020. This study utilized the toxicity levels of user content to identify toxicity changes by the user within the same community, across multiple communities, and over time. As for the toxicity detection performance, the BERT model achieved a 91.27% classification accuracy and an Area Under the Receiver Operating Characteristic Curve (AUC) score of 0.963 and outperformed several baseline machine learning and neural network models. The user behavior toxicity analysis showed that 16.11% of users publish toxic posts, and 13.28% of users publish toxic comments. However, results showed that 30.68% of users publishing posts and 81.67% of users publishing comments exhibit changes in their toxicity across different communities, indicating that users adapt their behavior to the communities' norms. Furthermore, time series analysis with the Granger causality test of the volume of links and toxicity in user content showed that toxic comments are Granger caused by links in comments.

INTRODUCTION

Online social media platforms enable users to communicate with each other in various ways, like sharing and publishing different types of content (Mondal et al., 2017). Unfortunately, the rapid growth of online communication on social media platforms has caused an explosion of malicious content in the form of harassment, profanity, and cyberbullying (Hu et al., 2013). A survey by Pew Research Center (Vogels, 2020) showed that 41% out of 10,093 American adults were personally harassed online, and 25% experienced severe forms of harassment. Moreover, 55% of the survey participants considered online harassment a major problem. This concern was also shared by online moderators (Cheng et al., 2015), noting that posts and comments on many social media platforms can easily take a dark turn and become *toxic*. Therefore, there is a need for solutions that identify toxic content and limit its presence on online platforms.

One challenge with studying online toxicity is the multitude of forms it takes (Davidson et al., 2017). These forms include hate speech, which refers to offensive content that targets a specific trait in a group of people (Silva et al., 2016); harassment, which occurs when a user deliberately aggravates other users online (Cheng et al., 2015); and cyberbullying, which means targeting and intimidating victims through online communication (Bowler et al., 2015).

The previous classifications of toxicity forms show that toxic content often contains insults, threats, and offensive language, which, in turn, contaminate online platforms (Mondal et al., 2017) by preventing users from engaging in discussions or pushing them to leave (Newell et al., 2016). Thus, several online platforms

have implemented prevention mechanisms, such as blocklists (Jhaver et al., 2018), that block specific accounts from interacting with users. Other approaches to preventing toxicity include deploying human moderators and bots to remove toxic content (Chandrasekharan et al., 2018). These efforts, however, are not scalable enough to curtail the rapid growth of toxic content on online platforms (Davidson et al., 2017). There is also the psychological distress associated with exposing human moderators to firsthand accounts of toxicity (Rodriguez and Rojas-Galeano, 2018). These challenges call for developing effective automatic or semiautomatic solutions to detect toxicity from a large stream of content on online platforms.

Users of social media platforms have various reasons for spreading harmful content, like personal or social gain (Squicciarini et al., 2014). Studies show that publishing toxic content (i.e., toxic behavior) is contagious (Tsikerdekis and Zeadally, 2014; Rodriguez and Rojas-Galeano, 2018); the malicious behavior of users can influence non-malicious users and leads them to misbehave, which affects the overall well-being of online communities. As an example of toxic behavior (Alfonso and Morris, 2013), one Reddit user named *Violentacrez* created several communities on controversial topics such as gore, and his followers mimicked this behavior by creating communities with highly offensive content as well. This influence-based culture (Johnson, 2018) that users follow in online communities motivates studies like the current study to investigate the problem of toxic online posting and commenting behavior. Fueled by cases reported in (Alfonso and Morris, 2013), this study focuses on the toxic behavior of users on Reddit. In particular, this research investigates the toxic cross-community behavior of users, which refers to publishing toxic content in more than one community.

This study argues that the toxicity of users' content may *change* based on the environment (i.e., community) in which they participate. Therefore, the focus is to investigate changes in toxicity in two types of content that Reddit describes as follows:

- *Post*: is the top-level submission of a user that can be either a post, link, video, image, or poll.

- *Comment*: is the response of another user or the poster to a post or a comment.

This study uses an extensive collection of more than 2.293 billion published content, including 87 million posts and 2.205 billion comments, from more than 1.2 million unique users who published content in more than 107,000 unique subreddits from June 2005 to April 2020.

LITERATURE REVIEW

Studies of online toxicity typically tackle the problem of hate from three main perspectives: (1) toxic behavior characterization, (2) toxic behavior detection, and (3) toxic behavior in online communities.

Toxic Behavior Characterization

Investigating human behavior is essential for organizations that rely on users to drive business, understand the dynamics of online communities, and prevent hate (Mathew et al., 2020; Yin and Zubiaga, 2021). Negative behavior of humans in online spaces involves a lack of inhibition, including online aggressiveness that would not exist in similar situations offline (Lapidot-Lefler and Barak, 2012). Suler (2004) introduced the phrase "toxic disinhibition" and defined it as the inhibition loss of users who act violently online, which holds no benefits and leads users to violate conventional coexistence rules (Suler, 2004). A typical form of toxic disinhibition is flaming behavior, which involves using hostile expressions to refer to other users in online communication. Textual features of flaming behavior include harsh language, negative connotations, sexual harassment, and disrespectful expressions (Pelicon et al., 2021). The definition of toxic disinhibition, or toxic behavior, varies based on the users, the communities, and the types of interactions (Shores et al., 2014). For instance, toxic behavior can consist of cyberbullying and deviance between players in massively multiplayer online games (MMOGs) (Shores et al., 2014; Kordyaka et al., 2020) or incivility between social media platform users (Maity et al., 2018; Pronoza et al., 2021), among other scenarios. In this work, we define *toxic behavior* in online communities as disseminating (i.e., posting) toxic content with hateful, insulting, threatening, racist, bullying, and vulgar language (Mohan et al., 2017).

Toxic Behavior Detection

There are two known methods for detecting toxic behavior on online platforms. The first relies on social network analysis (SNA) (Wang and Lee, 2021). The study of Singh et al. (2020) exemplifies SNA

usage and content-based analysis to detect cyberbullying (a form of toxic behavior) on Twitter. The study investigates the Momo Challenge, a fake challenge spread on Facebook and other social media platforms to entice younger users to commit violent acts. Researchers collected incidents related to the challenge by tracking 5,615 users' network graphs and 7,384 tweets using the Momo Challenge hashtag and relevant keywords. Findings showed that a small number of users employed keywords related to the Momo Challenge to cause cyberbullying events, whereas the majority used the keywords to warn other users about the dangerous challenge. Techniques involving SNA are suitable for detecting toxic behavior patterns and targeted attacks like cyberbullying. However, these techniques must analyze the involved users' social profiles or relations to detect toxic behavior, which may not be available on platforms other than social media websites. Therefore, the second and most common method avoids this limitation by detecting toxic behavior in user-generated content (Djuric et al., 2015).

Analyzing user-generated content involves detecting toxicity; this is a heavily investigated problem (Davidson et al., 2017; Ashraf et al., 2021; Obadimu et al., 2021). To detect toxic content, some studies (Nobata et al., 2016) build machine learning models that combine various semantic and syntactic features. At the same time, other studies use deep multitask learning (MTL) neural networks with word2vec and pretrained GloVe embedding features (Kapil and Ekbal, 2020; Sazzed, 2021). As for open-source solutions, Google offers the Perspective API (Georgakopoulos et al., 2018; Mittos et al., 2020), which allows users to score comments based on their perceived toxicity (Carton et al., 2020). The API uses pretrained machine learning models on crowdsourced labels to identify toxicity and improve online conversations (Perspective, 2017).

By using the outcomes of previous studies (Wulczyn et al., 2017; Georgakopoulos et al., 2018), this work evaluates the performance of classical machine learning models (Davidson et al., 2017) and neural network models (Del Vigna et al., 2017) to detect toxicity at two levels from user content.

Toxic Behavior in Online Communities

Online platforms continuously strive to improve user engagement through various forms of interaction. Websites such as Facebook and Reddit offer users the freedom to create communities of their own volition to interact with similar groups of users (Johnson, 2018). Despite the great interest in promoting healthy interactions among users in online communities, platforms struggle with the toxic behavior of some unsolicited users (Shen and Rose, 2019). This problem was evident on Reddit (Almerekhi et al., 2020; Massanari, 2017), where (Chandrasekharan et al., 2017a) found that some of the topics discussed by communities were incredibly toxic, leading to the 2015 ban of two hateful communities due to users' fears that these groups would infect other communities. The study found that the ban successfully prevented hate groups from spreading their toxicity to other communities. Nevertheless, this ban broke one of Reddit's core self-moderation policies, which exasperated users who sought total freedom on Reddit.

In a similar vein, Mohan et al. (2017) investigated the impact of toxic behavior on the health of communities. The study defines health as user engagement relative to community size and measures toxicity with a commercial lexicon from Community Sift to filter toxic words and phrases. By analyzing 180 communities, the study found a high negative correlation between community health and toxicity. Additionally, the study showed that communities require stable toxicity levels to grow in size without declining health. Despite these findings, the study did not consider users when investigating toxicity and viewed communities through content, not content creators.

As for cases in which toxic behavior arises between communities on different platforms, Chandrasekharan et al. (2017b) proposed a solution that relies on building a Bag of Communities (BoC). The research identified the abusive behavior of users in nine communities from Reddit, 4chan, MetaFilter, and Voat. By computing cross-platform post similarity, the proposed model achieved 75% accuracy without any training data from the target platform. Moreover, the BoC model can achieve an accuracy of 91% after seeing 100,000 human-moderated posts, which outperforms other domain-specific approaches. However, the study focused on cross-platform abusive behavior through content analysis without accounting for the users who behaved in an abusive or toxic manner.

Research Questions

Given the literature review discussed earlier, this study aims to answer the following research questions:

RQ1: *How can the toxicity levels of users' content and users across different communities be detected?*

RQ2: *Does the toxicity of users' behavior change (a) across different communities or (b) within the same community?*

RQ3: *Does the toxicity of users change over time across different communities?*

METHODOLOGY

In our study, investigating toxic behavior on Reddit requires a rigorous process that starts with obtaining the corpus from Reddit to detect toxicity and ends with finding insights from users' behavior.

Data Collection Site

Reddit is an online community with over 2.8 million sub-communities[1] that cover various topics from news to entertainment, incorporating a mix of cultures (Massanari, 2017). Those sub-communities are commonly known as "subreddits", denoted with the prefix "r/". The main activities that registered users (often called Redditors) perform include a) creating subreddits, b) submitting posts (i.e., sharing content in the community), c) commenting on the posts of others, and d) replying to comments in discussion threads (Choi et al., 2015; Kumar et al., 2018).
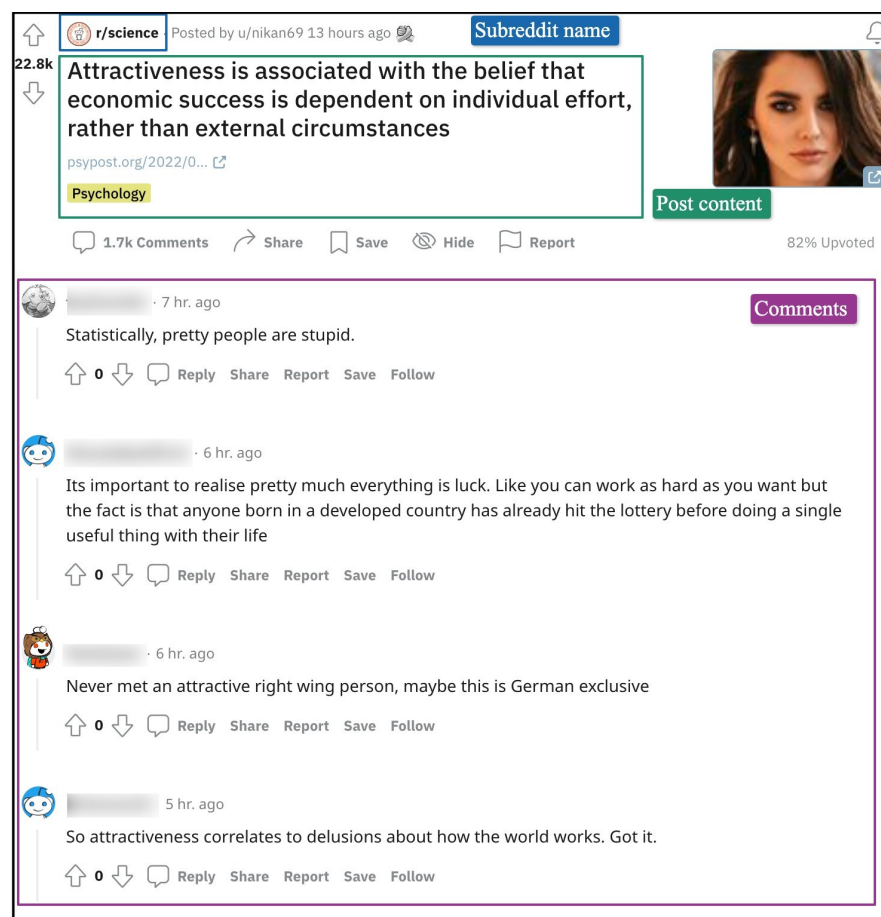
Figure 1. A Reddit post from the subreddit "r/science" with its associated discussion threads.

Figure 1 shows a post along with responses (i.e., comments) from "r/science." We can see in the figure that despite the strict moderation in this community, comments like "Statistically, pretty people are stupid" might be perceived by some Redditors as toxic.

---

[1] https://www.oberlo.com/blog/reddit-statistics; retrieved on 25 May 2022

Obtain Corpus

In this study, we targeted users in the top 100 subreddits (ranked by subscriber count). These subreddits account for a major proportion of Reddit's overall activity because they attract the largest number of active users (Choi et al., 2015). First, we compiled a list of subreddits using two popular external services: RedditList[2], and RedditMetrics[3]. We used both websites to curate a list of the top 100 largest safe-for-work subreddits based on subscriber count. In Appendix A, Tables A1 and A2 show the top 100 subreddits sorted by total subscribers. Note that while the subreddit r/announcements holds the highest number of subscribers, we removed it from our list because it did not serve our study purpose, as most of the users of the subreddit use it to consume content. Moreover, the subscriber values were retrieved on Aug. 29, 2017, while the total posts and comments are from June 2005 to April 2020. While there might be some differences among the top 100 subreddit rankings over the years, we used the list from 2017 because our study aims to understand the toxic behavior of a subgroup of users across multiple communities over time. Therefore, this specific list does not harm the research goal. Instead, the list brings unique opportunities for tracking the toxic behavior of a user subgroup over time.

Since this study focuses on users and their content, we cleaned our user collection by dropping any deleted or removed users (i.e., users with a removed or deleted "author" field). This process yielded 3,208,002 users who posted and 5,036,095 users who commented from 2005 through August 2017 within the top 100 subreddits. Additionally, we excluded bot users (i.e., automated accounts) to avoid potential biases in the subsequent analysis by using a publicly available list of 320 bots on Reddit[4]. Since the available bot list is outdated, it potentially misses newer bot accounts. Thus, for this work, we used the Pushshift API[5] to retrieve a list of accounts with a minimal comment reply delay. Setting the comment reply delay to 30 seconds allowed us to find more bot accounts that quickly reply to other users. We removed additional bot accounts by combining the bot list and Pushshift API list. When conducting this study, we found 37 bot accounts that produce around 2% of automated content. The massive volume of bot-generated content reaffirms the importance of removing bots in the data-cleaning phase of this study.

Since our research focuses on the toxic cross-community behavior of users, each user must participate in at least two different subreddits from the list of top 100 subreddits. Therefore, we filtered our original list of users to remove users who only participate in a single subreddit. This filtering process returned 577,835 users who posted (18% of the 3,208,002 users) and 890,913 users who left comments (17.7% of the 5,036,095). Furthermore, the intersection of these user lists yielded 241,138 users that performed both acts of posting and commenting. Overall, our dataset has 1,227,610 unique users who post and comment on Reddit. Lastly, we built our final collection of posts and comments with the Pushshift API (Baumgartner, 2017) by extracting user content from all subreddits. In other words, we started with a group of users participating in multiple communities, extracted their content from the top 100 subreddits, and then extracted their content from all other subreddits. As a result, we extracted 87,376,912 posts from 76,650 subreddits and 2,205,581,786 comments from 79,076 subreddits. To summarize, our collection has 2,292,958,698 posts and comments from 107,394 unique subreddits made by a group of cross-community users from June 2005 to April 2020.

Figure 2 shows the Cumulative Distribution Function (CDF) of the number of subreddits where a user left posts (A) and comments (B). Once a user participates in multiple subreddits (we already removed users who participated in a single subreddit, which is around 80% of users), the number of subreddits they participate in quickly grows. Findings from Figure 2 indicate that users who participate in less than 10 subreddits are more than 80% and 50% in terms of posting and commenting, respectively. Also, participation through commenting seemed to be easier than posting; in Figure 2, users who left comments in more than 20 subreddits are higher than 20%. In summary, the user collection in this study captured a substantial number of cross-community interactions and thus was appropriate for examining toxic behavior across multiple communities.

Label Dataset

To investigate the toxicity of users, we required a reliable machine learning model to detect the toxicity of user content. However, before building the detection model, we first created a set of relevance judgments (i.e., labels) that determine if a particular comment is toxic or not. Before conducting this study, we

---

[2] http://redditlist.com, retrieved on Aug. 29, 2017
[3] https://frontpagemetrics.com/top, retrieved on Aug. 29, 2017
[4] https://www.reddit.com/r/autowikibot/wiki/redditbots; retrieved on 13 May 2019.
[5] https://pushshift.io; retrieved on 22 May 2019.

(A)

(B)
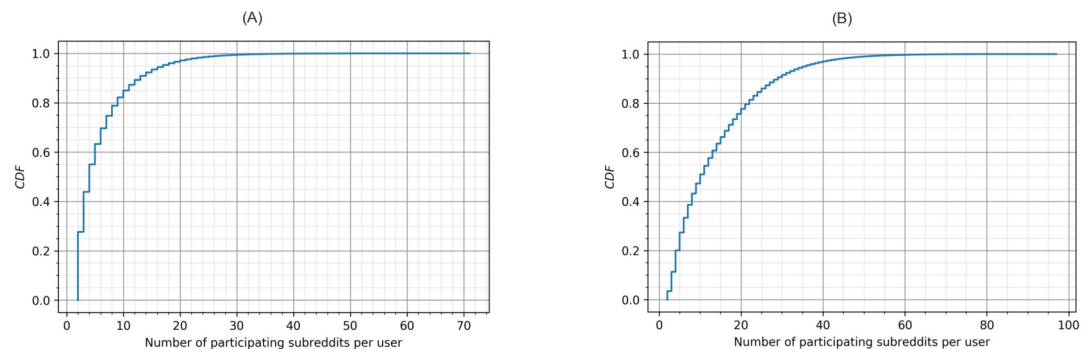
Figure 2. Cumulative Distribution Function of the participating subreddits per user in (A) posts and (B) comments.

found a few publicly available toxicity detection datasets, such as the Wikipedia comments training set that Google's Jigsaw/Conversation AI released on Kaggle [6]. However, we found that the dataset targets a classification problem that does not align well with our problem. Therefore, we created a labeled collection with data from our collection of comments.

To conduct our labeling experiment, we randomly sampled 10,100 comments from r/AskReddit, one of the largest subreddits in our collection. First, we used 100 comments to conduct a pilot study, after which we made minor modifications to the labeling task. Then, we proceeded with the remaining 10,000 comments to conduct the complete labeling task. We selected 10,000 comments to ensure that we had both a reasonably-sized labeled collection for prediction experiments and a manageable labeling job for crowdsourcing. For labeling, we recruited crowd workers from Appen[7] (formerly known as Figure Eight). Appen is a widely used crowdsourcing platform; it enables customers to control the quality of the obtained labels from labelers based on their past jobs. In addition to the various means of conducting controlled experiments, this quality control makes Appen a favorable choice compared to other crowdsourcing platforms.

We designed a labeling job by asking workers to label a given comment as either toxic or nontoxic according to the definition of a toxic comment in the Perspective API (Perspective, 2017). If a comment was toxic, we asked annotators to rate its toxicity on a scale of 2, as either (1) slightly toxic or (2) highly toxic. To avoid introducing any bias to the labeling task, we intentionally avoided defining what we consider highly toxic and slightly toxic and relied only on crowd workers' judgment on what the majority of annotators perceive as the correct label (Vaidya et al., 2020; Hanu et al., 2021). Nonetheless, we understand that toxicity is highly subjective, and different groups of workers might have varying opinions on what is considered highly or slightly toxic (Zhao et al., 2022). Therefore, annotators had to pass a test by answering eight test questions before labeling to ensure the quality of their work.

Additionally, we used 70 test questions for quality control during the labeling process. The test questions for the labeling task were comments that were undoubtedly toxic or nontoxic. Such questions were given to annotators at random intervals to detect spammers and ensure the quality of the obtained labels. One of the labeling task settings was to get the agreement of three annotators for each comment, meaning that at least two of the three annotators had to agree for a comment to be labeled. Moreover, crowd workers were required to spend a minimum of 30 seconds on each page while labeling and had a minimum labeling accuracy of 80% based on prior experience with Appen. Unfortunately, Appen does not provide any demographic statistics about the participating workers. Therefore, we only know that our workers had an accuracy $\geq 80\%$ and could understand English.

To assist workers with the labeling task, we provided a link to the discussion thread of the comment on Reddit to establish the context of the discussion. Regarding the job design, we split the job into 2,000 pages, each with five rows. As for costs, we paid USD 490.2 for the pilot (USD 9.48) and the complete

---

[6]https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge; retrieved on Feb. 13, 2018

[7]https://appen.com; retrieved on 10 June 2022.

labeling task (USD 471.24). The details of the labeling job are in Appendix B, where we show the labeling job instructions in Figure B1 and the labeling job test questions in Figure B2.

To assess the quality of the obtained labels from crowdsourcing, we computed a set of reliability measures[8] based on the labels from every worker for all classes. In this context, we define reliability as the proportion of perceived non-chance agreement to potential non-chance agreement (Gwet, 2014). To compute reliability, we used Equation 1, where $P_{\circ}$ is the percent of observed agreement and P_c is the percent of chance agreement. Generally, there are three main approaches to estimate the agreement by chance between two or more annotators across bifurcated categories: (a) individual-distribution-based approach like Cohen's Kappa (Zhao et al., 2013), (b) category-based approach such as Bennett et al.'s S score (Bennett et al., 1954), and (c) average-distribution-based approaches, such as the ones adopted by Gwet's gamma coefficient (Gwet, 2014).

$$Reliability = \frac{P_{\circ} - P_c}{1 - P_c} \tag{1}$$

Some of the reliability measures and results are in Table 1. Each measure takes into account various parameters based on its specific implementation. It is worth noting that while the observed agreement is high, Fleiss's Kappa is relatively low; this problem was heavily discussed by Feinstein et al. (Feinstein and Cicchetti, 1990) and Salminen et al. (Salminen et al., 2019), with the low Kappa score attributed 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. Given the class imbalance in our dataset, representative of the online discussion community, we use Gwet's measure in our research. Lastly, we sampled 100 random comments from our training set and manually labeled them for their toxicity. Then, we measured the agreement between our labels and crowd workers. Our findings showed that there is an agreement of 92%, which means that crowd labels are suitable for training machine learning models.

Table 1. Various reliability measure results on the relevance judgments obtained from Appen. ICC is Intraclass Correlation Coefficient

Measure	Obtained Value
Observed Agreement	0.848
Gwet's gamma coefficientcitep (Gwet, 2014)	0.697
Average rater consistency ICC	0.738
Average rater agreement ICC	0.737
Bennett et al. Score (Bennett et al., 1954)	0.697
Conger's kappa coefficient	0.483
Scott's pi, Krippendorff's alpha & Fleiss's kappa coefficient	0.482

Out of the 10,100 labeled comments, we removed three duplicate comments and 14 comments modified by moderators (i.e., the actual content of the comments was either removed or heavily edited by automatic bots or moderators). Of the 10,083 labeled comments, 86.81% were labeled as nontoxic, while the remaining 13.82% were labeled toxic (2.78% highly toxic and 11.04% slightly toxic). Some of the labeled comments from our collection are in Table 2. Due to space limitations, we only show relevant portions of the comments in the examples from Table 2.

Build Models

An integral part of our methodology involves building a robust prediction model that can classify posts and comments based on the crowdsourced labels into three classes, which are (a) nontoxic, (b) slightly toxic, and (c) highly toxic. The following subsections describe the features extracted from the labeled dataset and the classification models we evaluated to predict toxicity at different levels.

Extracting Features

For the prediction task, we utilized various features that characterize the semantic properties of text. First, we examined n-gram features at different configurations; then, we extracted an advanced set of features

---

[8] http://mreliability.jmgirard.com; retrieved on December 15, 2017.

Table 2. Examples of comments from every class as labeled by Appen workers.

Comment	Class
...Ex. Megan Fox is less hot to me because I have heard she is a fucking cunt. Fuck all of you Jersey haters. I harbor no hate for any states, except those in the south... Fuck that. I love seeing bitches lose their shit and go full retard on each other.	Highly toxic
Fuck I hate people that delete or not have a facebook account... She always does this hideous (like her face) fake laugh and says that I'm such a dumb blonde... ...Other people do well so they must be evil rather than hard working smart sonsabitches, huh?	Slightly toxic
I have pretty extreme ADHD and I do it to stay focused... I feel you on this one...with the exception of hard boiled egg in ramen... What I am saying is that the word "talent" is not that well defined...	Nontoxic

based on word embeddings, followed by a set of NLP-based features derived from the comments text. The following subsections explain each of the feature categories in more detail.

N-gram Features Before computing all the features, we cleaned the collection by removing new lines, URLs, and Reddit-specific [9] symbols like bullet points, quotes, strikethrough, spoilers, and coding parts from Reddit text. Furthermore, we split the collection text into tokens based on spaces and normalized all capital letters to lowercase letters. Then, we extracted 3,000 feature vectors from multiple variations of n-gram representations, including unigram features, bigram features, TF-IDF features, and n-gram features with a token range from 3 to 5 (Yin et al., 2009).

Word Embedding Features For embedding features, we created vectors in a low-dimensional space using a distributed memory model from the vector representations of the cleaned tokens in the collection (Le and Mikolov, 2014). So, we used Python's Gensim library [10] to build a skip-gram model and train the embeddings using a window size of 10, hierarchical softmax, and negative sampling of 10 noisy words. Then, we used the model to generate 300 word2vec feature vectors. Lastly, we trained another skip-gram model of window size 15, negative sampling of seven noisy words, and a learning rate of 0.03 to represent sentences as 300 doc2vec features.

NLP-based Features In addition to the previous features, we computed 37 shallow features based on natural language processing (NLP) techniques. Table 3 shows a summary of the list of features divided by the type of calculations we performed to obtain such features. We found that all the NLP-based features typically involve counting tokens or measuring the ratios of counts. Some of the features in Table 3 are adopted from Salminen et al. (2018), where they used similar features to identify hateful posts on Facebook and YouTube.

Table 3. The list of 37 NLP-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**	$\begin{cases} a = \text{counts of words, capitals, stop words, unique words, punctuations, nouns, verbs, adjectives} \\ b = \text{text length} \end{cases}$ $\begin{cases} a = \text{counts of capitals, characters (without spaces), stop words, unique words, punctuations,} \\ \text{profane words, nouns, verbs, adjectives} \\ b = \text{count of words} \end{cases}$

Classify Content

Classification Based on Classical Machine Learning

The classification approach considered several issues that persisted in the collection, such as the skewness of the classes. Since the labeled collection is highly skewed (86.81% of the comments are non-toxic), we

---

[9] https://github.com/LoLei/redditcleaner; retrieved on Oct. 16, 2021
[10] https://radimrehurek.com/gensim/; retrieved on May 20, 2018

311 had to address the class imbalance issue. One way to address this issue is to apply the Synthetic Minority
312 Over-sampling Technique (SMOTE) and Tomek links (an under-sampling technique) (Batista et al.,
313 2004). SMOTE performs over-sampling, and links are cleaned by under-sampling using Tomek links.
314 The next step in feature-based classification was to apply feature transformation by following a simple
315 min-max scaling approach to normalize all features. Then, we performed feature selection to reduce the
316 dimensionality of large feature vectors like n-grams. The last step in the classification procedure involved
317 performing a grid search for parameter tuning followed by repeated stratified cross-validation over five
318 folds.

319 *Classification Based on Neural Networks*
320 Recurrent neural networks (RNNs) are artificial neural networks designed to process a sequence of inputs
321 with temporal characteristics. One famous example is the long short-term memory (LSTM) neural network
322 consisting of an input gate, a cell, a forget gate, and an output gate. Another neural network type is the
323 convolution neural network (CNN or ConvNet), which is commonly used to analyze images. However, a
324 CNN can also be employed in other applications to detect data patterns (Johnson and Zhang, 2016).
325   In natural language processing, one of the most prominent deep learning networks is the trans-
326 former (Devlin et al., 2019), which handles sequences of ordered data to solve problems like machine
327 translation. The difference between RNNs and transformers is that the latter do not require data to be
328 processed in order. With this advantage, transformer models can be faster than RNNs during the training
329 phase due to the parallelization of the data processing phase (Vaswani et al., 2017). A well-known
330 transformer model is BERT, which consists of a multilayer bidirectional transformer encoder (Devlin
331 et al., 2019).

332 Explore Results
333 *Toxicity Judgments of User Content*
334 To determine user toxicity, we compute the percentages of highly toxic content and slightly toxic content.
335 Combining these provides a general judgment of a user's toxic behavior, regardless of the toxicity level.
336 Furthermore, combining toxicity levels compensates for the skewness of the dataset by increasing the
337 amount of data that represents what is considered toxic. For instance, if a user u creates three posts,
338 one labeled highly toxic, one slightly toxic, and one nontoxic, this user u is 67% toxic. Then, we use
339 percentiles to describe users based on the proportion of toxicity in their generated content. The quartile
340 values, which include the 25th, 50th, 75th, and 100th percentiles, capture the distribution of users based
341 on the toxicity proportions in their content. For instance, if the 25th percentile is 10, it means that 25% of
342 the time, the toxicity proportions in users' posts are below 10.

343 *Toxicity Changes of Users in Subreddits*
344 Prior studies of the behavior of online users found that temporal features can help characterize undesirable
345 behavior like trolling (Cheng et al., 2015). Thus, investigating the temporal aspects of toxic behavior is an
346 interesting problem from the perspective of Mathew et al. (2019), Kaakinen et al. (2018), and Cheng et al.
347 (2015). Toward this goal, we studied toxicity changes by computing the toxicity difference of user content
348 across all subreddits. In Equation 2, we illustrate how to calculate the change (Δ) of toxicity per user. To
349 get the percentage of toxic content, we combined the counts of highly toxic and slightly toxic content
350 that users made within a particular subreddit. Then, we divided this figure by the total number of content
351 items posted by these users within that same subreddit. Next, we computed the difference by subtracting
352 the highest toxicity proportion from each user's lowest toxicity within a particular year. Finally, from
353 Equation 2, we computed the differences in users' content from posts and comments over the years.

$$\Delta_u = \max_s(\frac{N(c_{toxic})_u^s}{N(c_{total})_u^s}) - \min_s(\frac{N(c_{toxic})_u^s}{N(c_{total})_u^s}) \tag{2}$$

354 $N(c_{toxic})_u^s$ is the number of toxic content (comments or posts) by user u in subreddit s, and $N(c_{total})_u^s$ is
355 total number of content (comments or posts) by user u in subreddit s.
356   For instance, if a subreddit s has a total of 20,000 posts, and a user u posted 1,000 slightly toxic and
357 800 highly toxic posts, we add both highly and slightly toxic posts to get a total of 1,800 toxic posts. Then,
358 we divide the total number of toxic posts by the total number of posts in s to get a toxicity percentage
359 of 0.09. With this procedure, we continue to get all the toxicity percentages of u to calculate Δ like

360 Equation 2. Obtaining the toxicity percentage of all users within subreddits in the posts and comments
361 collections is necessary for subsequent analysis in our study.

362 ***Link Analysis of User Content***

363 Another way to investigate user behavior is by looking at the links in their content. Since our collection is
364 rich in metadata, we extracted the URL field that includes the link (if any) that accompanies a post. Then,
365 we searched for all the URLs from comments text to build another version of our dataset that includes an
366 ID, a time stamp, and URLs. We included the time stamp in this version of the dataset to conduct some
367 statistical hypothesis tests, such as the Granger causality test (Mukherjee and Jansen, 2017) to identify
368 relationships between URLs in user content and the toxicity of said content.

FINDINGS

Classification Results

371 In the following subsections, we show the results of several experiments that build and evaluate machine
372 learning models for detecting toxicity at different levels.

Classical Classification Models

374 For this part of the analysis, we computed features to build and tune four widely used classic machine
375 learning models: Logistic Regression, Random Forest, Decision Tree, and XGBoost, using the 10,083
376 comments from the labeled collection. We chose these four algorithms for their extensive usage in
377 previous research on hate detection (Badjatiya et al., 2017; Salminen et al., 2018).

378 To handle class imbalance, we used SMOTE and Tomek Links on the training portion of the dataset
379 (0.80 of the labeled collection). Then, we transformed features by scaling all values to a range between 0
380 and 1. Additionally, we used the Random Forest algorithm to perform the classification, ranking, and
381 selection of features (Vens and Costa, 2011). The results depicted in Table 4 show the precision, recall,
382 F_1, AUC, and classification accuracy of every classifier, where accuracy measures the number of correctly
predicted data items divided by the total number of predicted items.

Table 4. The classification performance of each feature category across four different classifiers. P: precision, R: recall.

Features	Logistic Regression					Random Forest					Decision Tree					XGBoost				
	P	*R*	*F₁*	*AUC*	*ACC.*	*P*	*R*	*F₁*	*AUC*	*ACC.*	*P*	*R*	*F₁*	*AUC*	*ACC.*	*P*	*R*	*F₁*	*AUC*	*ACC.*
Unigram	0.649	**0.692**	0.668	0.878	85.1	0.653	0.443	0.479	0.882	84.1	0.618	0.606	0.611	0.763	85.2	0.715	0.653	**0.679**	0.913	**88.1**
Bigram	0.572	0.389	0.379	0.625	57.1	0.416	0.347	0.327	0.609	81.5	0.499	0.374	0.379	0.547	80.9	0.572	0.413	0.435	0.668	82.1
N-gram (3-5)	0.358	0.394	0.292	0.558	39.2	0.386	0.406	0.274	0.569	39.1	0.392	0.342	0.104	0.515	13.3	0.410	0.378	0.384	0.594	76.9
TFIDF	0.624	0.671	0.645	0.892	84.6	0.643	0.594	0.606	0.883	85.3	0.602	0.620	0.609	0.782	84.7	0.673	0.636	0.653	0.910	87.3
NLP	0.393	0.446	0.365	0.660	54.5	0.419	0.399	0.406	0.644	75.3	0.370	0.399	0.231	0.568	26.4	0.424	0.377	0.383	0.620	78.4
Word2vec	0.477	0.579	0.492	0.789	67.4	0.578	0.456	0.487	0.799	81.1	0.393	0.441	0.384	0.645	57.8	0.573	0.522	0.543	0.810	81.3
Doc2vec	0.498	0.593	0.523	0.828	73.8	0.570	0.531	0.548	0.822	81.9	0.426	0.474	0.434	0.652	66.4	0.561	0.551	0.556	0.815	81.3
All Features	0.610	0.641	0.624	0.893	84.1	0.662	0.509	0.552	0.884	84.9	0.592	0.578	0.584	0.735	84.6	**0.732**	0.636	0.671	**0.924**	87.8

383

384 Findings from Table 4 show the best performing classification model is XGBoost, followed closely by
385 Logistic Regression. As for the best features, the results show that models perform best on the unigram
386 features. However, all the features combined through concatenation with XGBoost showed the highest
387 precision and AUC scores at 0.732 and 0.924, respectively. On the other hand, with Logistic Regression,
388 unigram features achieved a recall score of 0.692. As for the F_1 and accuracy, XGBoost achieved the
389 highest scores of 0.679 and 88.1% on the unigram features. The grid search of XGBoost showed that the
390 best learning rate is 0.3, and the best number of estimators is 300. Moreover, feature selection reduced
391 the dimensionality of all the combined features from 12,637 to 1,406, where the top selected features
392 belonged to unigram and word embedding feature categories. This outcome aligns with the prior work
393 done by (Nobata et al., 2016), where their best-performing feature categories on all their datasets were the
394 n-gram and distributional semantic features.

Neural Network Models

396 Despite the outstanding performance of classic machine learning models, studies found that some neural
397 network architectures can outperform classical machine learning models, especially with capturing long-
398 range dependencies from textual data in hate detection problems (Badjatiya et al., 2017). Therefore, we
399 chose to experiment with varying configurations of BERT as a basis for our trials with CNNs, RNNs
400 (LSTM and biLSTM), and transformer networks.

- **CNN:** This model used a convolution layer along with global max pooling and batch normalization layers to normalize the layer dimensions and speed up the performance of the model (Zhou et al., 2016). The network deployed a learning rate of 0.00002. The optimizer was adam, and the maximum sequence length for tokenizing the training set was 384. The embedding features were from a pretrained BERT-medium model [11].

- **bidirectional LSTM:** The model used a bidirectional LSTM layer with embedding features from BERT-medium. Additionally, the model had average pooling layers with dense layers and was trained using the same learning rate and sequence size as the previous CNN model.

- **CNN+LSTM:** This model consisted of four channels with convolution layers, global max pooling, and batch normalization. In addition, the end of each channel has an LSTM layer. The final model consists of the combined channels with added drop-out layers. The same BERT-medium features were used in this model with the same configurations.

- **LSTM+CNN:** This model used a bidirectional LSTM layer followed by a series convolution, global max pooling, and batch normalization layers. Like the previous models, BERT-medium was used to obtain the feature vectors.

- **fine-tuned BERT:** This transformer model used the uncased (i.e., lower case) base model of BERT, which consists of 12 layers (also called transformer blocks), 768 hidden layers, 12 attention heads, and 110 million parameters. To fine-tune the model, we used a custom focal loss (Lin et al., 2017) function with gamma=2 and alpha=7 to account for class imbalance. Additionally, we computed class weights from each class's distribution of data points and used them to improve the training performance. As for the learning rate, we set it to 0.00003 and used the Adam optimizer.

We evaluated the performance of the neural networks on the labeled training set, where 80% of the data was for training while the rest was for testing and validation purposes. Then, we used the dataset's testing portion to evaluate the models' performance by measuring macro precision, recall, F_1, AUC, and accuracy scores, as in Table 5. The findings show that, clearly, the fine-tuned BERT model outperforms all neural network models, as it achieves a precision score of 0.7930, recall of 0.8034, an F_1 score of 0.7952, AUC score of 0.9629, and accuracy score of 91.27%.

One drawback of using a BERT model is that training was relatively slow. However, this issue can be solved by adjusting the batch size configuration. Comparing the performance of neural network models in Table 5 with the classical classification models in Table 4, the results showed that the fine-tuned BERT model outperformed all of the other models. Even though the performance of the BERT model might look too good to believe, the neural-network-based model has shown its high performance in toxic-classification tasks. For example, Google Jigsaw hosted a toxicity detection challenge to classify toxic comments from Wikipedia in 2018, and the first ranked team reported an AUC score of 0.9890 [12].

In summary, our fine-tuned BERT model will be used to detect toxicity in the remainder of this study.

Table 5. Performance of the neural network models in terms of the macro precision, recall, F_1, AUC, and accuracy scores.

	Evaluation Metrics				
Neural network models	*Precision*	*Recall*	*Macro F1*	*AUC*	*Accuracy*
CNN	0.6600	0.6172	0.6092	0.9231	87.12%
BiLSTM	0.6222	0.6629	0.6216	0.9336	86.42%
CNN+LSTM	0.6645	0.6966	0.6724	0.9380	87.51%
LSTM+CNN	0.7431	0.7212	0.7261	0.9570	89.99%
fine-tuned BERT	**0.7930**	**0.8034**	**0.7952**	**0.9629**	**91.27%**

---

[11]`https://huggingface.co/google/bert_uncased_L-8_H-512_A-8`; retrieved on Oct. 20, 2021
[12]`https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/52557`; retrieved on March 25, 2022.

436 ***Transferability of Models across Subreddits***

437 To use the models to predict the toxicity of all the Reddit comments, we must ensure that the model
438 trained by the data from r/AskReddit is transferrable to data from other subreddits (Fortuna et al., 2021).
439 Toward this goal, we obtained a random sample of 1,000 comments from the remaining 99 subreddits
440 (besides r/AskReddit). Then, we used crowdsourcing to label the comments for their level of toxicity.
441 Finally, we used the same labeling job we described earlier to obtain the ground truth labels. Comparing
442 the results from the prediction model and the crowdsourcing workers showed that the agreement between
443 them was 94.2%, meaning that the model trained on r/AskReddit can be generalizable to other subreddits.

444 **Detecting and Determining the Toxicity of Content and Users**

445 To answer RQ1 (*How can the toxicity levels of users' content and users across different communities*
446 *be detected?*), we infer the toxicity of the entire posts and comments collection by using our fine-tuned
447 BERT model. Post toxicity was detected by concatenating post titles and body sections (if they existed).
448 As for comments, toxicity was predicted directly from the comment text. The results of running the
449 prediction model on the entire collection of 87,376,912 posts and 2,205,581,786 comments are in Table 6.
450 The results show that, collectively, 17.61% of the posts were toxic (i.e., both highly toxic and slightly
451 toxic) and that the remaining 82.39% were nontoxic.

Table 6. Prediction results of the toxicity levels of the posts and comments of users.

Class	Posts (%)	Comments (%)
Highly toxic	1,794,115 (2.05%)	133,588,229 (6.06%)
Slightly toxic	6,364,092 (7.28%)	254,531,824 (11.54%)
Nontoxic	79,218,705 (90.66%)	1,817,233,194 (82.39%)

452 After obtaining the toxicity levels of user posts, we applied our method to judge user toxicity and get
453 the total number of users (and their toxicity percentages) in every quartile, as shown in Table 7. For users
454 who leave posts, the table shows that, in the 25th percentile, 26.27% of users had toxicity proportions
455 in the range (1%, 5%]. As for the 50th percentile, 25.56% of users had toxicity proportions that fell in
456 the range (5%, 9%]. Subsequently, in the 75th percentile, the toxicity proportions for 24.91% of users
457 were in the range (9%, 15%]. Additionally, Table 7 shows that, in the 100th percentile (i.e., the maximum
458 quartile), the toxicity proportions in 23.26% of users were in the range (15%, 100%]. Therefore, among
459 the four quartiles, the 25th percentile had the largest number of users; the average toxicity was 3.40%.

460 In the comments collection, for users in the 25th percentile, 26.55% had toxicity proportions in the
461 range (1%, 11%]. Concerning the 50th percentile, 23.64% of users had toxicity proportions with a range
462 of (11%, 16%]. The findings show that, in the 75th percentile, 24.96% of users had toxicity proportions
463 with a range of (16%, 22%]. Lastly, our results show that, in the 100th percentile, 24.85% of users had
464 toxicity proportions in the range (22%, 100%]. Simply put, in the 25th percentile of users who leave
465 comments, about 27% of the users have an average toxicity of 7.77%.

Table 7. Judgment of users' toxicity based on their predominant behavior. The judgments include the total number of users, their percentage (%), and the toxicity range of their posts and comments.

Users judgment	Posts (%) - toxicity range	Comments (%) - toxicity range
25th percentile	124,056 (26.27%) - (1%, 5%]	234,899 (26.55%) - (1%, 11%]
50th percentile	120,705 (25.56%) - (5%, 9%]	209,133 (23.64%) - (11%, 16%]
75th percentile	117,651 (24.91%) - (9%, 15%]	220,799 (24.96%) - (16%, 22%]
100th percentile	109,821 (23.26%) - (15%, 100%]	219,891 (24.85%) - (22%, 100%]

466 **Changes in Users' Toxicity across Communities**

467 Since some users do not show consistent toxic (or nontoxic) behavior in their content, with RQ2 (*Does the*
468 *toxicity of users' behavior change (a) across different communities or (b) within the same community?*),
469 we examine the content-based changes in users' toxicity. Here, we check if a change (or a multitude of
470 changes) in the toxicity of users occurs within the same subreddit or across different subreddits.

To examine toxicity changes in this study, we devised two different change conditions to look for in the users' collection. These conditions come from the two possible judgments of users' posting behavior based on their content's toxicity. Based on our methodology, we judge users based on their contributions as 1) toxic (slightly toxic and highly toxic) or 2) nontoxic. Based on these judgments, the conditions that we identified are as follows:

- **Condition 1:** Change in the toxicity of a user's contribution from nontoxic to toxic. (**NT→T**)

- **Condition 2:** Change in the toxicity of a user's contribution from toxic to nontoxic. (**T→NT**)

Additionally, this experiment checked whether the conditions were met within the same subreddit or across different subreddits. Given the criteria for investigating the change in toxicity, we examined the entire history of users' content in the posts and comments collections. First, we sorted content from oldest to newest based on time stamps. Then, we used the toxicity prediction labels that we obtained from our prediction model to check for the change conditions. For example, suppose the first (i.e., oldest or earliest) post of a user is nontoxic and the subsequent (i.e., following or newer) post is toxic. In that case, this user exhibits a change in toxicity due to condition 1. This experiment considers whether the change happened in the same subreddit or across multiple subreddits. Thus, we flagged every user based on the exhibited condition and identified the location of the change on Reddit. Subsequently, we used a majority voting mechanism to get the number of users that exhibited changes due to each condition and their locations on Reddit. We performed majority voting by getting a list of all the toxicity changes and their locations for every user. If most of the total changes for a user fall under a specific condition or location, then a user change is due to this particular condition. Moreover, any user who shows a single change in toxicity due to any condition was removed to avoid any issues that might arise by posting a single toxic post. For instance, if a user had two posts that showed a change in any of the conditions, we did not include it in the study. With this approach, we found that users can show at least two changes due to the one or two conditions at different locations on Reddit.

The majority voting technique allowed us to count users based on the overall change in their content. This approach resulted in 177,307 (30.68%) posting users and 727,587 (81.67%) commenting users that show changes due to conditions 1 or 2. To further analyze these users, the results depicted in Table 8 show the distribution of users who satisfy conditions within the same subreddit or across multiple subreddits in both posts and comments.

Table 8. Total number of users (and their %) that satisfy the conditions and their locations on Reddit in the posts and comments collections. NT: Nontoxic; T: Toxic.

	Submission users			Comment users		
Conditions	**Total**	**Same subreddit**	**Multiple subreddits**	**Total**	**Same subreddit**	**Multiple subreddits**
1. NT→T	23,946 (5.11%)	3,668 (15.32%)	16,377 (68.39%)	10,500 (1.19%)	2,608 (24.84%)	6,412 (61.07%)
2. T→NT	10,123 (2.16%)	1,468 (14.50%)	4,631 (45.75%)	7,891 (0.89%)	1,987 (25.18%)	4,572 (57.94%)
1. NT→T & 2. T→NT	435,000 (92.74%)	15,869 (3.65%)	419,131 (96.35%)*	866,228 (97.92%)	48,070 (5.55%)	818,158 (94.45%)**

Note: 50.80% of 419,131* and 68.63% of 818,158** show change in the same and multiple subreddits.

In Table 8, we measured the percentage of users that satisfy each condition along with the combined conditions and came up with compelling observations from each collection. For example, starting with the posts collection, we found that users show the most change due to conditions 1 and 2, where 96.35% of these users show a change across multiple subreddits. Additionally, 5.11% of these users show a change due to condition 1, where 68.39% of them show changes across multiple subreddits; similarly, the majority of users that satisfy condition 2 also show change across different subreddits. Specifically, 45.75% of these users show changes over multiple communities, meaning that most of the changes in toxicity among users that post occur in different communities.

As for the comments collection, just like the posts collection, users show the most change due to conditions 1 and 2, where 1.19% of these users show a change within the same subreddit. Moreover, 0.89% of these users show change due to condition 2, where 57.94% of them show changes across different subreddits. In other words, most commenting users show changes in their toxicity across different subreddits. So, concerning RQ2, findings show that engaging with multiple communities can cause users to exhibit changes in their toxic posting behavior.

To further illustrate the changes in toxic behavior, we count the total number of changes per user without considering the majority voting technique we used in Table 8. Then, we plot a histogram of

516 the counts of changes that occur due to condition 1, condition 2, and both conditions combined. The
517 histogram in Figure 3 shows that over time, most of the changes are in condition 1, where posting users
518 change their behavior from nontoxic to toxic.
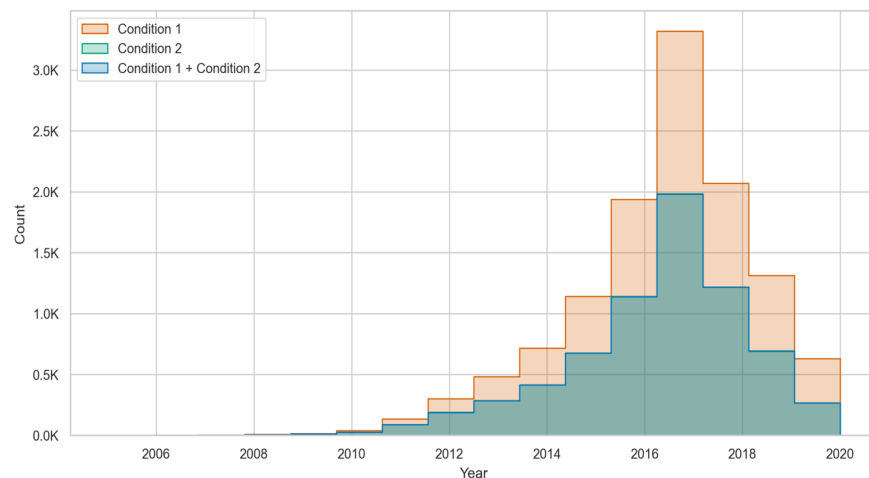
Figure 3. The count of toxicity changes over time in posting users from condition 1 (NT→T), condition 2 (T→NT), and both conditions combined.

519 Similarly, Figure 4 shows that for commenting users, most of the changes occur due to condition 1.
520 However, the gap between the counts of changes due to varying conditions is smaller in commenting users.
521 Furthermore, we found that posting users can show up to 37,800 changes in toxicity while commenting
522 users can show up to 295,912 changes due to both conditions. These high numbers of changes suggest
523 that users change their toxicity when their volume of contributions increases. In fact, these values result
     from having at least two changes (i.e., four posts or comments) from different conditions.
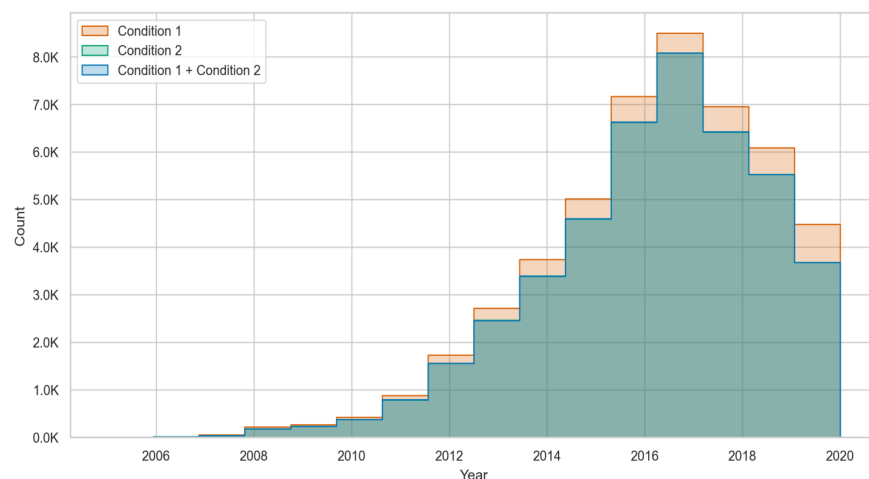
Figure 4. The count of toxicity changes over time in commenting users from condition 1 (NT→T), condition 2 (T→NT), and both conditions combined.

524

Changes in Users' Toxicity Over Time

526 Our collection's large volume of temporal data allows us to investigate toxicity changes over time.
527 Therefore, we chose yearly intervals to answer RQ3 (*Does the toxicity of users change over time across*
528 *different communities?*) and note observed changes in toxicity by computing the difference in the toxicity
529 of every user across all the subreddits.

With Equation 2, we computed users' toxicity percentages across all subreddits. Then, we calculated the change in toxicity (Δ) in every pair of years for posting and commenting users. Subsequently, we used scatter plots to visualize the change in toxicity per user across subreddits, which we then converted to heatmaps with varying smoothing parameters. The heatmap plots in Figure 5 show the distribution of toxicity in the posts and comments with smoothing at 64 neighbors. Sub-figures (A) and (B) show the heatmap plots for the posts from the years 2007–2008 and 2018–2019, respectively, while figures (C) and (D) show the heatmap plots for the comments collected from the years 2007–2008 and 2018–2019. To clarify the observations, we removed users who showed no change in their toxicity in the plots (i.e., users with Δ=0). Due to space limitations, we only show the heatmap plots from two pairs of years representing the collection's beginning and ending periods. The overall temporal analysis of the content shows that over time, changes in the toxicity of users' posts disperse across the participating communities, as illustrated by the increase in dark color in sub-figures (B) and (D) from Figure 5.
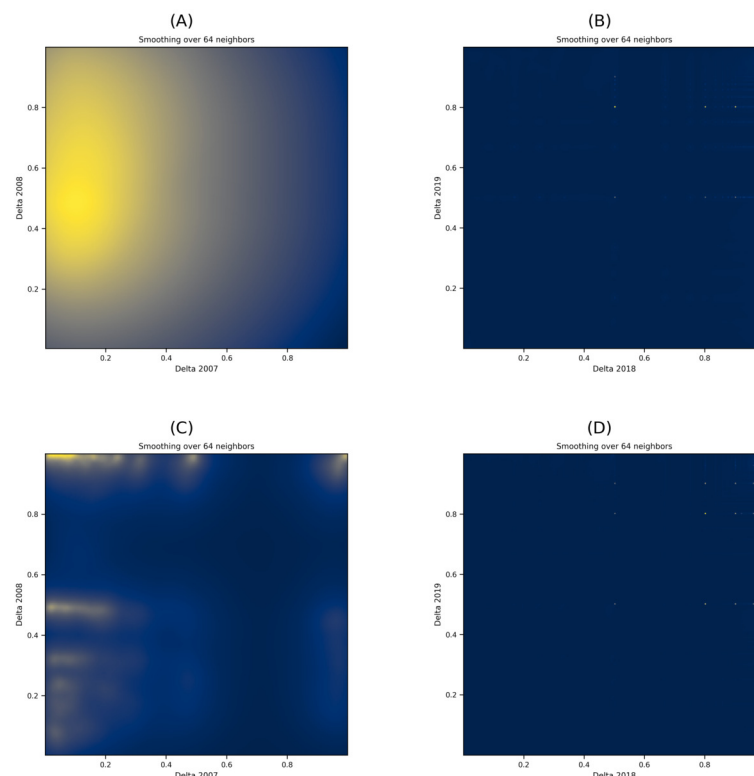
Figure 5. Heatmap plots of the Δ in user posts and comments over two pairs of years. The dark color in the heatmap plot denotes scattered deltas while the light colors denotes concentrated deltas in specific locations.

In other words, over time, more users diffuse their toxic behavior to a large number of varying communities. To further support this finding with users who leave comments, we conducted a dependent T-test on the commenting users' deltas for 2007 and 2019 (i.e., their initial delta and final delta). Results show that at $p < 0.001$, the t-statistic value of the users is 57.031, indicating a significant change in the posting behavior of users, which any of the conditions mentioned earlier can describe (change from toxic to non-toxic or change from non-toxic to toxic within the same subreddit).

Lastly, we visualize the Δ values of posting users in Figure 6, where we show the total change per year and interpolate the change by computing the smoothed rolling average on intervals of three years. The average line shows that changes peak in 2017 but drop after this point, suggesting some form of stability in the behavior of posting users. Similarly, Figure 7 shows that the average toxicity change in commenting users peaks in 2018 and drops slightly after this year. Unlike posting users, commenting users continue to show high amounts of change despite the drop after 2018.
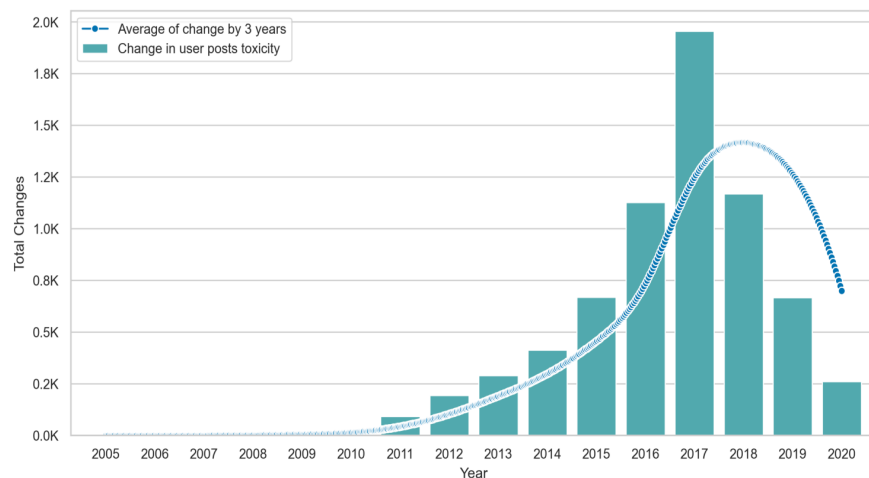
Figure 6. The total amount of Δ in posting users content over time with an interpolation of Δ averages across three year intervals.
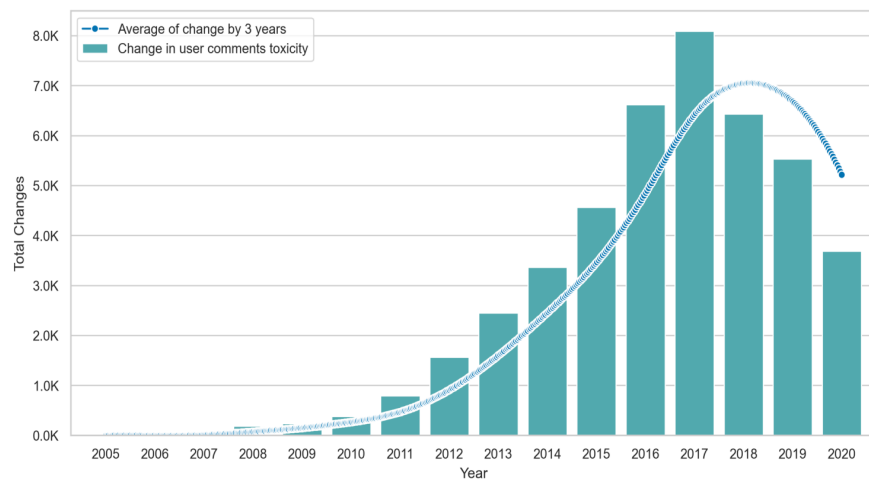
Figure 7. The total amount of Δ in commenting users content over time with an interpolation of Δ averages across three year intervals.

Changes in Toxicity and Links

Originally, Reddit was a news-sharing website where users posted links to news websites or various multimedia content to instigate discussions with other users. Hence, most of Reddit's earlier content (primarily posts) contained links. However, links are not limited to posts, as users can also include different types of links in their comments. Since our earlier investigation of toxic behavior focused on the textual content of user content, it is only natural to examine links to identify any correlation between toxicity in user content and certain types of links. To begin investigating links, we performed a preliminary exploratory analysis on the entire collection to identify the number of links and the percentage of links from the total number of posts and comments per year. The statistics illustrated in the top portion of Figure 8 show the total number of posts, toxic posts, and links in each year, while the bottom portion of the figure shows the corresponding normalized (i.e., scaled) totals using the minimum and maximum values from the totals. The accompanying values from Figure 8 are in Appendix C, where we also show the percentage of toxic posts in Table C1. Statistics from Table C1 show that between 2005 and 2012, more than 50% of user posts contained links, which means that, indeed, posts from the earlier years of Reddit contained a significant amount of links. Upon further investigation of the links in user content, we found external links that redirect to websites outside Reddit and internal links that redirect to a Reddit

570   user, post, community, or multimedia content uploaded on Reddit servers. Additionally, we found that some of the links in posts refer to videos, images, or other types of identifiable media.
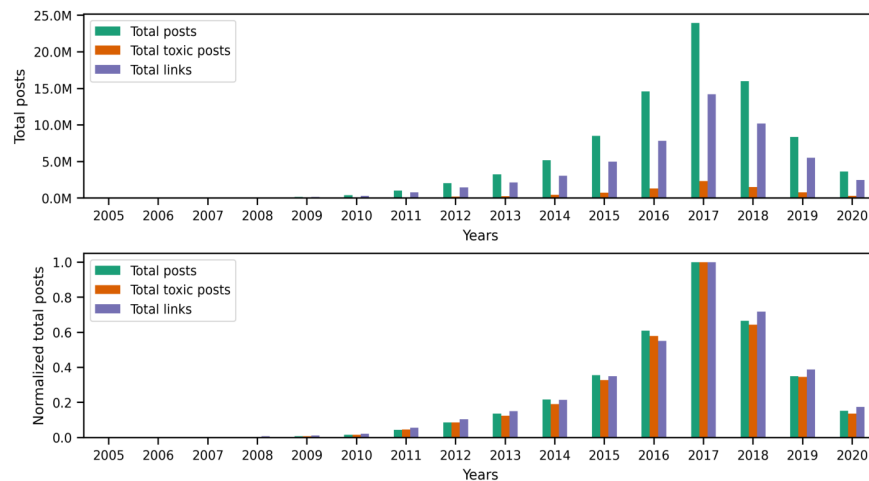
Figure 8. The total number of posts, toxic posts, and links in every year followed by the normalized totals using the min-max scale.

571
572   Our findings from the posts collection in Figure 9 and Table C2 show that around 84% of the links
573   in posts are external. As for the remaining internal links, we found that in posts, they typically link to
574   an image uploaded to the i.redd.it domain, or a video uploaded to the v.redd.it domain, or a Reddit user.
575   Figure 9 also shows the total number of links with identifiable media types. We used the mimetypes
576   python module to guess the media types from the link's text representation (i.e., path). So, if a link address
577   ends with .mp4, mimetypes identify it as a video without examining the link's content. In posts, we found
578   that the most identifiable media type happens to be images, so we calculated the total number of links
579   with images and the percentage of images from the known media types in links. The results in Table C2
       show that most of the links in posts from later years contain images. As for the comments collection, the
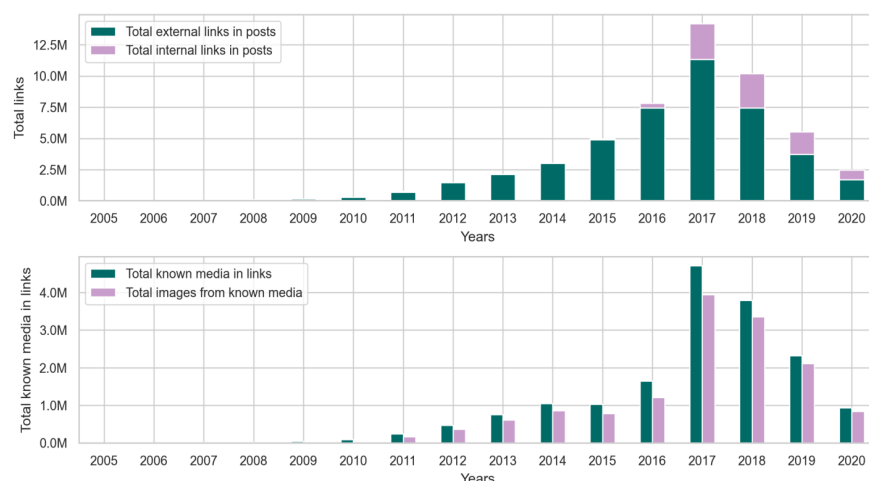
Figure 9. The distribution of internal and external links, followed by the total number of known media types and image links from the posts collection.

580
581   statistics in Figure 10 and Table C3 show that the percentage of links in comments is significantly less
582   than that of links in posts. However, this outcome does not diminish the fact that more than 110 million
583   comments contain links, which is about 5% of the entire comments collection. Furthermore, just like the
584   posts collection, Figure 11 shows that comments from earlier years in the collection contain more links

585 than comments in later years.
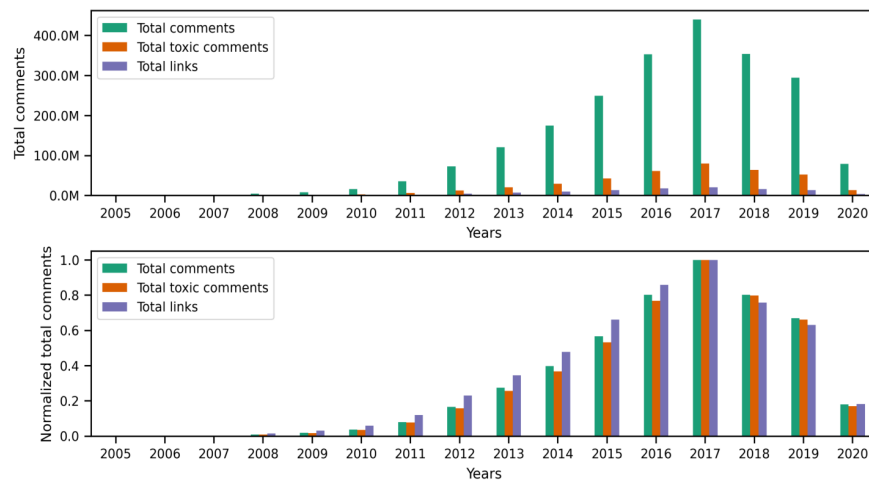
Figure 10. The total number of comments, toxic comments, and links in every year followed by the normalized totals using the min-max scale.

586 The results in Table C4 show that around 90% of the links in comments are external, and out of all
587 the media types identifiable in these links, images seem to appear the most in user comments. However,
588 when comparing the percentage of images in posts and comments, around 83% of links in posts contain
589 images, while 72% of links in comments contain images. This observation makes sense because many
communities on Reddit, such as r/cringepics, require users to post images in the community.
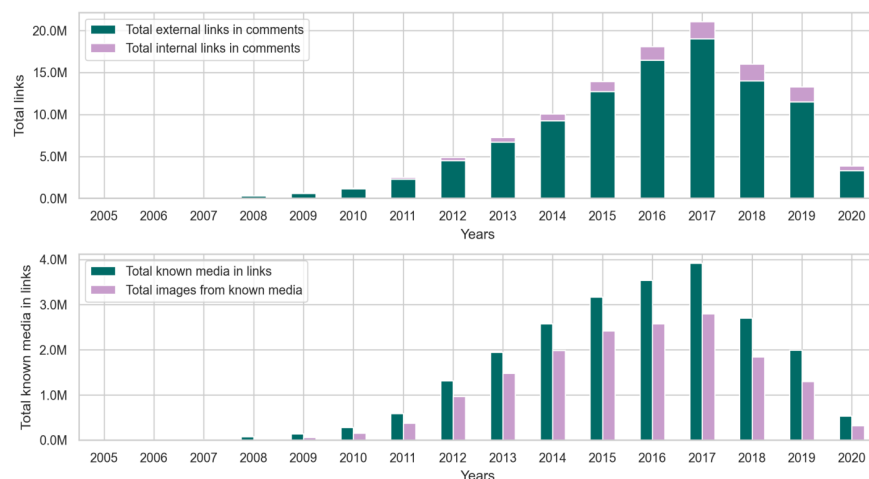
Figure 11. The distribution of internal and external links, followed by the total number of known media types and image links from the comments collection.

590
591 After performing the preliminary exploratory analysis of links in the collection, we used the Granger
592 causality test to find correlations between toxic behavior and links in posts and comments. First, we
593 conducted a test between the volume of content (X) in each collection and the volume of toxic (both
594 highly and slightly toxic) content (Y). Then, we conducted another test between the volume of links in
595 each collection (Z) and the volume of toxic (both highly and slightly toxic) content (Y). While our original
596 intention was to perform the Granger causality test on user posts as well, we found that, since almost
597 99% of posts contain links, running the test will not provide us with valuable insights on the relationship
598 between links in posts and toxicity. Moreover, our yearly time series in user posts did not produce a
599 stationary series, which does not satisfy the requirement to conduct the Granger causality test. Therefore,

we limit our experiments to user comments. Table 9 shows the F-statistic, p-value, and selected minimum
lags in years from conducting the Granger causality test on the volumes of comments, toxic comments,
and links in comments. The causality of toxicity in user comments can be observed in Table 9, where the
p-value < 0.05 for the volume of comments and links. In other words, the volume of comments and links
in comments influences the volume of toxic comments in the collection.

Table 9. Results of Granger causality for the comments collection at a minimum lag in years.

	F-statistic	p-Value	Lags (years)
(X,Y): X→Y	7.6306	0.014	2
(Z,Y): Z→Y	10.3849	0.014	3

IMPLICATIONS AND CONCLUSIONS

In this research, using over 10 thousand labeled comments, we trained feature-based and neural network-
based models to infer the toxicity of user posts. We then used a fine-tuned BERT model to analyze a
large-scale Reddit collection of more than 2 billion posts from more than 1.2 million users over 16 years
for a detailed analysis of the toxic behavior of users in multiple online communities. Our contributions
are three-fold:

- First, to our knowledge, we built one of the biggest labeled datasets of Reddit toxic comments
  and made it publicly available for further research. Additionally, compared to other binary labeled
  datasets, our dataset contains three levels of toxicity for each comment, ranging from non-toxic to
  slightly toxic to highly toxic.

- Second, by systematic comparisons of common feature-based models and neural network-based
  models, we demonstrate that a fine-tuned BERT model performs best for toxicity detection in posts
  and comments from Reddit.

- Third, our work is one of the first large-scale studies that investigate toxic behavior across multiple
  communities. We start with a list of cross-community users from the top 100 subreddits and expand
  our collection by obtaining posts and comments from more than 107,000 subreddits to reveal how
  users behave across communities from the perspective of toxicity.

Implications

Our work has several implications for the safety and moderation of online communities. These implications
include the following:

Early Detection of Changes in Toxicity:

The dissemination of toxicity in online communities impacts the positive experience many users seek
when using social media platforms. Several research studies showed that users could negatively influence
each other when interacting in online communities (Kwon and Gruzd, 2017; Zhang et al., 2018; Cheng
et al., 2017).

This type of negative behavior can continue to spread and harm online communities. Monitoring
the change in users' toxicity can be an early detection method for toxicity in online communities. The
proposed methodology can identify when users exhibit a change by calculating the toxicity percentage in
posts and comments. This change, combined with the toxicity level our system detects in users' posts,
can be used efficiently to stop toxicity dissemination. Furthermore, our methodology supports detecting
toxicity early in online communities from users' toxicity. In an active setting, users' toxicity percentages
can issue early alerts to online community moderators (bots or humans) so they can investigate potential
toxicity incidents and take necessary actions to mitigate the further spread of toxicity in communities.

Aid Moderators with Toxicity Changes:

Judging the toxicity of user content may not always be ideal for preventing the spread of hate and
incivility (Rajadesingan et al., 2020). Our study showed that users changed their posts' toxicity within and
across different communities. This change can result from fluctuations in the users' feelings or changes in
the atmosphere of their communities. This change, coupled with the toxicity of the users' content, can

643 create an accurate assessment method to prevent the spread of toxicity. For instance, instead of banning a
644 user for a tasteless contribution they left once, moderators can consider the users' predominant toxicity and
645 that of their previous content. This approach will prevent automated bots and moderators from excessively
646 penalizing or banning users. This sophisticated user- and content-based toxicity assessment allows
647 moderators to control toxicity and detect malicious users who deserve banning from online communities.
648 Trolls and the like (Cheng et al., 2017) can also be prevented from polluting online communities by using
649 our recommended method to judge users based on their content's predominant toxicity.

650 Moreover, the rules and norms of communities can be changed to prevent the spread of toxicity (Chan-
651 drasekharan et al., 2018). For example, when users show a rapid change in the toxicity of their behavior,
652 the moderation ecosystem might raise alerts/reminders for any breaches that do not conform to the
653 community norms. Lastly, our study suggests that one way to limit the spread of toxicity is by limiting
654 the spaces (i.e., communities) in which users can participate. To illustrate this finding, in Figure 12 we
655 show the correlation between the total amount of toxic posts (a), comments (b) and the total number
656 of communities that users participate in over time. The figures show a positive correlation between the
657 increase in the number of communities and the increase in toxicity. Ultimately, we cannot guarantee that
658 this is the only reason behind the increase in toxic content, yet we argue that increasing communities
659 could allow users to spread toxic content.
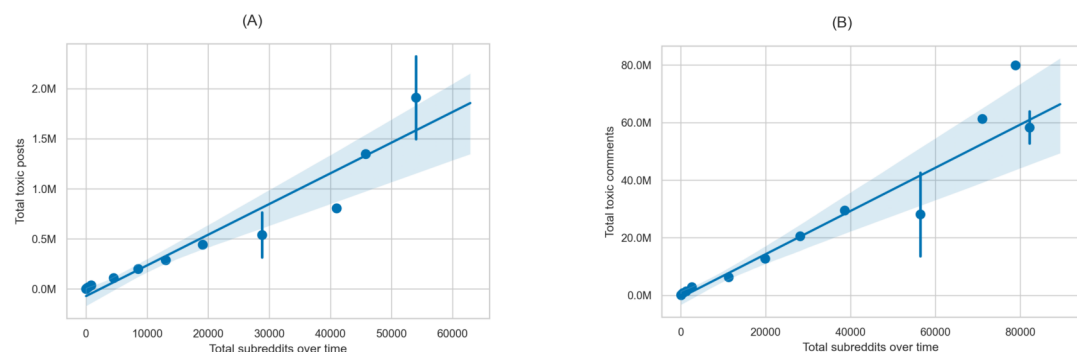
Figure 12. Correlation between the total number of participating subreddits over time and (a) the total number of toxic posts and (b) the total number of toxic comments.

660 **Limitations and Future Work**
661 Since our research focuses mainly on text analysis to detect toxicity, one limitation is that toxicity takes
662 different forms (e.g., images, videos, and sound clips). While more sophisticated techniques are required
663 to examine and analyze such content (Rafiq et al., 2015), multimedia submissions also have text titles
664 that we studied in this work. Another limitation of this work is that it does not fully consider bias in
665 the toxicity of the labels we obtained through crowdsourcing (Vaidya et al., 2020; Hanu et al., 2021).
666 However, since toxicity is a subjective matter, our study performed toxicity detection in a simplified
667 manner without accounting for subjectivity (Zhao et al., 2022). Lastly, we note that our study did not
668 tackle any contextual or categorical characteristics of toxic content (Radfar et al., 2020). That is partially
669 due to the heterogeneous nature of most Reddit communities, making it extremely difficult to capture
670 their context to judge different types of content, such as profanity in certain Not Safe For Work (NSFW)
671 communities (Madukwe and Gao, 2019).

672 Upon investigating the toxic posting behavior of users, we came across several ideas that can lead
673 to interesting future research directions. One of the ideas focuses on different scenarios involving users
674 joining new communities and considering the changes in their toxicity to these new communities (Cheng
675 et al., 2015; Choi et al., 2015; Cheng et al., 2017). Another take on the problem of toxic posting behavior
676 can focus on specific topics within each community (e.g., controversial topics or hot news) to study how
677 they trigger toxicity within users, as opposed to noncontroversial or regular topics (e.g., entertainment
678 news or funny stories). Besides focusing on various topics within online communities, one can also
679 study the temporal characteristics that foster the evolution of toxic communities from a few users with

680 predominantly toxic posts. Considering the factors that contribute to the rapid growth of such toxic
681 communities is also necessary for providing moderators and platform designers with the right tools to
682 prevent toxicity from contaminating online communities.

683 **Conclusions**

684 In this research, we investigated users' toxic cross-community behavior based on the toxicity of their posts
685 and comments. Our fine-tuned BERT model achieved a classification accuracy of 91.27% and an average
686 F_1 score of 0.79, showing a 2% and 7% improvement in performance compared to the best-performing
687 baseline models based on neural networks. We addressed RQ1 by running a prediction experiment on the
688 posts and comments from our Reddit collection. The analysis showed that 9.33% of the posts are toxic,
689 and 17.6% of the comments are toxic. We answered RQ2 by investigating the changes in the toxicity
690 of users' content across communities based on two primary conditions. First, our analysis showed that
691 30.68% of posting users showed changes in their toxicity levels. Moreover, 81.67% of commenting users
692 showed changes in their toxicity levels, mainly across multiple communities. Moreover, we found through
693 answering RQ3 that, over time, toxicity disperses with an increase in the number of participating users and
694 the frequency of cross-community participation. This finding is helpful because it can provide community
695 moderators with leads to help them track patterns from active users to prevent them from spreading toxic
696 content online.

697 Lastly, we conducted a Granger causality test between the volume of comments, the volume of links
698 in comments, and the volume of toxicity. We found that links in comments can influence toxicity within
699 those comments. This research addresses a prominent issue in social media platforms: toxic behavior
700 negatively impacts other users' experience. Thus, we believe it is necessary to conduct more research on
701 users' toxic behavior to help us understand the behavior's dynamics.

702 **REFERENCES**

703 Alfonso, F. and Morris, K. (2013). The most influential people on
704 reddit in 2013. Available at `https://www.dailydot.com/irl/`
705 `reddit-top-10-2013-quickmeme-unidan-boston-bombers/` (accessed 19 April
706 2019).
707 Almerekhi, H., Kwak, H., and Jansen, B. J. (2020). Investigating toxicity across multiple reddit communi-
708 ties, users, and moderators. In *Companion Proceedings of the Web Conference 2020*, pages 294–298,
709 New York, NY, USA. ACM.
710 Ashraf, N., Zubiaga, A., and Gelbukh, A. (2021). Abusive language detection in youtube comments
711 leveraging replies as conversational context. *PeerJ Comput. Sci.*, 7:e742. Publisher: PeerJ Inc.
712 Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in
713 tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages
714 759–760, Republic and Canton of Geneva, CHE. ACM.
715 Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods
716 for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29.
717 Baumgartner, J. (2017). Directory Contents. Available at `https://files.pushshift.io/`
718 `reddit/` (accessed 10 May 2021).
719 Bennett, E. M., Alpert, R., and Goldstein, A. (1954). Communications through limited-response question-
720 ing. *Public Opinion Quarterly*, 18(3):303–308.
721 Bowler, L., Knobel, C., and Mattern, E. (2015). From cyberbullying to well-being: A narrative-based
722 participatory approach to values-oriented design for social media. *Journal of the Association for
723 Information Science and Technology*, 66(6):1274–1293.
724 Carton, S., Mei, Q., and Resnick, P. (2020). Feature-based explanations don't help people detect
725 misclassifications of online toxicity. In *Proceedings of the International AAAI Conference on Web and
726 Social Media*, volume 14, pages 95–106. AAAI Press.
727 Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. (2017a).
728 You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of
729 the ACM on Human-Computer Interaction*, 1:31:1–31:22.
730 Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., and
731 Gilbert, E. (2018). The internet's hidden rules: An empirical study of Reddit norm violations at micro,
732 meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2:32:1–32:25.

Chandrasekharan, E., Samory, M., Srinivasan, A., and Gilbert, E. (2017b). The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3175–3187, New York, NY, USA. ACM.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1217–1230, New York, NY, USA. ACM.

Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015). Antisocial behavior in online discussion communities. In *Proceedings of the 9th International Conference on Web and Social Media*, pages 61–70. AAAI Press.

Choi, D., Han, J., Chung, T., Ahn, Y.-Y., Chun, B.-G., and 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*, pages 233–243, New York, NY, USA. ACM.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International Conference on Web and Social Media*, pages 512–515. AAAI Press.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the 1st Italian Conference on Cybersecurity*, pages 86–95. CEUR-WS.org.

Devlin, J., Chang, M.-W., Lee, K., and 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*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 29–30, New York, NY, USA. ACM.

Feinstein, A. R. and 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., and Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., and Plagianakos, V. P. (2018). Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, pages 35:1–35:6, New York, NY, USA. ACM.

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

Hanu, L., Thewlis, J., and Haco, S. (2021). How ai is learning to identify toxic online content. Available at `https://www.scientificamerican.com/article/can-ai-identify-toxic-online-content/` (accessed 10 June 2022).

Hu, X., Tang, J., Zhang, Y., and Liu, H. (2013). Social spammer detection in microblogging. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2633–2639. AAAI Press.

Jhaver, S., Ghoshal, S., Bruckman, A., and Gilbert, E. (2018). Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction*, 25(2):1–33.

Johnson, B. G. (2018). Tolerating and managing extreme speech on social media. *Internet Research*, 28(5):1275–1291.

Johnson, R. and Zhang, T. (2016). Supervised and semi-supervised text categorization using lstm for region embeddings. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, pages 526–534. JMLR.org.

Kaakinen, M., Oksanen, A., and Räsänen, P. (2018). Did the risk of exposure to online hate increase after the november 2015 paris attacks? a group relations approach. *Computers in Human Behavior*, 78:90–97.

Kapil, P. and Ekbal, A. (2020). A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 210:106458.

Kordyaka, B., Jahn, K., and Niehaves, B. (2020). Towards a unified theory of toxic behavior in video games. *Internet Research*, 30(4):1081–1102.

Kumar, S., Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*, pages 933–943, Republic and Canton of Geneva, CHE. ACM.

Kwon, K. H. and 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.

Lapidot-Lefler, N. and Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, 28(2):434–443.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, pages 1188–1196.

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

Madukwe, K. J. and Gao, X. (2019). The thin line between hate and profanity. In Liu, J. and Bailey, J., editors, *AI 2019: Advances in Artificial Intelligence*, pages 344–356. Springer International Publishing.

Maity, S. K., Chakraborty, A., Goyal, P., and Mukherjee, A. (2018). Opinion conflicts: An effective route to detect incivility in Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 2(117:1–117:27).

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

Mathew, B., Dutt, R., Goyal, P., and Mukherjee, A. (2019). Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, pages 173–182, New York, NY, USA. ACM.

Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., and Mukherjee, A. (2020). Hate begets hate: A temporal study of hate speech. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

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

Mohan, S., Guha, A., Harris, M., Popowich, F., Schuster, A., and Priebe, C. (2017). The impact of toxic language on the health of Reddit communities. In *Proceedings of the 30th Canadian Conference on Artificial Intelligence*, pages 51–56. Springer.

Mondal, M., Silva, L. A., and Benevenuto, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 85–94, New York, NY, USA. ACM.

Mukherjee, P. and Jansen, B. J. (2017). Conversing and searching: the causal relationship between social media and web search. *Internet Research*, 27(5):1209–1226.

Newell, E., Jurgens, D., Saleem, H. M., Vala, H., Sassine, J., Armstrong, C., and Ruths, D. (2016). User migration in online social networks: A case study on Reddit during a period of community unrest. In *Proceedings of the 10th International Conference on Web and Social Media*, pages 279–288. AAAI Press.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153, Republic and Canton of Geneva, CHE. ACM.

Obadimu, A., Khaund, T., Mead, E., Marcoux, T., and 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):102660.

Pelicon, A., Shekhar, R., Škrlj, B., Purver, M., and Pollak, S. (2021). Investigating cross-lingual training for offensive language detection. *PeerJ Comput. Sci.*, 7:e559.

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., and Rosso, P. (2021). Detecting ethnicity-targeted hate speech in russian social media texts. *Information Processing & Management*, 58(6):102674.

Radfar, B., Shivaram, K., and Culotta, A. (2020). Characterizing variation in toxic language by social context. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):959–963.

Rafiq, R. I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., and Mattson, S. A. (2015). Careful what you share in six seconds: Detecting cyberbullying instances in vine. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 617–622, New York, NY, USA. ACM.

Rajadesingan, A., Resnick, P., and Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 557–568. AAAI Press.

Rodriguez, N. and Rojas-Galeano, S. (2018). Fighting adversarial attacks on online abusive language moderation. In *Proceedings of the 5th Workshop on Engineering Applications*, volume 915, pages 480–493. Springer International Publishing.

Salminen, J., Almerekhi, H., Kamel, A. M., Jung, S.-g., and Jansen, B. J. (2019). Online hate ratings vary by extremes: A statistical analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 213–217. ACM.

Salminen, J., Almerekhi, H., Milenković, M., Jung, S.-g., An, J., Kwak, H., and Jansen, B. J. (2018). Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the 12th International Conference on Web and Social Media*, pages 330–339. AAAI Press.

Sazzed, S. (2021). Identifying vulgarity in Bengali social media textual content. *PeerJ Comput. Sci.*, 7:e665. Publisher: PeerJ Inc.

Shen, Q. and Rose, C. (2019). The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 58–69, Florence, Italy. ACL.

Shores, K. B., He, Y., Swanenburg, K. L., Kraut, R., and Riedl, J. (2014). The identification of deviance and its impact on retention in a multiplayer game. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1356–1365, New York, NY, USA. ACM.

Silva, L. A., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. (2016). Analyzing the targets of hate in online social media. In *Proceedings of the 10th International Conference on Web and Social Media*, pages 687–690. AAAI Press.

Singh, S., Thapar, V., and Bagga, S. (2020). Exploring the hidden patterns of cyberbullying on social media. *Procedia Computer Science*, 167:1636–1647.

Squicciarini, A. C., Dupont, J., and Chen, R. (2014). Online abusive users analytics through visualization. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 155–158, New York, NY, USA. ACM.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326.

Tsikerdekis, M. and Zeadally, S. (2014). Online deception in social media. *Communications of the ACM*, 57(9):72–80.

Vaidya, A., Mai, F., and Ning, Y. (2020). Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):683–693.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Vens, C. and Costa, F. (2011). Random forest based feature induction. In *IEEE 11th International Conference on Data Mining*, pages 744–753. IEEE Computer Society.

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 2021).

Wang, X. and Lee, E. W. (2021). Negative emotions shape the diffusion of cancer tweets: toward an integrated social network–text analytics approach. *Internet Research*, 31(2):401–418.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399, Republic and Canton of Geneva, CHE. ACM.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., and Edwards, L. (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7.

898 Yin, W. and Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and
899   solutions. *PeerJ Comput. Sci.*, 7:e598. Publisher: PeerJ Inc.
900 Zhang, J., Danescu-Niculescu-Mizil, C., Sauper, C., and Taylor, S. J. (2018). Characterizing online public
901   discussions through patterns of participant interactions. *Proceedings of the ACM on Human-Computer*
902   *Interaction*, 2:198:1–198:27.
903 Zhao, X., Liu, J. S., and Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of*
904   *the International Communication Association*, 36(1):419–480.
905 Zhao, Z., Zhang, Z., and Hopfgartner, F. (2022). Utilizing subjectivity level to mitigate identity term bias
906   in toxic comments classification. *Online Social Networks and Media*, 29:100205.
907 Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., and Xu, B. (2016). Text classification improved by integrating
908   bidirectional lstm with two-dimensional max pooling. In *Proceedings of the 26th International*
909   *Conference on Computational Linguistics*, pages 3485–3495, Osaka, Japan. ACL.

910

APPENDIX A

Table A1. The top 100 (1-50) subreddits ranked by the total number of subscribers along with the total number of posts and comments from our dataset.

Rank	Subreddit	Subscribers	Posts	Comments
1	r/funny	17,934,343	1,226,923	38,340,865
2	r/AskReddit	17,829,339	3,571,863	176,219,659
3	r/todayilearned	17,658,854	379,675	24,982,380
4	r/science	17,599,931	133,623	4,387,503
5	r/worldnews	17,522,963	613,496	31,844,773
6	r/pics	17,517,767	905,005	35,795,690
7	r/IAmA	17,215,966	34,839	7,650,945
8	r/gaming	16,860,176	762,776	25,502,647
9	r/videos	16,758,252	838,316	23,327,414
10	r/movies	16,468,377	407,215	17,824,942
11	r/aww	15,907,509	629,697	8,902,569
12	r/Music	15,879,988	400,517	4,567,182
13	r/gifs	15,051,932	175,824	12,300,726
14	r/news	14,994,220	896,236	25,720,157
15	r/explainlikeimfive	14,671,688	343,303	6,037,936
16	r/askscience	14,587,860	253,372	1,609,707
17	r/EarthPorn	14,197,666	92,811	1,178,788
18	r/books	13,699,914	83,049	2,786,204
19	r/television	13,617,822	129,974	6,005,034
20	r/LifeProTips	13,212,746	113,033	3,541,704
21	r/mildlyinteresting	13,170,674	319,910	7,369,887
22	r/space	12,915,830	77,426	2,126,598
23	r/Showerthoughts	12,797,856	944,483	9,315,365
24	r/DIY	12,753,699	52,804	1,647,926
25	r/Jokes	12,604,471	290,732	2,985,167
26	r/sports	12,552,016	129,084	2,422,532
27	r/gadgets	12,518,386	43,995	1,420,475
28	r/tifu	12,504,386	53,468	3,271,900
29	r/nottheonion	12,451,389	112,276	4,024,654
30	r/InternetIsBeautiful	12,433,228	29,919	338,448
31	r/photoshopbattles	12,363,401	93,001	624,593
32	r/history	12,356,415	48,904	1,105,441
33	r/food	12,351,889	181,129	2,401,415
34	r/Futurology	12,332,702	88,915	3,437,276
35	r/Documentaries	12,298,293	47,042	1,539,521
36	r/dataisbeautiful	12,293,355	36,084	2,016,558
37	r/listentothis	12,243,720	143,301	379,788
38	r/UpliftingNews	12,213,060	56,064	1,495,984
39	r/personalfinance	12,212,893	139,922	4,497,818
40	r/GetMotivated	12,135,696	50,356	935,431
41	r/OldSchoolCool	12,083,017	107,293	3,061,358
42	r/philosophy	12,081,626	23,948	713,078
43	r/Art	11,868,112	151,648	926,225
44	r/nosleep	11,678,653	23,966	607,853
45	r/creepy	11,665,125	46,469	1,147,867
46	r/WritingPrompts	11,612,067	206,881	948,065
47	r/TwoXChromosomes	11,215,698	52,283	3,002,720
48	r/Fitness	6,186,196	115,007	4,672,605
49	r/technology	5,551,587	267,264	9,225,651
50	r/WTF	4,861,274	239,730	19,032,277

Table A2. The top 100 (51-100) subreddits ranked by the total number of subscribers along with the total number of posts and comments from our dataset.

Rank	Subreddit	Subscribers	Submissions	Comments
51	r/bestof	4,772,718	41,570	1,869,232
52	r/AdviceAnimals	4,322,195	491,179	17,734,082
53	r/politics	3,468,561	955,613	60,494,833
54	r/atheism	2,096,408	147,972	8,890,785
55	r/europe	1,526,462	123,920	5,707,891
56	r/interestingasfuck	1,385,740	73,778	3,595,720
57	r/woahdude	1,345,016	58,643	1,514,561
58	r/leagueoflegends	1,118,408	550,318	22,450,314
59	r/gameofthrones	1,116,208	110,096	4,023,561
60	r/pcmasterrace	1,103,955	392,241	11,081,807
61	r/BlackPeopleTwitter	1,073,938	67,891	5,072,940
62	r/reactiongifs	1,038,629	86,617	1,488,555
63	r/trees	1,006,481	287,696	6,251,863
64	r/Unexpected	965,760	45,269	1,404,524
65	r/Overwatch	948,162	329,076	6,781,358
66	r/oddlysatisfying	905,675	53,992	1,683,927
67	r/Android	897,620	113,772	6,668,499
68	r/wholesomememes	840,077	36,283	1,003,541
69	r/Games	839,529	158,645	8,505,029
70	r/programming	826,809	51,209	3,208,560
71	r/4chan	819,656	38,445	2,338,530
72	r/nba	805,171	295,232	28,721,576
73	r/facepalm	791,286	36,112	1,955,898
74	r/cringepics	780,791	30,813	2,132,087
75	r/me_irl	779,311	435,999	2,081,621
76	r/relationships	774,812	61,107	5,975,894
77	r/sex	761,247	39,179	2,305,426
78	r/pokemon	760,949	120,266	3,992,989
79	r/fffffffuuuuuuuuuuuu	759,747	53,360	2,372,289
80	r/lifehacks	755,376	9,845	480,229
81	r/Frugal	741,976	24,157	1,611,244
82	r/soccer	736,005	231,161	20,685,778
83	r/tattoos	732,943	29,924	470,543
84	r/pokemongo	730,140	115,926	2,434,360
85	r/comics	726,976	85,320	1,128,707
86	r/OutOfTheLoop	688,156	55,979	1,249,050
87	r/malefashionadvice	684,010	56,821	2,314,703
88	r/CrappyDesign	667,846	78,494	1,376,476
89	r/StarWars	658,622	121,300	3,685,999
90	r/YouShouldKnow	644,359	8,068	577,679
91	r/AskHistorians	637,383	106,763	629,207
92	r/buildapc	635,055	281,713	5,049,608
93	r/nfl	626,637	197,693	33,747,257
94	r/HistoryPorn	626,507	33,838	647,963
95	r/RoastMe	622,922	23,436	1,808,683
96	r/loseit	613,079	47,978	1,388,423
97	r/FoodPorn	612,361	39,742	570,383
98	r/AnimalsBeingJerks	605,103	11,596	429,629
99	r/dankmemes	598,376	238,182	2,003,893
100	r/rickandmorty	586,805	51,538	1,017,558

911 **APPENDIX B**

Figure B1. The labeling task instructions that we provided to crowd workers.

His entire argument boils down to these 4 factors: *Continental differences in wild plant and animal species available for domestication *Factors affecting rates of diffusion and migration within continents *Factors affecting rates of diffusion and migration between continents *Continental differences in area/population size Explain to me how that is NOT environmental determinism.

Is this comment toxic? (required)

Please select at least one of these.

○ Toxic
○ Not toxic

ⓘ Choose one answer only

then the "Luke, I am your father" kind of means nothing. Though it didn't really to mean since I saw then in 4,6,5 order. With 6 being many, many times. That's back when things were on tv.

Is this comment toxic? (required)
○ Toxic
◉ Not toxic

ⓘ Choose one answer only

Too many people live there maybe? Or maybe it's because everyone but the creepy basement nerds skips town after 25 years only to show up at the holidays.

Is this comment toxic? (required)
◉ Toxic
○ Not toxic

ⓘ Choose one answer only

How toxic is the comment? (required)

	1	2	
Slightly Toxic	◉	○	Highly Toxic

Online blackjack is not played quite the same as in casinos (usually), as the deck (and its normally a single deck) is shuffled every hand. This removes any ability to card count, and therefore is stacked in the casino's favour.

Is this comment toxic? (required)
○ Toxic
◉ Not toxic

ⓘ Choose one answer only

Figure B2. The validation questions that crowd workers had to pass before beginning labeling.

912 **APPENDIX C**

Table C1. The total number of posts, toxic posts, and links in every year followed by the normalized totals using the min-max scale.

Year	Original Counts			Normalized Counts		
	Posts	Toxic (%)	Links (%)	Posts	Toxic	Links
2005	1,690	81 (4.79)	1,690 (100)	0.0	0.0	0.0
2006	10,917	895 (8.2)	10,917 (100)	0.0004	0.0004	0.0006
2007	47,556	4,372 (9.19)	47,556 (100)	0.0019	0.0018	0.0032
2008	105,825	10,407 (9.83)	101,357 (95.78)	0.0043	0.0044	0.007
2009	188,171	18,344 (9.75)	167,255 (88.88)	0.0078	0.0079	0.0117
2010	385,611	37,612 (9.75)	305,616 (79.26)	0.016	0.0162	0.0214
2011	1,058,576	108,253 (10.23)	791,010 (74.72)	0.0441	0.0466	0.0556
2012	2,052,406	201,278 (9.81)	1,472,632 (71.75)	0.0856	0.0866	0.1036
2013	3,247,906	287,250 (8.84)	2,133,632 (65.69)	0.1355	0.1236	0.1501
2014	5,176,179	442,283 (8.54)	3,054,626 (59.01)	0.2159	0.1904	0.215
2015	8,532,341	762,746 (8.94)	4,978,694 (58.35)	0.3559	0.3284	0.3505
2016	14,613,378	1,346,608 (9.21)	7,822,994 (53.53)	0.6097	0.5797	0.5508
2017	23,967,825	2,322,698 (9.69)	14,201,243 (59.25)	1.0	1.0	1.0
2018	15,968,062	1,496,033 (9.37)	10,201,097 (63.88)	0.6662	0.6441	0.7183
2019	8,379,712	804,135 (9.6)	5,523,004 (65.91)	0.3496	0.3462	0.3888
2020	3,640,757	315,212 (8.66)	2,468,515 (67.8)	0.1518	0.1357	0.1737

Table C2. The total number of internal links, external links, known media type links, and image links from the posts collection.

Year	Categories of links		Contents of links	
	Internal	External	Known media (%)	Images (%)
2005	0	1,690	565 (33.43)	9 (1.59)
2006	0	10,917	3,392 (31.07)	149 (4.39)
2007	0	47,556	13,001 (27.34)	1,267 (9.75)
2008	57	101,300	28,890 (28.5)	3,272 (11.33)
2009	197	167,058	45,240 (27.05)	8,270 (18.28)
2010	282	305,334	93,788 (30.69)	38,803 (41.37)
2011	81,426	709,584	245,682 (31.06)	166,934 (67.95)
2012	550	1,472,082	471,553 (32.02)	364,726 (77.35)
2013	1,078	2,132,554	751,834 (35.24)	610,537 (81.21)
2014	30,615	3,024,011	1,052,430 (34.45)	856,061 (81.34)
2015	70,917	4,907,777	1,032,840 (20.75)	788,774 (76.37)
2016	389,853	7,433,141	1,643,321 (21.01)	1,211,723 (73.74)
2017	2,862,721	11,338,522	4,710,886 (33.17)	3,940,218 (83.64)
2018	2,757,591	7,443,506	3,788,301 (37.14)	3,354,431 (88.55)
2019	1,803,981	3,719,023	2,323,104 (42.06)	2,109,346 (90.8)
2020	771,499	1,697,016	934,850 (37.87)	842,602 (90.13)

Table C3. The total number of comments, toxic comments, and links in every year followed by the normalized totals using the min-max scale.

	Original Counts			Normalized Counts		
Year	Comments	Toxic (%)	Links (%)	Comments	Toxic	Links
2005	310	26 (8.39)	38 (12.26)	0.0	0.0	0.0
2006	169,608	17,553 (10.35)	19,052 (11.23)	0.0004	0.0002	0.0009
2007	849,828	119,477 (14.06)	74,055 (8.71)	0.0019	0.0015	0.0035
2008	4,573,561	795,377 (17.39)	338,229 (7.4)	0.0104	0.01	0.0161
2009	8,494,022	1,450,220 (17.07)	674,146 (7.94)	0.0193	0.0181	0.032
2010	16,384,988	2,830,121 (17.27)	1,269,001 (7.74)	0.0372	0.0354	0.0602
2011	35,473,547	6,260,102 (17.65)	2,531,916 (7.14)	0.0806	0.0783	0.1202
2012	72,943,244	12,704,594 (17.42)	4,883,585 (6.7)	0.1657	0.1589	0.2318
2013	121,155,630	20,530,863 (16.95)	7,262,543 (5.99)	0.2752	0.2569	0.3447
2014	175,223,888	29,461,807 (16.81)	10,063,833 (5.74)	0.398	0.3686	0.4777
2015	249,496,457	42,597,414 (17.07)	13,934,125 (5.58)	0.5667	0.5329	0.6614
2016	352,996,950	61,346,847 (17.38)	18,100,953 (5.13)	0.8017	0.7675	0.8591
2017	440,297,137	79,930,465 (18.15)	21,068,587 (4.79)	1.0	1.0	1.0
2018	353,701,991	63,881,250 (18.06)	15,995,772 (4.52)	0.8033	0.7992	0.7592
2019	294,450,367	52,818,942 (17.94)	13,309,224 (4.52)	0.6688	0.6608	0.6317
2020	79,370,258	13,603,534 (17.14)	3,867,698 (4.87)	0.1803	0.1702	0.1836

Table C4. The total number of internal links, external links, known media type links, and image links from the comments collection.

	Categories of links		Contents of links	
Year	Internal	External	Known media (%)	Images (%)
2005	0	38	11 (28.95)	0 (0.0)
2006	1	19,051	4,679 (24.56)	556 (11.88)
2007	2	74,053	17,253 (23.3)	3,420 (19.82)
2008	12,299	325,930	74,095 (21.91)	23,315 (31.47)
2009	54,658	619,488	141,166 (20.94)	60,645 (42.96)
2010	99,215	1,169,786	280,332 (22.09)	154,372 (55.07)
2011	206,306	2,325,610	590,792 (23.33)	378,633 (64.09)
2012	370,074	4,513,511	1,317,284 (26.97)	970,935 (73.71)
2013	558,468	6,704,073	1,944,145 (26.77)	1,479,230 (76.09)
2014	779,987	9,283,846	2,575,111 (25.59)	1,988,173 (77.21)
2015	1,181,543	12,752,582	3,171,927 (22.76)	2,416,944 (76.2)
2016	1,630,462	16,470,491	3,538,101 (19.55)	2,577,966 (72.86)
2017	2,055,250	19,013,337	3,920,151 (18.61)	2,801,227 (71.46)
2018	1,974,185	14,021,587	2,702,087 (16.89)	1,840,742 (68.12)
2019	1,776,779	11,532,445	1,990,511 (14.96)	1,300,974 (65.36)
2020	533,875	3,333,823	536,813 (13.88)	323,796 (60.32)