Users' Behavioral and Emotional Response to Toxicity in Twitter Conversations

Ana Aleksandric[1,2], Sayak Saha Roy[1], Hanani Pankaj[1], Gabriela Mustata Wilson[2], Shirin Nilizadeh[1,2]

[1] University of Texas at Arlington, Department of Computer Science and Engineering, Arlington, TX, USA
[2] University of Texas at Arlington, Multi-Interprofessional Center for Health Informatics, Arlington, TX, USA
axa8470@mavs.uta.edu, sayak.saharoy@mavs.uta.edu, hjp6626@mavs.uta.edu, gabriela.wilson@uta.edu, shirin.nilizadeh@uta.edu

Abstract

Prior works have shown connections between online toxicity attacks, such as harassment, cyberbullying, and hate speech, and the subsequent increase in offline violence, as well as negative psychological effects on victims. These correlations are primarily identified through user studies conducted via virtual environments, simulations, and questionnaires. However, no work has investigated how, in practice and authentically, people react to online toxicity both emotionally, showing anger, anxiety, and sadness, and behaviorally in terms of engaging with and responding to toxicity instigators, considering conversations as a whole and the relation between emotions and behaviors. This data-driven study investigates the effect of toxicity on Twitter users' behaviors and emotions by considering confounding factors, such as account identifiability, activity, and conversation structure and topic. We collected about 80K Twitter conversations and identified those with and without toxic replies. Performing statistical tests along with propensity score matching, we investigated the *causal association* of receiving toxicity and users' responses. We found that authors of conversations with toxic replies are more likely to engage in conversations, reply in a toxic way, and unfollow toxicity instigators. In terms of users' emotional responses, we found that sadness and anger after the first toxic reply are more likely to increase as the amount of toxicity increases. These findings not only emphasize the negative emotional and behavioral effects of online toxicity on social media users but also, as demonstrated in this paper, can be utilized to build prediction models for users' reactions, which could then aid the implementation of proactive detection and intervention measures helping users in such situations.

Introduction

Social media is rampant with toxic content, characterized by offensive language, trolling, and the propagation of hate speech. These attacks predominantly aim to silence, insult, or demoralize individuals, particularly those from marginalized communities (Tahmasbi et al. 2021; Fredericks and Bradfield 2021). These attacks are usually targeted, e.g., as part of a smear campaign to damage or call into question someone's reputation (Hannan 2018). They can also be coordinated using other communication mediums and implemented by many users (Zannettou et al. 2020a; Tahmasbi et al. 2021). Previous literature highlighted some of the consequences of online harassment, cyberbullying, and trolling on the psychological well-being of victims (Kowalski et al. 2014), who expressed psychological distress, suicidal ideation (Giumetti and Kowalski 2022), self-harming behaviors, depression and anxiety (Hellfeldt, López-Romero, and Andershed 2020). In addition to emotional responses, victims may exhibit some specific and possibly negative behavioral reactions to such attacks. For example, an online cyberbullying study (Erişti and Akbulut 2019) found that victims primarily engaged in four types of behavioral reactions: *avoidance*, *revenge*, employing *countermeasures*, and *negotiation*. These findings are mostly identified through user studies conducted via questionnaires (Ortega et al. 2012), virtual experiments (Wright et al. 2009), or simulations (Alhujailli et al. 2020). Yet, data obtained from social media presents a unique opportunity to observe users' authentic emotions and behavioral reactions in online discussions, serving as a robust foundation for conducting data-driven studies to understand the effects of toxicity on users.

To the best of our knowledge, no work has conducted a large longitudinal data-driven study to examine the effects of toxic content on social media users' emotions and online behavior, considering conversations as a whole and the relation between emotions and behaviors. In this paper, our main research question is to understand how social media users *respond* to toxic content in terms of their behavioral actions on Twitter (Now X), as well as emotions captured in their replies. Firstly, we leverage the behavioral reactions (*avoidance*, *revenge*, *countermeasures*, and *negotiation*) defined in the literature (Erişti and Akbulut 2019), as a framework for creating meaningful groupings of behavioral responses to toxic content. In particular, we examine if the users try to *avoid* further encounters with toxic content by ignoring toxic replies, tend to *negotiate* by posting comments in conversations, even *take revenge* by responding in a toxic way, or *employ countermeasures* by unfollowing the toxicity instigators. Furthermore, prior work has shown that sadness and anger (Mameli et al. 2022) as well as anxiety (Kwan et al. 2020) are among the most prevalent emotional responses to cyberbullying. Therefore, we examine the impact of toxicity on the emotions of anger, sadness, and anxiety of users who received toxic replies and engaged in such conversations.

We collected a sample of 79,799 Twitter conversations from August 14th to September 28th, 2021. Then, we identified conversations with toxic replies where each conversation was represented as a reply tree where two tweets are connected if one is a reply to another. In our analysis, we consider factors that could have an impact on users' responses, including the number of toxic replies in conversations, the structure of conversations, the location of the toxic content in the conversations tree, and the conversation topic. We also explore the effects of account-specific attributes, including their online visibility, identifiability, activity level, and emotions expressed in their initial tweets. We formulated eight hypotheses to understand the *causal association* of toxicity and user behavioral and emotional responses: **H1**: Root authors of conversations with toxic replies are more likely to engage in conversations. **H2**: Root authors of conversations with toxic replies are more likely to engage in conversations if they are posted by a larger number of toxicity instigators. **H3**: Root authors of conversations with toxic replies are more likely to respond back in a toxic way. **H4**: Root authors of conversations with toxic replies are more likely to respond back in a toxic way if toxic replies are posted by a larger number of toxicity instigators. **H5**: Root authors of conversations with toxic replies are more likely to unfollow toxicity instigators. **H6**: A larger amount of toxicity increases users' anxiety. **H7**: A larger amount of toxicity increases users' anger. **H8**: A larger amount of toxicity increases users' sadness.

Our study yields multiple important findings. From the regression analysis, we observed that different users respond differently to toxic content. For example, in terms of *avoidance*, we demonstrate that 36% of users of conversations with toxic replies ignored toxic content and did not show any behavioral reactions. In terms of *negotiation*, we found users of conversations with toxic replies compared to others are more likely to engage in the conversations (60.7% vs. 48.04%) ($p < 0.0001$), and 11.6% of users employed *countermeasures* by unfollowing toxicity instigators ($p < 0.0001$). Also, the results indicate that the location of toxic content in the conversation and the user's account characteristics can affect users' reactions. For example, verified accounts are less likely to engage in the conversation or respond in a toxic way compared to non-verified accounts.

Predicting users' reactions to a potentially toxic conversation can significantly aid in moderation strategies by providing timely intervention and support to the affected victims. Thus, we utilized several features that were found to be effective in our aforementioned regression study to train classification models that can predict user behavior in response to toxic replies. Our models show great performance, having accuracy of 94%, 82%, and 92% for identifying whether the tweet author will further engage in the conversation, engage in a toxic manner, or unfollow the toxic comment instigator. Furthermore, our emotional investigation indicates that while the amount of toxicity does not play a significant role in changing the anxiety of users when receiving toxic replies, higher toxicity leads to users being more likely to express sadness and anger. Moreover, expressing emotions before the first toxic reply is likely to lead to boosting such emotions in the rest of the conversation. The data and the code can be found here, where we provided all unique tweet IDs: https://zenodo.org/records/10904849. Also, we make sure the dataset sticks to the FAIR guidelines (FORCE11 2020).

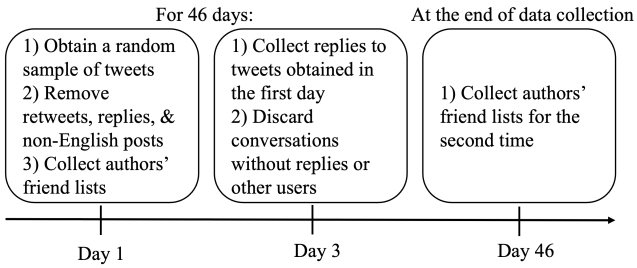

Figure 1: Data collection process.

Related Work

Prevalence of Social Media Victimization. Since abuse and toxicity are often carried out to humiliate or manipulate targeted individuals (Parent, Gobble, and Rochlen 2019; Varjas et al. 2010) , social media is often considered to be the chief outlet for housing such attacks due to high visibility (Yang, Ye, and Wang 2021; Duffy and Hund 2019) and more opportunities to remain anonymous (Schlesinger et al. 2017). Prior work has examined the association between abuse and on-the-ground "trigger" events, *e.g.,* terrorist attacks, and political events (Olteanu et al. 2018), suggesting that they lead to an increase in hateful comments (Zannettou et al. 2020a). The prevalence and characteristics of hate speech have been studied on specific web communities, such as r/Gab (Mathew et al. 2020) , 4chan's Politically Incorrect board (/pol/) (Hine et al. 2017), Twitter (Yousefi et al. 2023b; Zannettou et al. 2020b; Maarouf, Pröllochs, and Feuerriegel 2022), Whisper (Silva et al. 2016), and Parler (Kumarswamy, Singhal, and Nilizadeh 2023) while online abuse is normalized in several communities (Beres et al. 2021). A few efforts aimed to understand the characteristic differences between hate targets and hate instigators (ElSherief et al. 2018b; Ribeiro et al. 2018; ElSherief et al. 2018a) while others explored longitudinal behaviors of abusive accounts on Reddit (Kumar et al. 2022), and identified the circumstances when victims are more likely to respond to trolls (Sun and Shen 2021). Finally, the literature also examines the bystander effect on social media (Wong et al. 2021; Aleksandric et al. 2022), as well as the impact of toxic replies to conversations (Salehabadi 2019).

The Psychological and Emotional Impacts of Online Abuse. Established literature on the psychological impacts of online abuse is mostly focused on cyberbullying. A study determined that middle and high school students who had been cyber-bullied were more prone to exhibit self-harming behavior as well as suffer from depression, anxiety, and lower self-esteem (Eyuboglu et al. 2021). Victims might also respond by acceptance and self-blame (Veletsianos et al. 2018), with those having lower psychological endurance being more vulnerable to emotional outbursts.

Moreover, literature found that problems with emotion regulation increase the likelihood of individuals cyberbullying others or becoming the victim of cyberbullying (Arató et al. 2022). Perpetrators and victims show different sets of emotions, where victims are likely to express sadness, humiliation, and embarrassment (Gianesini and Brighi 2015). Also, anger received attention from researchers investigating cybervictimisation (Ak, Özdemir, and Kuzucu 2015), where anger is shown as the most common reaction to cyberbullying (Campbell et al. 2012) as well as sadness (Raskauskas and Stoltz 2007).

No prior data-driven studies aimed to evaluate the impact of online attacks on individual behavioral and emotional reactions in the online setting. Thus, this is the first observational study analyzing social media data to examine the effect of toxicity on users' online behaviors as well as expressing emotions of anger, anxiety, and sadness.

Data Collection

Figure 1 shows the pipeline used for collecting and processing our datasets.

Daily Collection of a Random Sample of Twitter Conversations: We used the Twitter API (Twitter 2022) to collect 1% random sample of tweets for 46 consecutive days, starting from August 14th till September 28th, 2021, and extracted English tweets that are not retweets or replies belonging to other conversations. For each initial tweet, we waited at least two days before collecting the replies of the conversation. For example, if we collected a random sample of tweets on September 1st, we would start collecting replies for each of these tweets on September 3rd. This is to give enough time so that the initial tweet can turn into a conversation. However, we discarded many tweets that did not receive any comments or had been deleted by the time we attempted to collect their replies. We also removed the conversations where the replies for the tweets were all posted by the author of the initial tweet.

Conversations and Reply Trees: As suggested in the previous literature, conversation structure plays a significant role in conversation dynamics (Saveski, Roy, and Roy 2021). Therefore, we collect and analyze the whole conversation, as we believe that the way the whole discussion unfolds has an impact on users' behavioral and emotional reactions in that conversation. We used Twarc (Twarc 2020), a Python wrapper for the official Twitter API, to obtain the entire conversation for each initial tweet, including its direct replies and nested replies (replies to replies). As it is shown in Figure 2, we represent each conversation as a *reply tree*, where one tweet is *child* of another tweet when one is a reply to the other. The initial tweets represent the roots of *reply trees*. We call the author of the root tweet as *root author*. We also define *direct replies* as the first level replies in a conversation, which is the set of replies to the root tweet. *Nested replies* refer to other levels of replies in a conversation other than the first layer, i.e., replies to replies. As shown in Figure 2, each *reply tree* has the following properties: *Size*, which indicates the number of tweets in the conversation; *Depth*, which is the depth of the conversation's deepest node; *Width*, which is the maximum number of nodes at any depth in the tree.

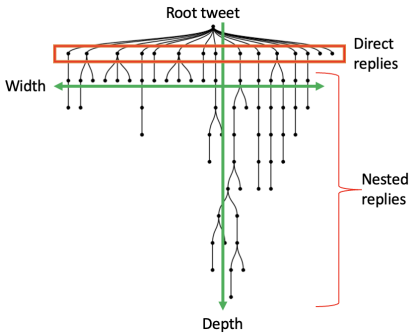

Figure 2: An example of a conversation tree.

Pre-processing and Filtering: Getting the replies for tweets of each day can take a long time, up to a couple of hours. Therefore, the first tweets collected for a certain day would have much less time to receive any comments than the last tweets obtained from that day. In order to solve this problem, we only kept replies sent in the first two days (48 hours) after the root tweet was posted. Thus, even though we would start collecting replies two days after the initial tweets were posted, due to the long time that collecting replies might take, some replies would be more than two days old. Therefore, discarding such replies ensures that all the replies collected were posted in the same time frame. Also, some tweets contained only links, images, and videos instead of text. Since our approach of detecting toxic tweets is text-based, such conversations were removed from the dataset. Moreover, certain *reply trees* were missing some replies due to the errors received during their data collection. Since these errors can have multiple reasons including, replies being deleted or hidden by their authors or root authors, we removed trees containing such errors from our dataset. Furthermore, we noticed that not all the root authors in our sample are unique, therefore to avoid duplication in our statistical analysis, we randomly selected a single conversation from each root author. **Our Conversation Dataset:** Finally, our dataset consists of 79,799 conversations with 528,041 tweets, posted by 328,390 unique users, out of which 79,799 are root authors.

As illustrated in Figure 1, we collected the followers and friends lists of all root authors every day, once the daily random sample was obtained. Thus, the lists collected at this time represent a snapshot of the authors' friends/followers lists before their tweets turned into a conversation. Furthermore, around 46 days after the first day when root tweets were collected, we again collected the list of friends and followers for all root authors. This is to analyze the impact of toxicity on users' online relationships. We could not collect the relationships right after obtaining the conversations, because of the number of API calls we could issue every day. This delay in collecting relationships imposes some limitations as users might end followership and friendships due to other events during this time. However, in our analysis, we compare these variables of users who received toxic replies and random authors, and seeing a difference can be an indicator of the impact of toxicity on Twitter relationships.

From our *unfriend* analysis, we had to discard 2,488 conversations, where 267 and 2,221 are conversations with and without toxic replies respectively because we were not able to obtain the friends' list for their root authors due to the authors either making their accounts private or deactivated.

Discovering Conversation Topics: The dataset used in this study contains a random sample of Twitter conversations which can potentially include a large number of topics. However, there might be some topics that provoke more emotional responses from the users involved in the discussion. For example, users might get more angry if they receive a toxic reply concerning their political views, or they might get sad if the toxicity is directed at their health-related decisions. Thus, a recent topic classification model (Antypas et al. 2022) used in previous studies (Leiter et al. 2024) was utilized to determine the main topic of the conversation by passing the text of the main tweet as the input. Note that this model has been fine-tuned for multi-label classification on 11,267 tweets yielding 19 discussion topics such as *news & social concern*, *diaries & daily life*, *business & entrepreneurs*, and others. Scores obtained per topic are in range from 0 to 1, where a higher score suggests that the text is more related to that topic.

Detecting Emotions: We used LIWC-22 (Boyd et al. 2022) to detect the emotions of each tweet in the dataset. This tool analyzes text to provide insights into the person's emotions, social and cognitive processes, etc. It has been used for psychological analysis of users online (Lyu et al. 2023) and it has shown a decent performance for detecting emotions in verbal expression (Kahn et al. 2007). It treats each tweet individually and provides scores for each post in range $0-99$, representing its relation to a specific attribute.

Detecting Toxic Replies: Many tools have been developed to detect toxicity in text (Androcec 2020). Recently, ChatGPT[1] has been used for detecting offensive language and hate speech (Huang, Kwak, and An 2023). Google's Perspective API has also been widely used for toxicity detection on social media (Chong and Kwak 2022). This API processes the given text input and provides output scores such as *Severe toxicity*, and *Toxicity*, in the range from 0 to 1, where a score closer to 1 means a higher severity for a specific attribute. Perspective API's scores have been evaluated in previous studies. Some works suggest considering a score greater than 0.5 (Habib and Nithyanand 2022) or 0.8 (Horta Ribeiro et al. 2021) as toxic. Thus, we conducted an experiment to compare the performance of Google's Perspective API and OpenAI API in the toxicity detection task.

Creating the Ground-truth Dataset: We initially obtained toxicity scores using Perspective API for each tweet once data collection was completed in 2021. However, it was not clear which threshold should be used to detect toxic replies in the dataset. To evaluate the scores, we manually extracted a random sample of 50 toxic conversations that contained at least one reply with the *Severe toxicity* score higher than 0.5, and 50 with no reply with the *Severe toxicity* score higher than 0.5. The total number of tweets included in the random sample was 943 (843 replies). Then, four annotators from different backgrounds manually labeled each tweet as 1 for toxic, and 0 if the reply is not toxic by observing the whole conversation, trying to capture the context of conversations. Two annotators coded 50 (25 toxic and 25 non-toxic) conversations, and the other two annotators coded the rest. The computed Cohen's Kappa score (Kvålseth 1989) was 0.5 showing a 93.5% agreement. Kappa score of 0.5 indicates a moderate agreement, suggesting that a comment being perceived as toxic can be subjective, as what might be found offensive by one person does not necessarily mean it will be perceived similarly by others. Finally, the ground-truth dataset consists of 64 (6.8%) toxic tweets belonging to 26 conversations and 879 (93.2%) non-toxic tweets.

Obtaining New Toxicity Scores: We re-ran Perspective API on the dataset as the new version of the API has been released (Lees et al. 2022). The goal was to find a threshold that reaches the highest accuracy compared to the manually labeled sample. For that, we used *Severe toxicity* and *Toxicity* attributes. As the scores provided by the API are from 0 to 1, we increased the testing value by 0.1 in each iteration to find the most accurate threshold for classification.

GPT Labels: We used the OpenAI API *gpt−3.5−turbo−16k* version to obtain the binary toxicity labels. We passed the following prompt *"Given the following post, determine if it is contextually toxic. Respond in an array format [toxic_value, explanation], where the first element is either '1' for yes or '0' for no, and the second element is the explanation for the value."* to the API with each reply individually. However, two parameters could be changed when passing prompts to the OpenAI API: *temperature* and *top_p*. According to the documentation[2], it is not suggested to change both parameters at the same time. Thus, as the *temperature* indicates the randomness of the output, we decided to test how different temperatures affect the accuracy of the classification. Once again, we increased it by 0.1 in each iteration to determine which temperature provided the best results. Obtained labels were compared with the manually labeled sample.

Evaluating the Performance: We compared the labels obtained from Perspective API and OpenAI to the manually labeled sample. Perspective's Toxicity attribute outperforms both Severe Toxicity and gpt−3.5−turbo−16k with the highest accuracy of 0.95 for the threshold of 0.6. In more detail, increasing the threshold from 0.1 to 0.6 for Perspective API also increased the accuracy in each iteration from 0.58-0.95 for Toxicity. Changing the temperature of OpenAI model from 0.1-0.9 did not significantly change its performance, and remained between 0.87-0.89. Therefore, we will use this threshold to detect toxic replies in the dataset. Any reply that shows a score equal to or higher than 0.6 in Toxicity is considered as *toxic*. We define **Conversations with Toxic Replies (CTR)** as conversations with toxic replies, which are shared by users other than the root author (7,205), and the conversations with no toxic replies are called **Conversations with No Toxic Replies (CNR)** (72,594).

[1]https://chat.openai.com/. Accessed: 2024-04-09.

[2]https://openai.com/research/overview. Accessed: 2024-04-01.

Variables and Statistical Models

We use causal inference and multivariate regression to study the associations between toxicity and Twitter users' responses. To provide a larger context for the interpretation of our analyses, we compare the root authors' emotions and behaviors of CTR and CNR.

Dependent Variables. These variables represent our understudy user responses, including: (1) *#root_author_replies*: a numeric variable indicating the number of root authors' replies. Some users might decide to *negotiate* about their point of view and participate in the conversation. (2) *#root_author_toxic_replies*: the number of root authors' toxic replies per conversation. This can capture the *revenge* behavior when the user responds to the toxicity in a toxic way. (3) *unfollowing*: a binary variable that shows if the user unfollows another user. This can be considered as a *countermeasure*, as the user tries to prevent further communication with the toxicity instigator. (4) *anxiety_after*: a numeric variable representing the average root author's anxiety after a toxic reply occurs in a conversation. In more detail, we sum emotion scores for all the root author's replies after the first toxic reply and then divide by the number of replies posted by the root author after the first toxic reply. The same procedure applies to other emotions. (5) *anger_after*: a numeric variable representing the average root author's anger after a toxic reply occurs in conversation. (6) *sadness_after*: a numeric variable representing the average root author's sadness after a toxic reply occurs in a conversation. Note that the variables 4-6 were rounded up to the closest integer for the analysis purposes. Also, they are computed only in CTR.

Independent Variables. The location of toxic replies in the tree structure might affect the users' behavioral and emotional reactions. For example, as shown in Figure 2, direct replies might be more visible compared to nested replies, as direct replies are immediate replies to the root tweet whereas nested replies might be found at the bottom of the tree. Therefore, we considered two independent variables: (1) *direct_toxicity*, a numeric variable defined as the ratio of the toxic direct replies to the total number of direct replies. (2) *nested_toxicity*, a numeric variable defined as the ratio of the toxic nested replies to the total number of nested replies.

Control Variables. When comparing the behavior of root authors in CTR and CNR, we controlled for features that might have influenced their behavior in conversations (i.e., confounding factors). In particular, we controlled for features that capture the structure of conversations, the relationship between root author and toxicity instigator, and users' *activity*, *visibility*, and *identifiability* on Twitter. **Online activity** includes *num_friends*, *num_tweets* and *account_age* (in years) as numeric variables. If a user posts many tweets then they might also tend to engage in conversations more, even if they receive toxic replies, or if a user is more established, i.e., been using Twitter for more years, then they might perceive toxicity and react to it differently. **Online visibility** includes *num_followers*, and *listed_counts* as numeric variables and *verified* as a binary variable. While other works (ElSherief et al. 2018b) have shown that online visibility and receiving hate are related, they might also influence users' reactions toward toxicity. For example, verified users or a user with more followers might get less negatively affected by receiving a few toxic replies, compared to a user who is not verified or has a few followers. **Identifiability** represent features from the user profiles that can help identify a user, including *description_length* (in chars) as a numeric variable, and *has_URL* and *has_location* as binary variables. We also captured *has_image* but noticed that all accounts are with profile images and therefore discarded this variable. We argue that it might be that more identifiable users are more likely to get negatively affected when they are under attack by others compared to anonymous users. Also, other works have shown that anonymous accounts tend to show abusive behavior more than others (Schlesinger et al. 2017; Correa et al. 2015; Zhang and Kizilcec 2014). **Emotions before a toxic reply**. Furthermore, we believe that if the users already expressed a certain emotion before the toxic reply, it might impact the way they respond to it. For example, if the user is already angry and receives a toxic comment, the user could get even more angry and respond in a toxic way. Therefore, we included the following control variables in the analysis: *anger_before*: a numeric variable representing the average root author's anger before a toxic reply occurs in conversation; *anxiety_before* representing a numeric variable indicating the average root author's anxiety before a toxic reply occurs in a conversation, and *sadness_before*: a numeric variable representing the average root author's sadness before a toxic reply occurs in a conversation. Once again, such variables are only computed for CTR and were used as control variables in H6-H8. Also, they were computed by chronologically ordering the tweets within the conversation and calculating averages before the toxic reply occurred. Also, we used *root_anxiety*, *root_anger*, and *root_sadness* indicating emotions of the root tweet in models testing H1-H5, as users might show different behaviors depending on the emotions expressed in their root tweet. **Conversation structure** includes *size*, *width*, and *depth* as numeric variables. These features can play a role in how users respond to toxicity, e.g., a toxic reply buried in a nested conversation might not have the same negativity level as having one in a short conversation. We controlled for **started_toxic**, a binary variable that indicates whether the root author initiates toxicity by posting a toxic root tweet or a first toxic reply. Such authors could respond differently as higher toxicity of the root leads to higher toxicity in the conversation (Salehabadi et al. 2022). Finally, 19 **topics** (Antypas et al. 2022) were used as control variables as they can affect users' responses. For example, political or gaming topics might trigger longer discussions (Hilvert-Bruce and Neill 2020; Bor and Petersen 2022).

Behavioral and Emotional Characteristics

We provide descriptive statistics about conversations, author accounts, and user's behavioral and emotional reactions.

Conversations' Characteristics: Table 1 compares the characteristics of CTR and CNR. Since the distribution of features is not normal, we present min, mean, median, and max values. It shows that the prevalence of toxic direct replies ($Mean = 0.4$) is higher compared to nested replies ($Mean = 0.08$). The Mean number of toxicity instiga-

	CTR/ CNR			
	Min	Median	Mean	Max
size	2/2	7/3	17.88/5.5	1,689/1,842
width	1/1	3/1	9.9/2.8	1,688/1,825
depth	2/2	4/3	4.8/3.14	198/88
#users	2/2	4/2	11.46/3.76	810/1,818
direct tox.	0/0	0.25/0	0.4/0	1/0
nested tox.	0/0	0/0	0.08/0	1/0
#toxicity	1/0	1/0	1.5/0	144/0
# conversations			7,205/72,594	

Table 1: Characteristics of conversations in our dataset.

	Root authors of CTR/CNR			
	Min	Median	Mean	Max
#root_author replies	0/0	1/0	2.24/1.06	313/504
#root_author toxic_replies	0/0	0/0	0.17/0.03	17/8
unfollow	0/0	0/0	0.04/0.09	1/1
anx._before	0/NA	0/NA	0.17/NA	25/NA
sad._before	0/NA	0/NA	0.46/NA	25/NA
ang._before	0/NA	0/NA	0.31/NA	33/NA
anx._after	0/NA	0/NA	0.11/NA	33/NA
sad._after	0/NA	0/NA	0.77/NA	80/NA
ang._after	0/NA	0/NA	0.26/NA	50/NA
#followers	0/0	$1K$/722	$79/17K$	$54/36M$
#friends	0/0	594/533	$2/1K$	$1/2M$
#tweets	1/1	$13/10K$	$38/31K$	$1/4M$
verified	0/0	0/0	0.08/0.04	1/1
desc._len.	0/0	77/73	81.2/78.3	183/199
listed_count	0/0	8/4	359.6/88.5	$210/129K$
has_url	0/0	0/0	0.49/0.47	1/1
has_location	0/0	1/1	0.8/0.8	1/1
acc._age	0/0	3/3	4.6/4.7	15/15
#users	7,205/72,594			

Table 2: Comparing users who received toxic replies and other root authors' behaviors and account characteristics.

tors in CTR is 1.5 while the maximum number is 144. The min, median, mean, and max values for conversation size in the whole dataset are 2, 3, 6.6, and 1,842, respectively. These stats show that most conversations are small, however, the size of 8,872 (11.1%), 291 (0.4%), and 25 (0.03%) of conversations is more than 10, 100, and 500. We also see that size, width, and depth values are higher for CTR, e.g., the median and mean of size for CTR are about 7 and 17.88, while those for CNR are about 3 and 5.5, aligning with the previous literature suggesting that toxic conversations are wider, deeper, and larger than non-toxic conversations (Saveski, Roy, and Roy 2021). Similarly, on average more users participate in CTR (Mean is about 11.46) compared to CNR (Mean is about 3.76). We ran Mann-Whitney U tests for all the variables and found that there are significant differences ($p < 0.05$) between these characteristics in CTR and CNR. For brevity, we do not present these results.

Root Authors' Behaviors: Table 2 shows the statistics about the user's reactions. Interestingly, Table 2 shows that the maximum *#root_author_replies* is higher in CTR compared to CNR (504 vs. 313), but the mean of *#root_author_replies* in CTR is approximately twice times the mean of *#root_author_replies* in CNR (1.06 vs. 2.24). In addition, root authors of CTR are engaged in a larger percentage of conversations compared to root authors of CNR (60.7% vs. 48.04%). Moreover, the percent of CTR where users responded with at least one toxic reply is 11.7%, while in CNR, that number is 2.7%. Furthermore, the percentage of all root authors' toxic replies that were immediate toxic replies to toxic comments is 8.13% (the toxic children of the toxic node in the conversation tree). The percent of all children nodes where the root author directly responded to a toxic tweet which is toxic in nature is 22.18%. Additionally, in 11.6% of CTR, root authors unfollowed at least one user who posted toxic comments on their posts. Mann-Whitney U tests showed that there is a statistically significant difference in engagement and responding in a toxic way between root authors of CTR and CNR ($p < 0.05$).

Root Authors' Account Characteristics: Table 2 shows that our dataset contains a higher percentage of *verified* accounts among the root authors of CTR (0.08) compared to authors of CNR (0.04), consistent with the prior studies (ElSherief et al. 2018b). In addition, the Mean values of followers, friends, tweets, and listed counts of CTR root authors are higher compared to these characteristics of CNR root authors, again consistent with prior studies (ElSherief et al. 2018b), indicating that CTR authors might be more active and visible accounts compared to CNR root authors. Moreover, a higher percentage of CTR authors have URLs specified on their profiles, and their Mean account age might indicate that their accounts are younger compared to CNR authors. We ran a Mann-Whitney test to compare the distributions of all account characteristics and found a significant difference ($p < 0.05$) between all account characteristics among CTR root authors and CNR root authors.

Root Authors' Emotions: Table 2 demonstrates that the mean *anxiety_before* and *anger_before* of root authors of CTR is higher than the mean *anxiety_after* (0.17 vs. 0.11) and *anger_after* (0.31 vs. 0.26). Interestingly, the maximum values of both *anger_after* ($Max = 50$) and *anxiety_after* ($Max = 33.3$) are higher than maximum values of *anger_before* ($Max = 33.33$) and *anxiety_before* ($Max = 25$), potentially indicating that such emotions are not likely to get elevated when receiving toxicity, but there might be extreme cases when they do. However, in the case of sadness, both the mean and max value are higher for *sadness_after* than *sadness_before*. Finally, sadness is the most prevalent emotion expressed by root authors of CTR. Mann-Whitney U tests showed that there is a statistically significant difference of distributions in the emotion of sadness of root authors before and after the first toxic reply ($p < 0.05$) ($Mean_{before} = 0.46$ vs. $Mean_{after} = 0.77$), while the differences between anger and anxiety before and after the first toxic reply do not show significance ($p > 0.05$).

Hypotheses Testing

This study investigates whether root authors of CTR are more or less likely to engage in certain behaviors. This is a causal question. While we might not achieve the ideal identification of causal relationships, we use techniques borrowed from causal inference literature instead of simple correlational analyses (Olteanu, Varol, and Kiciman 2017). We

cannot argue that we examine causal effects because other factors, such as users' personality traits, mindset, previous experiences, etc., could affect users' behaviors in conversations with toxic replies. However, by using the *propensity score matching*, it is reasonable to use the label "causal association" instead of causal effect to describe findings from an observational study (Hammerton and Munafò 2021).

Propensity score matching balances the treatment and control groups (root authors of CTR and root authors of CNR, respectively) on the confounding factors to make them comparable allowing to draw conclusions about the causal impact of treatment using observational data (Yao et al. 2017; Olteanu, Varol, and Kiciman 2017). We used *propensity score matching* based on users' characteristics and conversation topics to obtain balanced *treatment* and *control* groups where each author has an equal propensity of being exposed to the treatment (Rosenbaum and Rubin 1984; Olteanu, Varol, and Kiciman 2017). To calculate propensity scores, we used logistic regression with the binary dependent variable indicating whether the conversation contains toxic replies, while the root authors' account characteristics and conversation topics were considered predictors.

Our hypotheses are tested on three datasets with sizes 79,799, 266,937, and 3,408, which represent authors who engaged in conversations, the unique repliers in all the conversations, and root authors of CTR who engaged after the first toxic reply, respectively. After employing the propensity matching algorithm on the first two datasets (the third dataset includes CTR only), we obtained balanced datasets with sizes 14,410, and 147,112, respectively. We evaluated the balance between matched samples by examining the standardized mean differences (SMD) (Zhang et al. 2019). None of the variables showed an SMD above 0.1, suggesting that there are no significant differences in variables between the two groups, indicating good matching (Zhang et al. 2019). Finally, we used multivariate regression on these balanced treatment and control groups, to examine the relationship between receiving toxic replies and users' behaviors while considering the confounding factors. For numeric dependent variables, we used Poisson multivariate regression models, and for binary dependent variables, we employed logistic regressions. We applied Bonferroni correction (Armstrong 2014) to adjust p-values for multiple hypotheses testing. Dividing the p-value of 0.05 by the number of hypotheses tested on the same dataset yields a $p = 0.01$ (H1-H4) and $p = 0.02$ (H6-H8). The p-value remained at 0.05 for H5. The results are presented in Table 3; note that topics' coefficients were omitted due to brevity. Also, the coefficients of 0.0 are very small numbers close to 0.

H1: Root authors of CTR are more likely to engage in conversations. H1 aims to test whether CTR authors tend to *negotiate* their points of view by engaging in conversations. We measured engagement using *#root_author_replies* as the dependent variable. The results of the Poisson regression are shown in Table 3 (column H1). The relationship between *direct_toxicity* and *#root_author_replies* is negatively significant ($p < 0.0001$), meaning that root authors of CTR are less likely to engage in the conversation in case it contains toxic direct replies. However, there is a positive significant causal association between the *nested_toxicity* and the *#root_author_replies* ($p < 0.0001$). Such findings suggest that root authors tend to engage more if a big discussion develops including a larger percent of toxic nested replies, while they are less likely to engage with direct toxic replies to their root tweet. This is inline with findings of the previous literature that users are likely to engage more with replies that receive multiple comments (Yousefi et al. 2023a). Furthermore, conversation structure is also associated with user engagement. Table 3 suggests that users tend to engage more in deeper and wider conversations ($p < 0.0001$), again indicating that users engage in toxic conversations when a bigger discussion occurs. Interestingly, *verified* accounts tend to engage less in conversations compared to other users ($p < 0.0001$), which can be due to the great number of replies they receive on their posts, or that they believe *any attention is good attention* (Palomino and Varma 2020). The incidence rate ratio (IRR) indicates that the account being verified decreases the probability of engagement by 51%. Young and more identifiable accounts that provide a location and longer description tend to engage more in conversations compared to other users ($p < 0.0001$), which might indicate that such users care more about their account reputability. *In conclusion, this analysis supports H1 as it shows that there is a causal association between receiving toxic replies and root authors' engagement in the conversations.*

H2: Root authors of CTR are more likely to engage in conversations if receiving toxic replies from a larger number of toxicity instigators. Users might not react the same way if toxic replies are posted by a single user vs. many users. We defined our independent variable, *toxicity_instigators*, as the ratio of unique users who posted toxic replies to the total number of unique users in the conversation. We avoided using this variable together with other independent variables in the models due to multicollinearity, as a number of toxicity instigators is correlated with the number of toxic replies. The results of the Poisson model suggest that users are more likely to engage in the conversation if toxicity is coming from a larger number of users ($p < 0.0001$), and *therefore H2 is supported*. IRR suggests that for each point increase of *toxicity_instigators*, the expected *#root_author_replies* increases by 45.5% ($p < 0.0001$).

H3: Root authors of CTR are more likely to respond back in a toxic way. This hypothesis compares the engagement of CTR and CNR root authors in a toxic way. Even though root authors of CNR did not receive toxic replies, they can still respond to others in a toxic way. The results of regression analysis on *#root_author_toxic_replies* as dependent variable, illustrated in Table 3 (column H3), show a positive significant correlation between *#root_author_toxic_replies* and the *nested_toxicity* ($p < 0.0001$). That is, the more toxic nested replies posted by other users are involved in the conversation, root authors of CTR are more likely to respond back in a toxic way compared to authors of CNR, and therefore H3 is supported. IRR revealed that for each unit of increase in *nested_toxicity*, the *#root_author_toxic_replies* increases by a factor of 3.78 (278%) demonstrating a substantial impact on users' behavior. Interestingly, we do not see the same

	#root_author_replies		#root_author_toxic_replies		anxiety_after	anger_after	sadness_after
	H1 *partially supported*	H2 *supported*	H3 *partially supported*	H4 *supported*	H6 *rejected*	H7 *partially supported*	H8 *partially supported*
direct_toxicity	-0.6^{***}		-0.2^{**} (0.1)		-0.2 (0.2)	0.0 (0.1)	0.6^{***} (0.1)
nested_toxicity	0.5^{***}		1.3^{***} (0.1)		-1.6^{**} (0.6)	0.7^{*} (0.3)	-0.5^{*} (0.2)
toxicity_inst.		0.4^{***}		1.6^{***} (0.1)			
width	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	-0.0	-0.0^{***}	-0.0
depth	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0	-0.0	-0.0^{**}
started_toxic1	0.5^{***}	0.4^{***}	1.8^{***}	1.7^{***}	-1.1^{***} (0.2)	0.3^{**} (0.1)	0.2^{***}
followers	0.0^{**}	0.0^{**}	-0.0	-0.0	0.0	0.0	-0.0
#friends	0.0	0.0	-0.0^{*}	-0.0	0.0	-0.0^{**}	-0.0^{**}
#tweets	0.0	0.0^{*}	0.0	0.0	0.0	0.0	0.0^{***}
listed_counts	-0.0^{***}	-0.0^{***}	-0.0	-0.0	-0.0	0.0^{*}	-0.0^{*}
desc._length	0.0^{***}	0.0^{***}	0.0	0.0	-0.0	0.0^{**}	-0.0^{***}
verified1	-0.7^{***}	-0.6^{***}	-0.9^{**} (0.2)	-0.7^{**} (0.2)	0.1 (0.5)	-0.5 (0.3)	0.0 (0.3)
account_age	-0.0^{***}	-0.0^{***}	-0.0^{***}	-0.0^{***}	0.0	-0.0	-0.0^{***}
has_location1	0.1^{***}	0.1^{***}	0.0 (0.1)	0.0 (0.1)	0.9^{***} (0.2)	0.3^{**} (0.1)	0.7^{***} (0.1)
has_url1	-0.0^{**}	-0.0	0.0	0.1	-0.4^{**} (0.1)	-0.2 (0.1)	0.1^{**}
anxiety_before					0.1^{***}		
anger_before						0.1^{***}	
sadness_before							0.1^{***}
root_sadness	-0.0^{**}	-0.0^{**}	-0.0	-0.0			
Observations	14,410	14,410	14,410	14,410	3,408	3,408	3,408
Log Likelihood	$-30{,}968.6$	$-31{,}382.6$	$-5{,}452.6$	$-5{,}386.3$	$-1{,}531.2$	$-2{,}977.3$	$-6{,}984.4$
Note:							*p<0.02; **p<0.01; ***p<1e-04

Table 3: Results of hypotheses testing. The standard errors with 0.0 values are not listed in the table. Non-significant variables are omitted due to brevity.

trend for *direct_toxicity*, which might be because in threads, users tend to argue with each other by posting nested replies back and forth, increasing the number of toxic replies, while it does not necessarily happen with direct replies. Furthermore, root authors who initiated toxicity are more likely to engage in a conversation in a toxic way compared to other users ($p < 0.0001$). Similarly as in H1, young accounts are more likely to respond in a toxic way ($p < 0.001$), while the IRR shows that the account being *verified* decreases the expected number of *#root_author_toxic_replies* by 58% ($p < 0.01$). Also, users tend to respond in a toxic way in deeper conversations ($p < 0.0001$).

H4: Root authors of CTR are more likely to respond back in a toxic way if receiving toxic replies from a larger number of toxicity instigators. We used *toxicity_instigators* as the independent variable in Poisson regression model. *The results suggest that users are more likely to respond in a toxic way if toxicity is posted by a larger number of instigators ($p < 0.0001$), supporting H4.*

H5: Root authors of CTR are more likely to unfollow toxicity instigators compared to authors of CNR. Unfollowing can be an example of employing a *countermeasure*, as it can reduce the probability of receiving toxic comments on future posts. Running a logistic regression model where the binary dependent variable indicates whether the root author unfollowed the replier while the independent variable is the replier's number of toxic replies. **Additional control variables:** We used two sets of control variables in this model, including characteristics of the root author and characteristics of the replier. CTR root authors might perceive toxic replies differently depending on who the repliers are. For example, they might be more likely to unfollow instigators who have no identifiable accounts. *The results showed that there is a statistically significant positive correlation between unfollowing the replier and the replier's number of toxic replies within a conversation ($p < 0.001$), supporting H5.* The significance of control variables suggests that root authors are more likely to unfollow *identifiable* repliers that provide URL ($p < 0.001$) and location ($p < 0.001$). This could be because root authors feel more affected when receiving toxic replies from an identifiable person than from an anonymous account, making them unfollow such users.

Interestingly, none of the models yield a significant correlation between users' emotions and their behaviors, besides a negative correlation between the sadness of the root tweet and engagement ($p < 0.01$), indicating that users expressing sadness tend to engage less in the conversation. However, such findings are unexpected, as we expected that users expressing certain emotions might react more severely to toxic content and potentially elevate the conflict.

H6: A larger amount of toxicity will likely increase users' anxiety. According to the results presented in Table 3 (H6), there exists a negative significant relationship between nested_toxicity and anxiety_after ($p < 0.01$), while the association between direct_toxicity and anxiety_after does not show statistical significance *rejecting H6*. However, users who were already feeling anxious before the first toxic reply tend to express more anxiety after ($p < 0.0001$).

H7: A larger amount of toxicity will likely increase users' anger. As demonstrated in Table 3 (H7), the relationship between direct_toxicity and anger of users after the first toxic reply is not statistically significant ($p > 0.02$) while

	Unfollowed
num_toxic_replies	0.1^{***} (0.0)
replier num_followers	-0.0^{**} (0.0)
replier num_tweets	0.0^{***} (0.0)
replier listed_counts	0.0^{*} (0.0)
replier account_age	-0.0^{***} (0.0)
replier has_location	0.4^{***} (0.0)
replier has_url	0.3^{***} (0.0)
author num_followers	-0.0^{***} (0.0)
author num_friends	0.0^{***} (0.0)
author num_tweets	0.0^{***} (0.0)
author listed_counts	-0.0^{**} (0.0)
author description_length	-0.0^{***} (0.0)
author verified True	-0.7^{***} (0.1)
author account_age	-0.1^{***} (0.0)
author has_location	0.2^{***} (0.0)
author has_url	-0.1^{***} (0.0)
Observations	147,112
Note:	*p<0.05; **p<0.01; ***p<0.001

Table 4: Results of testing H5. Non-significant variables are omitted due to brevity.

Dependent vars.	Accu.	Precision	Recall	F1-Score
M1	0.94	0.93	0.95	0.94
M2	0.82	0.81	0.83	0.82
M3	0.92	0.87	0.86	0.93

Table 5: Performance of classification models.

there is a statistically significant positive relationship between nested_toxicity and anger_after ($p < 0.02$), *partially supporting H7*. Such findings potentially indicate that users might get angry if a bigger discussion develops where they receive a large portion of toxic nested replies. Similarly as in testing H6, we noticed that the association between the average anger of users before the first toxic reply and average anger after the first toxic reply is positive and significant ($p < 0.0001$), meaning that users who were angry are likely to express this emotion more after receiving toxicity.

H8: A larger amount of toxicity will likely increase users' sadness. As shown in Table 3 (H8), there is a statistically significant positive relationship between direct_toxicity and sadness of users after the first toxic reply ($p < 0.0001$). In other words, if the user receives more toxic replies on their main posts, such users tend to express more sadness. However, the model also shows a negative statistically significant relationship between nested_toxicity and sadness_after ($p < 0.02$), revealing that users tend to express less sadness if they receive a larger portion of toxic nested replies. Additional research might be required to get a better understanding of the reasons behind these findings, but we currently believe that users potentially experience direct toxic replies more personally compared to nested replies, leading to them expressing more sadness. Also, the model reveals that users showing more sadness before the first toxic reply tend to get even more sad after the first toxic reply occurs ($p < 0.0001$). Such findings signify that the amount of toxicity is likely to impact the amount of sadness users convey, *partially supporting H8*.

Prediction Models

Our findings indicate the significance of several attributes related to the engagement, profile characteristics, topics discussed as well as the overall structure of the toxic conversations which contribute to the behavior of the authors after receiving toxic comments on their tweets. We utilize these features to construct classification models to predict how the authors will react to the conversation (when has turned/is about to turn toxic), specifically, if they further engage in such conversations **(M1)** and if they do such in a toxic manner **(M2)**, or unfollow the toxic instigators entirely **(M3)**. Our models were trained on a system running on an Intel Xeon W Processor with 184GB of RAM and 4x NVIDIA A4000 GPUs. We utilize the same variables for the respective models, as highlighted in previous sections.

We trained *three* random forest classification models on 79,799 conversations, where 7,205 conversations were labeled as CTR and 72,594 as CNR using a threshold of 0.6 for *Toxicity* attribute score of Perspective API. To evaluate the classifier in the realistic scenario, we considered the whole unbalanced dataset, i.e., before applying the propensity matching algorithm. To balance the dataset, we utilized SMOTE (Blagus and Lusa 2013) which uses the k-nearest neighbors algorithm to generate synthetic samples by interpolating between existing minority class samples. Also, to further avoid over-fitting, we ran our models through a 10-fold cross-validation with 8 folds for training and 2 folds for testing in each iteration. For each model, we conducted a feature importance analysis and considered all, as well as the top N features (where N=5,10) that contributed the most to the respective classification for training. We finalize N=10 since it provides the best performance across all three models (as highlighted in Table 5) with accuracy scores of 0.94, 0.82, and 0.92 for Models 1, 2, and 3 respectively. Nonetheless, developing automated approaches can significantly aid in the early identification of behavior after being exposed to toxic comments and allow moderators to employ prompt intervention strategies to enhance user safety.

Broader Impacts and Limitations

Broader Impacts: We will share our unique dataset and classifiers with the research community. The dataset includes tweet IDs only, thus, it does not contain personally identifiable information. Our findings could eventually lead to the development of personalized moderation techniques. For example, sending comforting messages or encouraging followers to stand up for users might reduce the emotional impact of toxicity. Furthermore, content moderation could flag the posts for intensive negative emotions ensuring that harmful content is addressed promptly. However, such strategies should also take into consideration usual users' online behaviors; users who usually respond in a toxic way might not prefer to receive comforting messages, while users who negotiate prefer others to stand up for their opinion. Thus, future work might include examining different responses by taking more factors into account, e.g., users' personality traits and past reactions to toxicity.

Discussion and Limitations: We relied on the Perspec-

tive API model to identify toxic tweets, and our data thus inherits its shortcomings or biases (TeBlunthuis, Hase, and Chan 2024). The data collection process has some limitations. We only considered English tweets. Since the replies were collected two days after obtaining the root tweets, some of replied could have been deleted, not allowing us to collect full conversation trees. Despite the certain bias, we believe that obtaining and analyzing the large enough sample of 79,799 conversations should have provided reliable results. Moreover, studying the difference between toxic direct and nested replies and investigating the reasons behind root authors' distinct responses to them can help uncovering the dynamics of toxic conversations. In more detail, our results show that root authors are more likely to engage and respond back in a toxic manner if receiving toxic nested replies compared to receiving toxic direct replies. While one possibility is that users tend to engage with replies that receive multiple comments, another explanations could be that toxic nested replies are less likely to be directed to root authors compared to direct replies. Therefore, root authors might engage less with toxic nested replies, as they might not be directed to them. Such explanation can also justify the results of H8, where we show that root authors are more likely to express sadness when receiving toxic *direct replies* while they are less likely to do so when receiving toxic *nested replies*. Finally, studying the timelines of the accounts could help overall understanding of users' responses, which we consider for future work.

Conclusion

In summary, this large data-driven study examines the effect of toxicity on users' online behaviors and emotions. Firstly, we investigated users' behaviors in terms of *avoidance*, *countermeasures*, *negotiation*, and *revenge*, while also considering factors such as conversation structure, user characteristics, and conversation topics. Moreover, we examined the impact of toxicity on the emotions of anger, sadness, and anxiety of users who did engage in toxic conversations after receiving toxic replies. Finally, we built a prediction model to forecast the behavioral responses of users with high accuracy. This study can help develop efficient intervention mechanisms to help mitigate the negative consequences of receiving toxic replies on social media.

References

Ak, Ş.; Özdemir, Y.; and Kuzucu, Y. 2015. Cybervictimization and cyberbullying: The mediating role of anger, don't anger me! *Computers in human behavior*, 49: 437–443.

Aleksandric, A.; Singhal, M.; Groggel, A.; and Nilizadeh, S. 2022. Understanding the Bystander Effect on Toxic Twitter Conversations. *arXiv preprint arXiv:2211.10764*.

Alhujailli, A.; Karwowski, W.; Wan, T. T.; and Hancock, P. 2020. Affective and stress consequences of cyberbullying. *Symmetry*, 12(9): 1536.

Androcec, D. 2020. Machine learning methods for toxic comment classification: a systematic review. *Acta Universitatis Sapientiae, Informatica*, 12(2): 205–216.

Antypas, D.; Ushio, A.; Camacho-Collados, J.; Silva, V.; Neves, L.; and Barbieri, F. 2022. Twitter Topic Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, 3386–3400. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Arató, N.; Zsidó, A.; Rivnyák, A.; Péley, B.; and Lábadi, B. 2022. Risk and protective factors in cyberbullying: the role of family, social support and emotion regulation. *International journal of bullying prevention*, 4(2): 160–173.

Armstrong, R. A. 2014. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5): 502–508.

Beres, N. A.; Frommel, J.; Reid, E.; Mandryk, R. L.; and Klarkowski, M. 2021. Don't you know that you're toxic: Normalization of toxicity in online gaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

Blagus, R.; and Lusa, L. 2013. SMOTE for high-dimensional class-imbalanced data. *BMC bioinformatics*, 14: 1–16.

Bor, A.; and Petersen, M. B. 2022. The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American political science review*, 116(1): 1–18.

Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 1–47.

Campbell, M.; Spears, B.; Slee, P.; Butler, D.; and Kift, S. 2012. Victims' perceptions of traditional and cyberbullying, and the psychosocial correlates of their victimisation. *Emotional and Behavioural Difficulties*, 17(3-4): 389–401.

Chong, Y. Y.; and Kwak, H. 2022. Understanding toxicity triggers on Reddit in the context of Singapore. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1383–1387.

Correa, D.; Silva, L. A.; Mondal, M.; Benevenuto, F.; and Gummadi, K. P. 2015. The many shades of anonymity: Characterizing anonymous social media content. In *Ninth International AAAI Conference on Web and Social Media*.

Duffy, B. E.; and Hund, E. 2019. Gendered Visibility on Social Media: Navigating Instagram's Authenticity Bind. *International Journal of Communication (19328036)*, 13.

ElSherief, M.; Kulkarni, V.; Nguyen, D.; Wang, W. Y.; and Belding, E. 2018a. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

ElSherief, M.; Nilizadeh, S.; Nguyen, D.; Vigna, G.; and Belding, E. 2018b. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Erişti, B.; and Akbulut, Y. 2019. Reactions to cyberbullying among high school and university students. *The Social Science Journal*, 56(1): 10–20.

Eyuboglu, M.; Eyuboglu, D.; Pala, S. C.; Oktar, D.; Demirtas, Z.; Arslantas, D.; and Unsal, A. 2021. Traditional school bullying and cyberbullying: Prevalence, the effect on mental health problems and self-harm behavior. *Psychiatry research*, 297: 113730.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2024-04-01.

Fredericks, B.; and Bradfield, A. 2021. 'Waiting with Bated Breath': Navigating the Monstrous World of Online Racism. *M/C Journal*, 24(5).

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gianesini, G.; and Brighi, A. 2015. Cyberbullying in the era of digital relationships: The unique role of resilience and emotion regulation on adolescents' adjustment. In *Technology and youth: Growing up in a digital world*, 1–46. Emerald Group Publishing Limited.

Giumetti, G. W.; and Kowalski, R. M. 2022. Cyberbullying via social media and well-being. *Current Opinion in Psychology*, 101314.

Habib, H.; and Nithyanand, R. 2022. Exploring the Magnitude and Effects of Media Influence on Reddit Moderation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 275–286.

Hammerton, G.; and Munafò, M. R. 2021. Causal inference with observational data: the need for triangulation of evidence. *Psychological medicine*, 51(4): 563–578.

Hannan, J. 2018. Trolling ourselves to death? Social media and post-truth politics. *European Journal of Communication*, 33(2): 214–226.

Hellfeldt, K.; López-Romero, L.; and Andershed, H. 2020. Cyberbullying and Psychological Well-being in Young Adolescence: The Potential Protective Mediation Effects of Social Support from Family, Friends, and Teachers. *International Journal of Environmental Research and Public Health*, 17(1).

Hilvert-Bruce, Z.; and Neill, J. T. 2020. I'm just trolling: The role of normative beliefs in aggressive behaviour in online gaming. *Computers in Human Behavior*, 102: 303–311.

Hine, G. E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and its Effects on the Web.

Horta Ribeiro, M.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–24.

Huang, F.; Kwak, H.; and An, J. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, 294–297.

Kahn, J. H.; Tobin, R. M.; Massey, A. E.; and Anderson, J. A. 2007. Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American journal of psychology*, 120(2): 263–286.

Kowalski, R. M.; Giumetti, G. W.; Schroeder, A. N.; and Lattanner, M. R. 2014. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin*, 140(4): 1073.

Kumar, D.; Hancock, J.; Thomas, K.; and Durumeric, Z. 2022. Understanding Longitudinal Behaviors of Toxic Accounts on Reddit. *arXiv preprint arXiv:2209.02533*.

Kumarswamy, N.; Singhal, M.; and Nilizadeh, S. 2023. Impact of Stricter Content Moderation on Parler's Users' Discourse. *arXiv preprint arXiv:2310.08844*.

Kvålseth, T. O. 1989. Note on Cohen's kappa. *Psychological reports*, 65(1): 223–226.

Kwan, I.; Dickson, K.; Richardson, M.; MacDowall, W.; Burchett, H.; Stansfield, C.; Brunton, G.; Sutcliffe, K.; and Thomas, J. 2020. Cyberbullying and children and young people's mental health: a systematic map of systematic reviews. *Cyberpsychology, Behavior, and Social Networking*, 23(2): 72–82.

Lees, A.; Tran, V. Q.; Tay, Y.; Sorensen, J.; Gupta, J.; Metzler, D.; and Vasserman, L. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3197–3207.

Leiter, C.; Zhang, R.; Chen, Y.; Belouadi, J.; Larionov, D.; Fresen, V.; and Eger, S. 2024. Chatgpt: A meta-analysis after 2.5 months. *Machine Learning with Applications*, 100541.

Lyu, S.; Ren, X.; Du, Y.; and Zhao, N. 2023. Detecting depression of Chinese microblog users via text analysis: Combining Linguistic Inquiry Word Count (LIWC) with culture and suicide related lexicons. *Frontiers in psychiatry*, 14: 1121583.

Maarouf, A.; Pröllochs, N.; and Feuerriegel, S. 2022. The Virality of Hate Speech on Social Media. *arXiv preprint arXiv:2210.13770*.

Mameli, C.; Menabò, L.; Brighi, A.; Menin, D.; Culbert, C.; Hamilton, J.; Scheithauer, H.; Smith, P. K.; Völlink, T.; Willems, R. A.; et al. 2022. Stay safe and strong: characteristics, roles and emotions of student-produced comics related to cyberbullying. *International journal of environmental research and public health*, 19(14): 8776.

Mathew, B.; Illendula, A.; Saha, P.; Sarkar, S.; Goyal, P.; and Mukherjee, A. 2020. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–24.

Olteanu, A.; Castillo, C.; Boy, J.; and Varshney, K. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Olteanu, A.; Varol, O.; and Kiciman, E. 2017. Distilling the Outcomes of Personal Experiences: A Propensity-Scored Analysis of Social Media. In *Proceedings of the 2017*

ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17, 370–386. New York, NY, USA: Association for Computing Machinery. ISBN 9781450343350.

Ortega, R.; Elipe, P.; Mora-Merchán, J. A.; Genta, M. L.; Brighi, A.; Guarini, A.; Smith, P. K.; Thompson, F.; and Tippett, N. 2012. The emotional impact of bullying and cyberbullying on victims: A European cross-national study. *Aggressive behavior*, 38(5): 342–356.

Palomino, M. A.; and Varma, A. P. 2020. Any Publicity is Good Publicity: Positive, Negative and Neutral Tweets Can All Become Trends. In *2020 39th International Conference of the Chilean Computer Science Society (SCCC)*, 1–8. IEEE.

Parent, M. C.; Gobble, T. D.; and Rochlen, A. 2019. Social media behavior, toxic masculinity, and depression. *Psychology of Men & Masculinities*, 20(3): 277.

Raskauskas, J.; and Stoltz, A. D. 2007. Involvement in traditional and electronic bullying among adolescents. *Developmental psychology*, 43(3): 564.

Ribeiro, M.; Calais, P.; Santos, Y.; Almeida, V.; and Meira Jr, W. 2018. Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Rosenbaum, P. R.; and Rubin, D. B. 1984. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387): 516–524.

Salehabadi, N. 2019. *The impact of toxic replies on Twitter conversations*. The University of Texas at Arlington.

Salehabadi, N.; Groggel, A.; Singhal, M.; Roy, S. S.; and Nilizadeh, S. 2022. User Engagement and the Toxicity of Tweets. *arXiv preprint arXiv:2211.03856*.

Saveski, M.; Roy, B.; and Roy, D. 2021. The Structure of Toxic Conversations on Twitter. In *Proceedings of the Web Conference 2021*, 1086–1097.

Schlesinger, A.; Chandrasekharan, E.; Masden, C. A.; Bruckman, A. S.; Edwards, W. K.; and Grinter, R. E. 2017. Situated anonymity: Impacts of anonymity, ephemerality, and hyper-locality on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, 6912–6924.

Silva, L.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2016. Analyzing the targets of hate in online social media. In *Tenth International AAAI Conference on Web and Social Media*.

Sun, Q.; and Shen, C. 2021. Who would respond to A troll? A social network analysis of reactions to trolls in online communities. *Computers in Human Behavior*, 121: 106786.

Tahmasbi, F.; Schild, L.; Ling, C.; Blackburn, J.; Stringhini, G.; Zhang, Y.; and Zannettou, S. 2021. "Go eat a bat, Chang!": On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. In *Proceedings of the web conference 2021*, 1122–1133.

TeBlunthuis, N.; Hase, V.; and Chan, C.-H. 2024. Misclassification in Automated Content Analysis Causes Bias in Regression. Can We Fix It? Yes We Can! *Communication Methods and Measures*, 1–22.

Twarc. 2020. Collect Twitter Data with Twarc! https://scholarslab.github.io/learn-twarc/. Accessed: 2024-04-01.

Twitter. 2022. Twitter API. Accessed: 2024-04-01.

Varjas, K.; Talley, J.; Meyers, J.; Parris, L.; and Cutts, H. 2010. High school students' perceptions of motivations for cyberbullying: An exploratory study. *Western Journal of Emergency Medicine*, 11(3): 269.

Veletsianos, G.; Houlden, S.; Hodson, J.; and Gosse, C. 2018. Women scholars' experiences with online harassment and abuse: Self-protection, resistance, acceptance, and self-blame. *New Media & Society*, 20(12): 4689–4708.

Wong, R. Y. M.; Cheung, C. M.; Xiao, B.; and Thatcher, J. B. 2021. Standing up or standing by: Understanding bystanders' proactive reporting responses to social media harassment. *Information Systems Research*, 32(2): 561–581.

Wright, V. H.; Burnham, J. J.; Inman, C. T.; and Ogorchock, H. N. 2009. Cyberbullying: Using Virtual Scenarios to Educate and Raise Awareness. *Journal of Computing in Teacher Education*, 26(1): 35–42.

Yang, X.; Ye, H. J.; and Wang, X. 2021. Social media use and work efficiency: Insights from the theory of communication visibility. *Information & Management*, 58(4): 103462.

Yao, X. I.; Wang, X.; Speicher, P. J.; Hwang, E. S.; Cheng, P.; Harpole, D. H.; Berry, M. F.; Schrag, D.; and Pang, H. H. 2017. Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. *JNCI: Journal of the National Cancer Institute*, 109(8): djw323.

Yousefi, N.; Noor, N. B.; Spann, B.; and Agarwal, N. 2023a. Examining Toxicity's Impact on Reddit Conversations. In *International Conference on Complex Networks and Their Applications*, 401–411. Springer.

Yousefi, N.; Noor, N. B.; Spann, B.; and Agarwal, N. 2023b. Towards Developing a Measure to Assess Contagiousness of Toxic Tweets.

Zannettou, S.; ElSherief, M.; Belding, E.; Nilizadeh, S.; and Stringhini, G. 2020a. Measuring and characterizing hate speech on news websites. In *Proceedings of the 12th ACM Conference on Web Science*, 125–134.

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2020b. A quantitative approach to understanding online antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 786–797.

Zhang, K.; and Kizilcec, R. F. 2014. Anonymity in social media: Effects of content controversiality and social endorsement on sharing behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Zhang, Z.; Kim, H. J.; Lonjon, G.; Zhu, Y.; et al. 2019. Balance diagnostics after propensity score matching. *Annals of translational medicine*, 7(1).

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, in the Introduction.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes, in the Broader Impacts and Limitations section.
   (f) Did you discuss any potential negative societal impacts of your work? NA
   (g) Did you discuss any potential misuse of your work? NA
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, as we only share Tweet IDs of the dataset, it does not contain identifiable information.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, in the Analysis and Results section.
   (b) Have you provided justifications for all theoretical results? Yes, in the Analysis and Results section.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, in the Related Work section.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, explanations can be found in the Analysis and Results section.
   (e) Did you address potential biases or limitations in your theoretical framework? No, but we do acknowledge the limitation in the Broader Impacts and Limitations.
   (f) Have you related your theoretical results to the existing literature in social science? Yes, in the Introduction and Related work sections.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, in the Broader Impacts and Limitations section.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we provided a URL to Zenodo repository in the Introduction.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in the Prediction Models section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, in the Prediction Models section.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, in the Prediction Models section.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, in the Prediction Models section.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, in Prediction Models Section.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes, all the data sources can be found in the References.
   (b) Did you mention the license of the assets? No, as we are using all publically available assets.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, we provided the URL to the Zenodo repository in the Introduction
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, as we collected public data from social media.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we mentioned it in the Broader Impacts and Limitations section.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? We state in the Introduction that our dataset sticks to FAIR guidelines.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Zenodo repository includes a datasheet answering the required questions.
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA