

The Power of Social Norms: How Initial Responses to Toxicity Shape Conversations on Twitter

Ana Aleksandric^{1*}, Mohit Singhal^{2*}, Anne Groggel³, Shirin Nilizadeh⁴

¹Florida Atlantic University

²Northeastern University

³St. Mary's College of Maryland

⁴The University of Texas at Arlington

aaleksandric@fau.edu, m.singhal@northeastern.edu, agroggel@smcm.edu, shirin.nilizadeh@uta.edu

Abstract

Online harassment and abusive language continue to be a growing concern on social media platforms. In this study, we explore the power of group dynamics to shape the toxicity of Twitter conversations. First, we examine how the presence of others in a conversation can potentially diffuse Twitter users' responsibility to address a toxic reply. Second, we examine whether the toxicity of the first direct reply to a toxic tweet in conversations establishes group norms for subsequent replies. By doing so, we outline users participating in the conversation before the first toxic reply and the tone of initial responses to a toxic reply as explanatory factors that affect whether others feel uninhibited to post their own abusive or derogatory replies. We test this premise by analyzing a random sample of more than 187K tweets belonging to $\sim 9K$ conversations. This analysis of group dynamics is motivated by a larger body of scholarship on contagion of antisocial behavior and the power of establishing social norms that maintain rather than sanction toxicity. We find evidence that an increased number of users participating in the conversation before receiving a toxic tweet is negatively associated with the number of users who responded to the toxic reply in a non-toxic way. Furthermore, posting a toxic reply immediately after a toxic comment is negatively associated with users posting non-toxic replies and Twitter conversations becoming increasingly toxic. We argue that understanding how social media users respond to uncivil comments or abusive language reveals social norms as powerful social cues that can shape human behavior online.

Introduction

Content Warning: This study analyzes group dynamics in online toxic conversations. This paper provides demonstrative examples of user content that might include profane and hateful content that may be found offensive by some.

Social media and online communities allow individuals to freely express opinions, engage in interpersonal communication, and learn about new trends and news stories. Platforms such as Twitter (now known as X) hold promise for users to engage in rich and vibrant conversations with others from various backgrounds and cultures. Nevertheless,

*Work done at The University of Texas at Arlington.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

these platforms also serve as spaces for uncivil behavior. In particular, toxicity as explicit language, derogatory, aggressive, or disrespectful content has become endemic on online platforms (Anderson et al. 2016; Dutton 1996; Papacharissi 2002; Hwang et al. 2008; Zannettou et al. 2020). There is a growing concern regarding the prevalence of incivility over social media platforms and its impact on online communities (Rost, Stahel, and Frey 2016; Duggan 2014). How groups react to divisive behavior can reflect broader social norms online, whether they spread negative behaviors, call out racist or sexist behavior, or ignore toxic behaviors (Binns 2012).

Toxic behavior frequently occurs in the presence of other users whose actions can influence the dynamics of a social situation. For instance, they may actively engage with the perpetrator's behavior by posting toxic replies, endeavor to counteract toxicity by confronting such behavior or contributing positively to the conversation, or simply observe the interaction unfold (Aleksandric et al. 2024). Social norm theory provides a useful lens for understanding communication behavior on social media, particularly the persistence of toxicity. Norms reflect shared expectations about appropriate behavior, shaping the cultural "dos and don'ts" that guide interactions and regulate how individuals engage with content and other users online. In online communities, users are going to be guided both by their internalized discomfort with breaking social norms, their perceptions of what others view as acceptable, and by the behaviors they see others engaging in.

The belief that people behave differently in groups is a well-established social psychological tenet. In particular, the *bystander effect theory* in social sciences refers to the phenomenon where individuals are less likely to offer help or intervene in emergency situations when other people are present (Latané and Darley 1969; Darley and Latané 1968). This theory suggests that the presence of others can lead to diffusion of responsibility, where individuals believe that someone else will take action, resulting in a reduced likelihood of any single individual taking action themselves. The bystander effect has been studied extensively in psychology and sociology, and it highlights the complex social dynamics that influence human behavior in group settings (Latané and Darley 1968; Fischer et al. 2011).

However, less attention has been paid to how these dynamics play out in social media conversations, especially when a user becomes a target of toxic behaviors like harassment, hate speech, cyberbullying, or trolling. Previous literature has examined users' behavior in toxic conversations on social media (Aleksandric et al. 2024; Xia et al. 2020; Shen et al. 2020), highlighting that users tend to engage more in toxic than in non-toxic conversations. Also, Saveski et al. (Saveski, Roy, and Roy 2021) explored how the likelihood of a toxic reply differs depending on whether the parent post is toxic or not. Nevertheless, our understanding how the presence of others affects the user's behavior and their preference to encourage toxicity or stand up for the target remains underexamined.

Based on the bystander effect theory, the presence of others can diffuse one's sense of responsibility to help, with users believing another individual will act (Latané and Darley 1969; Darley and Latané 1968). Using social norms as a theoretical framework, we analyze a random sample of $\sim 9K$ *Twitter conversations* to explore how group dynamics can influence online behavior. We investigate how the presence of others in a conversation affects users' inclination to address toxic replies and how initial responses to such toxicity impact the overall tone of the conversation. To answer these questions, we conducted statistical tests while accounting for potential confounding factors, including users' account attributes, conversation structure, and topic of discussion.

Our findings suggest that there is a negative, statistically significant relationship between the number of conversation participants before the first toxic reply and the number of unique users who respond to a toxic reply in a non-toxic way. This indicates that the greater the pre-toxic participation, the fewer users tend to engage in a non-toxic way after the first toxic reply occurs. In addition, the results demonstrate that the toxicity levels of the initial responses to toxic replies tend to affect the tone of the remainder of the conversation by establishing a norm. Moreover, qualitative analysis was employed to investigate how often positive standing up, i.e., correcting misunderstandings, agreeing civilly, or defending an individual or a group of people, occurs in toxic conversations. Through qualitative analysis, we find that users rarely attempted to resolve the conflict, with only 19.7% standing up to resolve or stop the toxicity. Therefore, our findings suggest that group dynamics play an important role in shaping toxic threads, while some users' characteristics and discussion topics were also found relevant.

In summary, this study sheds light on how group dynamics influence online behavior, which could be an initial step for developing effective interventions against toxicity. In more detail, investigating how the presence of other users before the first toxic reply, as well as initial responses to toxicity, shape *Twitter conversations*, is crucial to this understanding. This comprehension can serve as a foundation for designing targeted interventions that leverage group dynamics to encourage positive engagement and mitigate the spread of harmful content. This contribution may help researchers understand how quickly antisocial norms can be established online and inspire future work to further inves-

tigate what factors lead users to adhere to or break social norms in addressing a toxic reply in civil conversations.

Related Work

Social Norms: Social norms are often conceptualized as prosocial, guiding interactions that benefit others or the collective while discouraging or sanctioning actions that cause harm (Heckathorn 1988). Foundational work has found that some individuals adhere to group social norms even when these perceptions conflict with their own (Asch 1956). But rather than promoting prosocial behavior, group dynamics can, in certain contexts, facilitate the spread of norm violations (Álvarez-Benjumea and Winter 2018). Observing antisocial behaviors, such as littering or jaywalking, can lead individuals to perceive as accepted norms increasing the likelihood that they will engage in similar behaviors themselves (Cialdini, Kallgren, and Reno 1991; Mullen, Copper, and Driskell 1990).

Research on cyberbullying has demonstrated that increasing the number of bystanders decreases intentions to intervene (Obermaier, Fawzi, and Koch 2016). However, other scholarship has shown that when individuals are aware they are visible to others, by using a webcam or making participants' screen-names more salient can reverse this effect (Van Bommel et al. 2012). Understanding the victim's perspective or empathizing with the target influence one's intentions of helping the victim (Paterson, Brown, and Walters 2019; Freis and Gurung 2013; Domínguez-Hernández, Bonell, and Martínez-González 2018). In online communities, the adoption of antisocial behaviors may be amplified by the anonymity and reduced social cues these settings provide (Suler 2004; Lee and Kim 2015), making individuals more likely to conform to group norms, even when those norms encourage harmful or antisocial conduct.

Detection and Classification. Empirical work on toxicity has employed machine learning-based detection algorithms to identify and classify offensive language, hate speech, and cyberbully (Zhang et al. 2016; Davidson et al. 2017; Koratana and Hu; Pitsilis, Ramampiaro, and Langseth 2018; Yin et al. 2023; Frenda et al. 2019). Works have used various methods ranging from lexical-based approaches (Markov et al. 2021; Wiegand et al. 2018) to deep neural networks (Mazari, Boudoukhani, and Djefal 2023; Roy et al. 2021; del Valle-Cano et al. 2023; Alshamrani et al. 2021; Chen, McKeever, and Delany 2019; Ribeiro and Silva 2019) Some recent works have used text and images together (Yang et al. 2019; Singh, Ghosh, and Jose 2017) as well as text and socio-cultural information (Vijayaraghavan, Larochelle, and Roy 2021) to detect hate speech. The state-of-the-art toxicity detection tool is available through Google's Perspective API (Google Perspective API 2021). Perspective API has been studied and used extensively in the previous literature (Kumarswamy, Singhal, and Nilizadeh 2025; Singhal et al. 2023; Salehabadi et al. 2022; Zannettou et al. 2020; Gröndahl et al. 2018; ElSherief et al. 2018; Saveski, Roy, and Roy 2021; Kumar et al. 2023; González-Pizarro and Zannettou 2023). Hence, we will use Google Perspective API to detect toxic tweets in conversations.

Hypotheses

Prior work has examined the contagion of online toxicity in a group setting, such as focusing on team-based predictors of toxicity (Shen et al. 2020). Our work adds to the literature by investigating the group dynamics in naturally formed Twitter conversations, without predefined team structures or performance-based roles. Within our study, in larger conversations where no one speaks out against toxic behavior, participants may infer an implicit acceptance of such exchanges, reinforcing the perception that hostile replies are socially acceptable. Therefore, we hypothesize:

H1: The number of users participating in a conversation before the first toxic reply is negatively associated with the number of users who post non-toxic replies after the first toxic reply.

We used the number of users who participated in the conversation before the first toxic reply as a proxy for users who are observing the conversation. The goal is to understand how the presence of a larger group shapes the remaining thread after the first toxic reply occurs. Typically, Twitter sends notifications to users when other accounts respond to their comments (X Help Center 2026), so users can keep observing what happens in the thread. Note that Twitter V2 provides the number of views a tweet has received. However, we did not use this metric because it is not a unique count (Twitter 2023) and counts multiple views of the same user. Thus, the number of followers of the root authors was used as a control variable, since they might also observe the conversation.

Reactions to toxic behaviors: Users observing the conversation can shape the reactions to toxic behavior online, whether they actively follow the perpetrator’s behavior by posting additional toxic replies or attempting to re-establish norms of decorum. Online settings provide researchers with an opportunity to investigate how quickly norms are established among group members and the degree to which one’s actions are expected to align with what is appropriate or expected within the group (Tajfel 2010). Once established, it is not surprising that individuals generally conform to group norms (Asch 1956). Other work shows that observing trolling behavior by others influences new users (Cheng et al. 2017). In other words, individuals may be more likely to post toxic replies after seeing others do so, believing it is the norm.

Peer conformity is positively associated with in-person bullying (Duffy and Nesdale 2009) and cyberbullying (Bleize et al. 2021; Bastiaensens et al. 2016). In this instance, we might expect the toxicity of tweets within the conversation to alter based on the introduction and reaffirmation of toxic content. For example, passive behavior by users who initially participated in the conversation may be perceived as implicit approval of hate speech, while individual users’ reactions, such as countering, are important in addressing online toxicity. For instance, when the first toxic tweet is met with negative sanctions, it may be perceived as inappropriate. In contrast, when initial toxicity is followed by further toxic replies, others may perceive such behavior as normative and follow suit. Thus, we test the following two hypotheses: **H2:** If a Twitter user posts a non-toxic re-

ply immediately after the toxic reply, then more users post non-toxic replies. **H3:** If a Twitter user posts a non-toxic reply immediately after the toxic reply, then the toxicity of the conversation after this reply is more likely to be non-toxic.

These hypotheses test the premise that passive behaviors of users already participating in the conversation can be perceived as implicit approval of toxicity. Even though prior work found that the language toxicity of a comment significantly increased the number of its replies (Xia et al. 2020), our analysis aims to shed light on whether the toxicity level of the initial response to the first toxic reply plays a significant role in toxicity levels of the remaining parts of the conversation.

Data Collection

Dataset and Cleaning. Our dataset comes from Twitter¹, one of the most widely used social media platforms. One part of our data originates from a recent study (Aleksandric et al. 2024) where the dataset consists of a random sample of public English tweets during the period of August 14th to September 28th, 2021 (*Dataset 1*). To expand our dataset, we used Twitter API V2 (Twitter 2022) to gather an additional random sample of posts from March 31st to April 23rd, 2023 (*Dataset 2*). A detailed data collection procedure for Dataset 2 is shown in Figure 1. We collected two datasets to investigate if our findings can be generalized over a period of time.

As described in Figure 1, two days after the collection of daily tweets, we would use Twarc (Twarc 2020) to collect complete conversations for each initial tweet in the daily random sample. Twarc is a command-line tool and Python library for collecting Twitter data. We removed English tweets that are retweets or replies in other public conversations to only obtain the full conversations. We also dismissed conversations in which all responses were posted by the author of the initial tweet, as no other users were involved. Moreover, the process of gathering conversation replies can take up to 24 hours, so some posts would have more time to acquire replies compared to others. Hence, we only saved replies received in the first 48 hours after the initial tweet was published, so the time frame for captured replies is consistent for all the conversations to reduce possible bias in the results. We used a threshold of 48 hours, as most of the replies occur closely after the initial tweet has been posted. We also removed conversations containing tweets or replies that only included links, images, or videos (named *empty tweets*) rather than text, as Perspective API is a text-based toxicity detection tool (Google Perspective API 2021). Furthermore, in certain cases, we were unable to extract the full conversation due to users hiding some of the replies or removing them. Such conversations were discarded from the dataset, as we would not certainly know whether the missing replies were toxic or who they belonged to. Finally, the cleaned dataset consists of 136,847 conversations containing 938,317 tweets.

¹The data was collected before Twitter was rebranded to X, while Twitter API V2 was available.

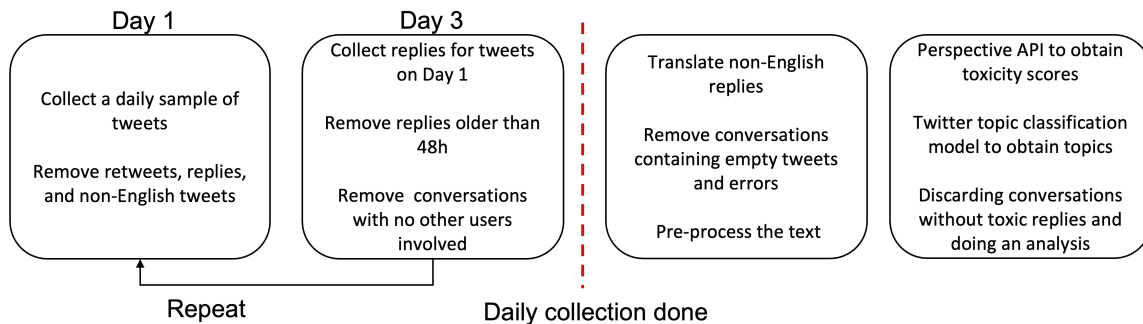


Figure 1: Data collection flow. The figure demonstrates the timeline used for a collection of a random sample of tweets and their replies, as well as the data pre-processing steps.

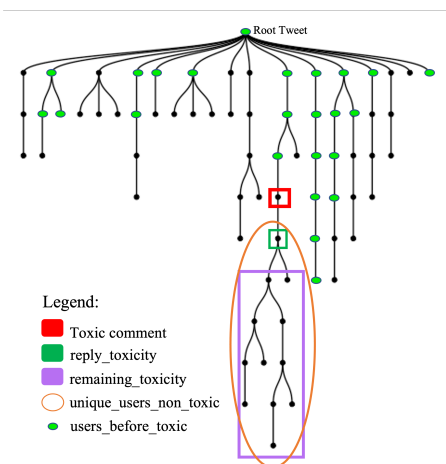


Figure 2: An example of a conversation tree.

Definitions. The user who posted the initial tweet is named *root author*, and each full conversation is represented as a tree that started with the initial tweet named the *root tweet*. As illustrated in Figure 2, each conversation is represented as a tree structure whose tweets (nodes) are connected when one is responding to another.

Obtaining Conversation Topic. Certain topics, such as political ones, can trigger greater toxicity or attract more attention from other users, potentially revealing that the bystander effect tends to occur more frequently in specific conflicts. Thus, we used the Twitter multi-topic classification model trained on TweetTopic dataset, which is shared and evaluated by the recent study (Antypas et al. 2022), to obtain coefficients for 19 relevant discussion points on social media. Some topic examples include *news & social concern, science & technology, diaries & daily life, business & entrepreneurs*, etc. This RoBERTa-based model is fine-tuned on the Twitter corpus and has been recently used by several works (Leiter et al. 2024; Hewitt et al. 2023; Cho et al. 2023; Towle and Zhou 2023). We passed our initial tweets to the model as input to obtain scores in the range 0-1, where higher scores indicate a higher probability that the tweet is linked to a specific topic.

Discovering Conversations with Toxic Replies. To identify all the toxic tweets in our dataset, we leveraged Google’s Perspective API (Google Perspective API 2021). Google Perspective API applies different machine learning models to score the toxicity of textual data. When a comment or text is passed as input to the Perspective API, the API returns scores for the requested attributes representing probability scores between 0 and 1. However, for this study, it is crucial to understand what such scores indicate. For example, a toxicity score of 0.8 means that 8 out of 10 people reading the comment would perceive the comment as toxic. To accomplish that, Google Perspective API is trained on millions of comments originating from multiple relevant sources, such as Wikipedia and The New York Times, across a range of languages (Google Perspective API 2021). These comments are annotated by 3-10 coders who speak the suitable language, followed by using their labels to train the API models (Google Perspective API 2021). Note that the models have been evaluated under the ROC curve and also checked for unintended biases for each of the identity groups (Performance Overview 2023).

In this paper, we consider scores for the *Toxicity* attribute since Google’s Perspective API defines a text having this attribute as a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion (Google Perspective API 2021). Note that before passing our tweets to the Perspective API, we cleaned the text of tweets by removing punctuation and URLs, and replacing emojis with appropriate text describing the emojis. Moreover, the fact that our initial tweets are written in English does not guarantee that their replies will also be in English. Thus, we translated non-English replies into English, using Google Trans API (Han 2024). There were around 287K replies that were not classified as English replies by Twitter. Finally, we created a binary variable indicating whether a tweet was toxic or not toxic. In more detail, a tweet is labeled as toxic if its *toxicity* score is higher than or equal to 0.7, according to Google’s Perspective API recommendation (Google Perspective API 2021). We acknowledge that Google’s Perspective API, as a toxicity detection tool, also has limitations (Nogara et al. 2025; TeBlunthuis, Hase, and Chan 2024; Sap et al. 2022), leading to the use of a stricter thresh-

old, intending to reduce bias in the results. However, prior research has demonstrated its effectiveness in identifying various forms of toxic language in generated text as well as on social media data (Ovalle et al. 2023; Kumarswamy, Singhal, and Nilizadeh 2025; Aleksandric et al. 2024).

Consequently, around 2.4% of the tweets in our dataset were considered as *toxic*, while 56.2% of the toxic tweets were posted by users other than root authors. The analysis focuses on toxic conversations only, as the bystander effect can only occur in cases where a toxicity attack exists within a conversation. Therefore, conversations that did not receive any toxic reply are discarded from the analysis, leaving the number of conversations at 10,455. In addition, we discarded 1,223 conversations that the root author initiated with the toxic reply (root tweet is toxic), as such conversations might not experience the same bystander effect as other conversations, which could possibly alter our results.

Also, note that some of the root authors appeared multiple times in our data, which could potentially introduce bias to our results since we use their account characteristics as control variables. Therefore, we randomly picked a single conversation from each user whose initial tweet repeatedly appeared in the dataset. Hence, the final dataset consists of 9,107 conversations consisting of 187,658 tweets posted by 118,609 unique users, where 5,115 conversations belong to dataset 1 and the rest of 3,992 conversations originated from dataset 2.

Methodology

We use multivariate regression analysis to test our hypotheses. Below, we explain our dependent, independent, and control variables in detail.

Independent Variables: The following independent variables are defined: (1) *users_before_toxic*: the number of unique users engaged in the conversation before the first toxic comment occurred. Examples of such tweets are colored in green in Figure 2. Note that even though some replies are at a higher level in the conversation tree, it does not necessarily mean that they occurred before the first toxic comment. For example, it is visible in Figure 2 that the second layer of the tree contains both green and unlabeled nodes. Even though some of the unlabeled nodes are positioned higher in the tree structure, they are still posted later than the green nodes in lower levels of the tree. Consequently, this variable was computed by chronologically ordering all tweets in the conversation from the oldest to the youngest and then calculating the number of unique users that posted tweets older than the first toxic tweet. Note that, on Twitter, with every new post, notifications are sent to users who have already participated in the discussion. For this variable, we could instead use the number of followers each victim has, but knowing that their accounts are public suggests that more people might be observing the conversation. Furthermore, followers might be observing the conversation; however, because they did not engage before the first toxic tweet, it is hard to claim they are interested in it. Therefore, we use the number of followers as a control variable in our regression models. (2) *reply_toxicity*: the toxicity score of the first

comment posted that is a direct reply to the first toxic comment in the conversation, e.g., circled in green in Figure 2.

Dependent Variables: To test our hypotheses, we defined the following dependent variables: (1) *unique_users_non_toxic*: the total number of unique users who posted non-toxic comments after the first toxic reply in the conversation thread. We focus only on the single conversation thread that emerged from the first toxic reply, as circled in orange in Figure 2. We did not use the chronological ordering of all tweets in the conversation because we would capture users involved in other conversation threads, even though they were not related to the first toxic comment. (2) *remaining_toxicity*: the ratio of all toxic replies that occurred in the conversation thread after the first toxic reply and the total number of replies in the thread. Note that this calculation excludes the *reply_toxicity*, and it is surrounded by the purple rectangle in Figure 2.

Control Variables: We controlled for several factors that could affect group behavior in Twitter conversations. We added controls for root authors' *activity*, i.e., *num_friends*, *num_tweets*, and *account_age*, because more active users might have different audiences. For example, if a user posts many tweets, followers might engage less with their tweets and be less likely to defend them against toxicity attacks. We also controlled for the *visibility*, i.e., *num_followers*, *listed_counts* and *verified*. For example, users with *verified* accounts or influencers with many followers might receive more help from others when they are under attack. Additionally, we controlled for profile characteristics, such as *description_length*, *has_URL*, and *has_location*. Users who provide less information on their profiles may receive less defense from other participants, as others may be less inclined to support anonymous users. Finally, we used the *width* and *depth* of the conversation tree to control for differences in conversation structure. *Depth* is the length from the root tweet to the conversation's deepest node, and *width* represents the maximum number of tweets at any level in the conversation tree. Finally, control variables included in all the models were conversation topics. There is a possibility that the topics of conversations influence how toxic conversations unfold, as well as whether users encourage or stand up against toxicity. The list of topics is as follows: *arts & culture*, *business & entrepreneurs*, *celebrity & pop culture*, *diaries & daily life*, *family*, *fashion & style*, *film tv & video*, *fitness & health*, *food & dining*, *gaming*, *learning & educational*, *music*, *news & social concern*, *other hobbies*, *relationships*, *science & technology*, *sports*, *travel & adventure*, and *youth & student life*. Each topic with a coefficient greater than 0.5 was set to 1, and 0 otherwise (converting it to binary), thereby determining which topics each conversation is associated with.

Dataset Characterization

Our final dataset consists of 9,107 conversations with 187,658 tweets, posted by 118,609 unique users. Table 1 shows the descriptive statistics of the variables used in the statistical models. Note that we display the statistics of the final (merged) dataset, including the statistics on each dataset

Table 1: Descriptive statistics of variables used in the analysis

	Variable	Merged Dataset				Dataset 1				Dataset 2			
		Min	Median	Mean	Max	Min	Median	Mean	Max	Min	Median	Mean	Max
Dependent Variables	remaining_toxicity	0	0	0.05	0.67	0	0	0.05	0.6	0	0	0.05	0.67
	unique_users_non_toxic	0	0	0.68	43	0	0	0.72	43	0	0	0.64	17
Independent Variables	users_before_toxic	1	2	5.13	463	1	2	4.6	385	1	2	5.8	463
	reply_toxicity	0.001	0.18	0.25	0.97	0.001	0.18	0.25	0.97	0.005	0.19	0.26	0.96
Control Variables	num_friends	0	589	2006	~ 1.5M	0	560	1847.1	~ 1.5M	0	642.5	2210.3	~ 610.5K
	num_tweets	1	~ 14K	~ 40K	~ 1.5M	1	~ 13K	~ 36K	~ 1.5M	1	~ 15.5K	~ 45K	~ 1M
	account_age	0	3	5.03	17	0	3	4.5	15	0	4	5.8	17
	num_followers	0	1,233	~ 97K	~ 55M	0	1,019	~ 58K	~ 54.4M	0	1613	~ 147K	~ 55M
	listed_counts	0	8	398.6	~ 218K	0	7	237.6	~ 102.8K	0	10	604.9	~ 218K
	verified	0	0	0.1	1	0	0	0.08	1	0	0	0.13	1
	description_length	0	78	81.29	193	0	74	79.55	183	0	83	85.52	193
	has_URL	0	0	0.5	1	0	0	0.5	1	0	0	0.5	1
	has_location	0	1	0.76	1	0	1	0.76	1	0	1	0.75	1
	width	1	3	11.39	1,688	1	3	9.97	1,688	1	3	13.22	853
	depth	1	3	4.23	197	1	3	4.3	197	1	3	4.13	73

separately. *Dataset 1* collected in 2021 includes 5,115 conversations, while *dataset 2* collected in 2023 includes 3,992 conversations. The collection of two datasets allows us to gain insights into overall group dynamics and to check whether the trend may have changed over time.

The minimum number of tweets in conversations is 2, and the maximum is 1,689 tweets. The mean number of tweets included in these conversations is 20.6. In more than half of conversations, the number of users engaged in the conversation before the first toxic reply (*users_before_toxic*) is 2, while the maximum number is 463. Also, the average of *reply_toxicity* is 0.25. The mean number of *unique_users_non_toxic* is 0.68, indicating that in more than half of the conversations, there are users who posted non-toxic comments after the first toxic reply. The deepest conversation in our dataset, i.e., *depth* is 197, while the maximum *width* is 1,688.

Interestingly, the maximum number of users posting non-toxic replies after a toxic reply in *dataset 1* is 47, which is higher than that in *dataset 2* (17). On the other hand, the maximum *users_before_toxic* in *dataset 2* is higher than that of *dataset 1* 463 vs. 385. Furthermore, the percentage of root authors in the two datasets who provided URLs and locations on their profiles does not differ significantly. However, Mann-Whitney tests indicate that all other root authors' characteristics differ significantly between the two datasets ($p < 0.05$). In addition, the differences in other dependent and independent variables between *dataset 1* and *dataset 2* are statistically significant ($p < 0.05$). In summary, the two conversation samples differed significantly, allowing us to examine whether our hypotheses hold across two time periods. The top five topics in the dataset were *celebrity & pop culture, film, tv & video, diaries & daily life, news & social concern*, and *sports*, which also matches the most prevalent topics in the two datasets separately.

Results

To test hypotheses H1 and H2, we employed a Poisson regression model as the distribution of *unique_users_non_toxic* does not follow a normal distribution, and it is a count variable. We ran a linear regression model to test hypothesis H3, as *remaining_toxicity* follows

a normal distribution. Additionally, we ran a Negative Binomial regression model that accounts for excess variance in count data, and our key findings remained consistent.

Higher levels of pre-toxic participation are associated with a lower number of users engaging in non-toxic responses following the toxic reply: The results obtained from a Poisson regression model are displayed in Table 2 (column H1) and reveal a statistically significant negative association between the *unique_users_non_toxic* and *users_before_toxic* ($p < 0.001$). In other words, we find that the greater the number of users participating in the conversation before the first toxic comment is significantly correlated with the lower the number of users posting non-toxic replies after that comment, in support of *H1*. Additionally, the results show that verified accounts are less likely to receive non-toxic comments from users after the first toxic reply than non-verified users, whereas the opposite is true for root authors who specified locations on their profiles. In other words, these results suggest that users are more likely to respond in a non-toxic way after toxicity when interacting with identifiable users, but are less likely to de-escalate when the root author is a verified account. It may be that highly visible accounts, such as verified users, are less likely to receive “standing up” responses to toxicity in their threads. Moreover, deeper conversations are more likely to have a higher number of users posting non-toxic replies. Interestingly, this implies that if a larger discussion develops, conversation participants tend to engage in a less toxic way.

Greater toxicity level of the initial response to the first toxic reply is associated with less non-toxic engagement.

In our second hypothesis, we posited that the first reply immediately after the first toxic reply might play an important role in how the rest of the conversation develops. For this analysis, we discarded conversations that do not contain a direct reply to toxic tweets, leaving the dataset with 3,959 conversations, consisting of 75,926 tweets posted by 37,627 unique users. *Reply_toxicity* was used as the independent variable in the Poisson regression model. Table 2 (column H2) shows that there is a negative statistically significant association between *reply_toxicity* and *unique_users_non_toxic* ($p < 0.001$). We find support for Hypothesis 2, with our results indicating that fewer users

posting non-toxic replies is associated with greater toxicity of the first reply to the initial toxic comment.

Toxic reply after first toxic reply is associated with more uncivil behavior: For our final hypothesis, H3, we examine how the toxicity of the first comment after the first toxic reply might determine the direction in which the conversation thread will emerge. In other words, if the first reply to a toxic reply is also toxic, the whole conversation thread might become more toxic. In this case, conversations that have less than two tweets after the first toxic reply are removed, leaving 2,230 conversations for analysis. Furthermore, *reply_toxicity* was used as an independent variable in the linear regression model, while the dependent variable used was *remaining_toxicity*. The model results (Table 2 - column H3) indicate that there is a positive statistically significant association between *reply_toxicity* and *remaining_toxicity* ($p < 0.001$). This shows that there is a correlation between the higher toxicity of the immediate comment after the first toxic reply, and the greater toxicity levels in the remaining of the conversation thread. Thus, we find evidence in support of H3.

To ensure that these findings remained consistent across samples, we ran models for both *dataset 1* from 2021 and *dataset 2* from 2023 individually. Once we determined that we found support for our hypotheses across datasets, the samples were combined and we presented findings from the merged dataset. This shows that our findings are not specific to a timeline and specific offline events and they can be generalized.

Table 2: Results of the regression analysis. Note that standard errors are presented in parentheses.

	Dependent variable:		
	unique_users_non_toxic	remaining_toxicity	
	H1: Poisson	H2: Poisson	H3 :OLS
users_before_toxic	-0.02*** (0.002)		
reply_toxicity		-0.37*** (0.058)	0.05*** (0.01)
Followers	0.0 (0.0)	-0.0 (0.0)	-0.0 (0.0)
Friends	-0.00001* (0.0)	0.00001* (0.0)	-0.0 (0.0)
Num_tweets	0.0 (0.0)	0.0 (0.0)	-0.0 (0.0)
Listed_count	0.0 (0.00001)	0.0 (0.00001)	0.0 (0.0)
VerifiedTrue	-0.45*** (0.06)	0.11 (0.06)	-0.01 (0.014)
Age	-0.01 (0.003)	0.004 (0.003)	-0.001 (0.001)
UrlTrue	-0.04 (0.03)	-0.01 (0.03)	-0.01 (0.01)
Description	0.00003 (0.0003)	0.001* (0.0003)	0.0001 (0.0001)
LocationTrue	0.11*** (0.03)	0.02 (0.03)	-0.01 (0.01)
Width	0.0001 (0.001)	0.0003 (0.0004)	-0.0002 (0.0001)
Depth	0.02*** (0.001)	0.01*** (0.001)	0.0001 (0.0003)
Observations	9,107	3,959	2,230
R ²			0.023

Note:

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Qualitative Analysis

Even though our regression analysis shows a negative association between pre-toxic user participation and the number of users posting non-toxic reply post-toxicity, additional qualitative analysis could provide insights into how often standing up with the intent to resolve the conflict occurs in our dataset.

Labeling of the first toxic replies. Firstly, we extracted a random sample of 200 conversations for the manual label-

ing, where at least one additional user was involved in the thread originating from the first toxic reply. The total number of such conversations is 3,959. This criterion ensures that standing up might be present, compared to conversations where no other users or replies are found after the first toxic reply. Then, two annotators labeled all the first toxic replies in these conversations with a binary label depending on whether it attacked an individual/group of people or not. The main idea is to filter out the conversations where the first toxic reply received a high toxicity score due to the usage of profanity words or attacks directed to non-human subjects, such as movies, anime, etc., as we are interested in positive standing up for human targets, as this scenario most closely resembles bullying compared to other scenarios mentioned. The two annotators selected for this task are computer science PhD students who are highly involved in social computing research, especially in content moderation and users' responses to online toxicity. The Cohen's Kappa (McHugh 2012) score for two sets of labels was 0.7, indicating substantial agreement. After the annotators met and resolved all conflicting labels, the total number of conversations in which the first toxic reply represented an attack against an individual or a group was 61.

Labeling positive standing up. The next step involved identifying conversations where positive standing up happens. Some examples of positive standing-ups are respectfully correcting misunderstandings, setting boundaries, disagreeing civilly, defending an individual or a group of people, etc. (Allison and Bussey 2016; Biernesser et al. 2023; Kärnä et al. 2011; Williford et al. 2012). As shown in our regression analysis, toxicity begets more toxicity. Therefore, negative standing up, such as counterattacking, responding with sarcasm, threatening other users, and similar, is likely to escalate a conflict further rather than resolve it. Hence, in this analysis, we are interested in how often positive standing up with the intent to resolve the conflict occurs. Once again, two annotators manually labeled a thread emerging from the first toxic replies in these 61 conversations, assigning a binary label to each reply indicating whether it represented a positive standing up or not. The total number of replies labeled was 1,501, including the first toxic replies. The intercoder agreement Cohen's Kappa score equals 0.6, suggesting moderate agreement between annotators. This implies that this task is not easy for humans, potentially due to many replies in some threads, making it hard for annotators to understand the context of the conversation. Annotators met to resolve any conflicting labels. The total number of standing-up replies was 20 out of 1,440, belonging to 12 (19.7%) conversations. These findings demonstrated that positive standing up with the intent to resolve the conflict rarely happens.

Representative examples of positive and negative standing up. Below, we provide examples of positive and negative standing up that were identified by the coders in our dataset. Note that the reply that is in *red* color signifies negative standing up, i.e., counterattacking and arguing their point of view by insulting others. Additionally, the reply in *green* color demonstrated positive standing up, in this case, respectfully stating their point of view.

First toxic reply: [MASK] You are not going to call the idiot an idiot He will fight the Austin ISD all the way to the scummy TX Supreme Court and then your son won't be safe Take a stand and call this pig out and his hypocrisy The moron is playing with Texans lives while getting the vaccine himself.

First response: [MASK] [MASK] Getting the vaccine is a personal choice Same with wearing a mask One can choose to do so or not Respect the right to choose.

Second response: [MASK] [MASK] No it's not Are you stupid It impacts others so it is not You don't live in a cave any more Is it your choice to shoot a gun on the street It's not Get a clue Kids need to show multiple vaccinations before going to school camp etc.

Discussion

The present study sheds light on the practical question of how norms of toxicity are established online. Online settings allow us to investigate how quickly such group norms are established and how closely one's actions conform to what is deemed appropriate or expected in the group (Tajfel 2010). Our study cannot measure users' perceptions of the group nor their internalization of norms. However, we use a framework of social norms to gain insight into behavior, such as the number of participants prior to the first toxic reply and subsequent posting patterns, as proxies for the social context in which users are impacted. These behavioral patterns reveal how social dynamics within a single conversation can shape the nature of replies, highlighting the need for further research on the formation of descriptive social norms online.

We do not claim that the bystander effect is directly present in our dataset, as it is difficult to determine how many users actually observe a given incident (e.g., a toxic reply). However, when individuals see that no one intervenes in a toxic exchange, they may interpret this silence as signaling that such behavior is acceptable or that intervention may be further met with more antisocial behavior. This can create a feedback loop in which users look to others for cues, and the absence of response reinforces norms that tolerate or even legitimize toxic or antisocial behavior. This dynamic may suppress helping behavior, not only through diffusion of responsibility, but also by shaping perceptions of what constitutes appropriate online replies.

Despite these robust findings, the scope of this study was limited. Firstly, our sample was restricted to tweets in English. Secondly, our dataset does not capture the 'true' number of users who observe the conversation. Furthermore, what different users see on their feeds might be algorithmically selected to recommend the root post or relevant replies. Due to this, it is impossible to assume that a typical user will consume all the threads in the conversation or all the replies that occurred before the first toxic comment. Thus, it is hard to estimate the number of conversation observers. However, we believe *users_before_toxic* can be a good proxy for the unique number of views because users who choose to participate are also more likely to observe how the rest of the conversation will unfold. Hence, they could potentially stand up for users who are attacked. In addition, the extent to which

social media users counter toxicity can be influenced by factors such as the extremity of the views expressed (Schieb and Preuss 2016). Finally, we acknowledge that Google's Perspective API as a toxicity detection also contains certain limitations (Nogara et al. 2025; TeBlunthuis, Hase, and Chan 2024; Sap et al. 2022). Despite these limitations, our results suggest that the perceptions of social group norms through increased conversational participants is associated with fewer users responding in a non-toxic manner to toxic replies. Furthermore, the similar trends observed across two datasets imply that this phenomenon was not chance-driven but rather reflects the true nature of online group social dynamics, which did not change over time.

Scholarship can build upon our work to investigate the decision-making process and the potential costs users face when considering whether to intervene or not to alter the tone of a Twitter conversation. For instance, users observing the conversation may struggle to empathize with targets of toxicity or hate speech because they lack insight into others' perspectives or feelings. The literature showed that the bystanders' empathic concern shapes motivation to act and intervene (Machackova, Dedkova, and Mezulanikova 2015). Some recent scholarship has examined how technological designs may encourage bystander interventions on cyberbullying online on a large scale (Taylor et al. 2019). Future work should examine interventions such as fostering bystanders' role-taking, strengthening perspective-taking, and empathy for targets of incivility (Davis and Love 2017).

Conclusion

In this study, we assessed a random sample of *Twitter Conversations* to understand how social norms and the toxicity level of the first response to a toxic reply can lead to the disinhibition of users' toxic replies. Motivated by theory of social norms and by the bystander effect theory, we find that increased conversational participants are associated with fewer Twitter users standing up to a toxic reply. We also highlight the importance of initial responses to toxic tweets within a conversation. Our results demonstrate that there is an association between posting a toxic reply immediately after an initial toxic comment and an increased likelihood of the remaining of the conversation being more toxic. Understanding how users respond to uncivil or abusive content helps reveal how social norms shape behavior online.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant No. 2309318. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We would like to thank Sayak Saha Roy and Nazanin Salehabadi for initiating the research and helping us in the initial phase of data collection.

References

- Aleksandric, A.; Roy, S. S.; Pankaj, H.; Wilson, G. M.; and Nilizadeh, S. 2024. Users' Behavioral and emotional response to toxicity in twitter conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 29–42.
- Allison, K. R.; and Bussey, K. 2016. Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review*, 65: 183–194.
- Alshamrani, S.; Abusnaina, A.; Abuhamad, M.; Nyang, D.; and Mohaisen, D. 2021. Hate, obscenity, and insults: Measuring the exposure of children to inappropriate comments in youtube. In *Companion Proceedings of the Web Conference 2021*, 508–515.
- Álvarez-Benjumea, A.; and Winter, F. 2018. Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, 34(3): 223–237.
- Anderson, A. A.; Yeo, S. K.; Brossard, D.; Scheufele, D. A.; and Xenos, M. A. 2016. Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research*, 30(1): 156–168.
- 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.
- Asch, S. E. 1956. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and applied*, 70(9): 1.
- Bastiaensens, S.; Pabian, S.; Vandebosch, H.; Poels, K.; Van Cleemput, K.; DeSmet, A.; and De Bourdeaudhuij, I. 2016. From normative influence to social pressure: How relevant others affect whether bystanders join in cyberbullying. *Social Development*, 25(1): 193–211.
- Biernesser, C.; Ohmer, M.; Nelson, L.; Mann, E.; Farzan, R.; Schwanke, B.; and Radovic, A. 2023. Middle School Students' Experiences with Cyberbullying and Perspectives Toward Prevention and Bystander Intervention in Schools. *Journal of School Violence*, 22(3): 339–352.
- Binns, A. 2012. DON'T FEED THE TROLLS! Managing troublemakers in magazines' online communities. *Journalism practice*, 6(4): 547–562.
- Bleize, D. N.; Tanis, M.; Anshütz, D. J.; and Buijzen, M. 2021. A social identity perspective on conformity to cyber aggression among early adolescents on WhatsApp. *Social Development*, 30(4): 941–956.
- Chen, H.; McKeever, S.; and Delany, S. J. 2019. The use of deep learning distributed representations in the identification of abusive text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 125–133.
- 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*, 1217–1230.
- Cho, I.; Takahashi, R.; Yanase, Y.; and Saito, H. 2023. Deep RL with Hierarchical Action Exploration for Dialogue Generation. *arXiv preprint arXiv:2303.13465*.
- Cialdini, R. B.; Kallgren, C. A.; and Reno, R. R. 1991. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology*, volume 24, 201–234. Elsevier.
- Darley, J. M.; and Latané, B. 1968. Bystander intervention in emergencies: diffusion of responsibility. *Journal of personality and social psychology*, 8(4p1): 377.
- Davidson, T.; Warmley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aai conference on web and social media*.
- Davis, J. L.; and Love, T. P. 2017. Self-in-self, mind-in-mind, heart-in-heart: The future of role-taking, perspective taking, and empathy. In *Advances in group processes*, volume 34, 151–174. Emerald Publishing Limited.
- del Valle-Cano, G.; Quijano-Sánchez, L.; Liberatore, F.; and Gómez, J. 2023. SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles. *Expert Systems with Applications*, 216: 119446.
- Domínguez-Hernández, F.; Bonell, L.; and Martínez-González, A. 2018. A systematic literature review of factors that moderate bystanders' actions in cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 12(4).
- Duffy, A. L.; and Nesdale, D. 2009. Peer groups, social identity, and children's bullying behavior. *Social development*, 18(1): 121–139.
- Duggan, M. 2014. *Online harassment*. Pew Research Center.
- Dutton, W. H. 1996. Network rules of order: Regulating speech in public electronic fora. *Media, Culture & Society*, 18(2): 269–290.
- ElSherief, M.; Kulkarni, V.; Nguyen, D.; Wang, W. Y.; and Belding, E. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media*.
- Fischer, P.; Krueger, J. I.; Greitemeyer, T.; Vogrincic, C.; Kastenmüller, A.; Frey, D.; Heene, M.; Wicher, M.; and Kainbacher, M. 2011. The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological bulletin*, 137(4): 517.
- Freis, S. D.; and Gurung, R. A. 2013. A Facebook analysis of helping behavior in online bullying. *Psychology of popular media culture*, 2(1): 11.
- Frenda, S.; Ghanem, B.; Montes-y Gómez, M.; and Rosso, P. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5): 4743–4752.

- González-Pizarro, F.; and Zannettou, S. 2023. Understanding and detecting hateful content using contrastive learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 257–268.
- Google Perspective API. 2021. <https://www.perspectiveapi.com/>.
- Gröndahl, T.; Pajola, L.; Juuti, M.; Conti, M.; and Asokan, N. 2018. All You Need is "Love" Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, 2–12.
- Han, S. 2024. Google Trans API.
- Heckathorn, D. D. 1988. Collective sanctions and the creation of prisoner's dilemma norms. *American Journal of Sociology*, 94(3): 535–562.
- Hewitt, J.; Thickstun, J.; Manning, C. D.; and Liang, P. 2023. Backpack language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9103–9125.
- Hwang, H.; Borah, P.; Namkoong, K.; and Veenstra, A. 2008. Does civility matter in the blogosphere? Examining the interaction effects of incivility and disagreement on citizen attitudes. In *58th Annual Conference of the International Communication Association, Montreal, QC, Canada*.
- Kärnä, A.; Voeten, M.; Little, T. D.; Poskiparta, E.; Alanen, E.; and Salmivalli, C. 2011. Going to scale: A nonrandomized nationwide trial of the KiVa antibullying program for grades 1–9. *Journal of consulting and clinical psychology*, 79(6): 796.
- Koratana, A.; and Hu, K. ????. Toxic Speech Detection.
- Kumar, D.; Hancock, J.; Thomas, K.; and Durumeric, Z. 2023. Understanding the behaviors of toxic accounts on reddit. In *Proceedings of the ACM Web Conference (WWW)*.
- Kumarswamy, N.; Singhal, M.; and Nilizadeh, S. 2025. Causal Insights into Parler's Content Moderation Shift: Effects on Toxicity and Factuality. In *Proceedings of the ACM on Web Conference 2025*, 3762–3771.
- Latane, B.; and Darley, J. M. 1968. Group inhibition of bystander intervention in emergencies. *Journal of personality and social psychology*, 10(3): 215.
- Latané, B.; and Darley, J. M. 1969. Bystander" apathy". *American Scientist*, 57(2): 244–268.
- Lee, S.-H.; and Kim, H.-W. 2015. Why people post benevolent and malicious comments online. *Communications of the ACM*, 58(11): 74–79.
- 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*, 16: 100541.
- Machackova, H.; Dedkova, L.; and Mezulanikova, K. 2015. Brief report: The bystander effect in cyberbullying incidents. *Journal of adolescence*, 43: 96–99.
- Markov, I.; Ljubešić, N.; Fišer, D.; and Daelemans, W. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 149–159.
- Mazari, A. C.; Boudoukhani, N.; and Djeflal, A. 2023. BERT-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing*, 1–15.
- McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3): 276–282.
- Mullen, B.; Copper, C.; and Driskell, J. E. 1990. Jaywalking as a function of model behavior. *Personality and Social Psychology Bulletin*, 16(2): 320–330.
- Nogara, G.; Pierri, F.; Cresci, S.; Luceri, L.; Törnberg, P.; and Giordano, S. 2025. Toxic Bias: Perspective API misreads German as more toxic. In *Proceedings of the international AAAI conference on web and social media*, volume 19, 1346–1357.
- Obermaier, M.; Fawzi, N.; and Koch, T. 2016. Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New media & society*, 18(8): 1491–1507.
- Ovalle, A.; Goyal, P.; Dhamala, J.; Jagers, Z.; Chang, K.-W.; Galstyan, A.; Zemel, R.; and Gupta, R. 2023. "I'm fully who I am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1246–1266.
- Papacharissi, Z. 2002. The virtual sphere: The internet as a public sphere. *New media & society*, 4(1): 9–27.
- Paterson, J. L.; Brown, R.; and Walters, M. A. 2019. The short and longer term impacts of hate crimes experienced directly, indirectly, and through the media. *Personality and Social Psychology Bulletin*, 45(7): 994–1010.
- Performance Overview. 2023. https://developers.perspectiveapi.com/s/about-the-api-model-cards?language=en_US&tabset=20254=3.
- Pitsilis, G. K.; Ramampiaro, H.; and Langseth, H. 2018. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.
- Ribeiro, A.; and Silva, N. 2019. INF-HatEval at SemEval-2019 Task 5: Convolutional neural networks for hate speech detection against women and immigrants on twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 420–425.
- Rost, K.; Stahel, L.; and Frey, B. S. 2016. Digital social norm enforcement: Online firestorms in social media. *PLoS one*, 11(6): e0155923.
- Roy, S. G.; Narayan, U.; Raha, T.; Abid, Z.; and Varma, V. 2021. Leveraging multilingual transformers for hate speech detection. *arXiv preprint arXiv:2101.03207*.
- Salehabadi, N.; Groggel, A.; Singhal, M.; Roy, S. S.; and Nilizadeh, S. 2022. User Engagement and the Toxicity of Tweets.
- Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5884–5906.

Seattle, United States: Association for Computational Linguistics.

Saveski, M.; Roy, B.; and Roy, D. 2021. The structure of toxic conversations on Twitter. In *Proceedings of the Web Conference 2021*, 1086–1097.

Schieb, C.; and Preuss, M. 2016. Governing hate speech by means of counterspeech on Facebook. In *66th ica annual conference, at fukuoka, japan*, 1–23.

Shen, C.; Sun, Q.; Kim, T.; Wolff, G.; Ratan, R.; and Williams, D. 2020. Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. *Computers in Human Behavior*, 108: 106343.

Singh, V. K.; Ghosh, S.; and Jose, C. 2017. Toward multi-modal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2090–2099.

Singhal, M.; Ling, C.; Paudel, P.; Thota, P.; Kumarswamy, N.; Stringhini, G.; and Nilizadeh, S. 2023. SoK: Content moderation in social media, from guidelines to enforcement, and research to practice. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, 868–895. IEEE.

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

Tajfel, H. 2010. *Social identity and intergroup relations*, volume 7. Cambridge University Press.

Taylor, S. H.; DiFranzo, D.; Choi, Y. H.; Sannon, S.; and Bazarova, N. N. 2019. Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–26.

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*, 18(3): 278–299.

Towle, B.; and Zhou, K. 2023. Model-Based Simulation for Optimising Smart Reply. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12030–12043.

Twarc. 2020. Collect Twitter Data with Twarc! <https://scholarslab.github.io/learn-twarc/>.

Twitter. 2022. Twitter API.

Twitter. 2023. About view counts. <https://help.twitter.com/en/using-twitter/view-counts>.

Van Bommel, M.; Van Prooijen, J.-W.; Elffers, H.; and Van Lange, P. A. 2012. Be aware to care: Public self-awareness leads to a reversal of the bystander effect. *Journal of Experimental Social Psychology*, 48(4): 926–930.

Vijayaraghavan, P.; Larochelle, H.; and Roy, D. 2021. Interpretable multi-modal hate speech detection. *arXiv preprint arXiv:2103.01616*.

Wiegand, M.; Ruppenhofer, J.; Schmidt, A.; and Greenberg, C. 2018. Inducing a lexicon of abusive words—a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1046–1056.

Williford, A.; Boulton, A.; Noland, B.; Little, T. D.; Kärnä, A.; and Salmivalli, C. 2012. Effects of the KiVa anti-bullying program on adolescents' depression, anxiety, and perception of peers. *Journal of abnormal child psychology*, 40: 289–300.

X Help Center. 2026. About the Notifications timeline. <https://help.twitter.com/en/managing-your-account/understanding-the-notifications-timeline>. Accessed: April 10, 2026.

Xia, Y.; Zhu, H.; Lu, T.; Zhang, P.; and Gu, N. 2020. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW2): 1–23.

Yang, F.; Peng, X.; Ghosh, G.; Shilon, R.; Ma, H.; Moore, E.; and Predovic, G. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the third workshop on abusive language online*, 11–18.

Yin, W.; Agarwal, V.; Jiang, A.; Zubiaga, A.; and Sastry, N. 2023. Annobert: Effectively representing multiple annotators' label choices to improve hate speech detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 902–913.

Zannettou, S.; ElSherief, M.; Belding, E.; Nilizadeh, S.; and Stringhini, G. 2020. Measuring and Characterizing Hate Speech on News Websites. In *12TH ACM WEB SCIENCE CONFERENCE*. ACM.

Zhang, J.; Kumar, R.; Ravi, S.; and Danescu-Niculescu-Mizil, C. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 136–141.

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 Hypotheses and Methodology sections.**
- (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 Discussion 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? **No, as we are not making the dataset public.**
- (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 Hypotheses section.**
- (b) Have you provided justifications for all theoretical results? **Yes, in the Results section.**
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **Yes, in the Discussion section.**
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **Yes, in the Results and Discussion sections.**
- (e) Did you address potential biases or limitations in your theoretical framework? **Yes, in Hypotheses and Discussion sections.**
- (f) Have you related your theoretical results to the existing literature in social science? **Yes, in the Discussion section.**
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **Yes, in the Discussion 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)? **NA**
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **NA**
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **NA**
- (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)? **NA**
- (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **NA**
- (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? **NA**
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.**
- (b) Did you mention the license of the assets? **No, as the dataset is publically available.**
- (c) Did you include any new assets in the supplemental material or as a URL? **No**
- (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **No, as we are using publically available Twitter data.**
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes, we mentioned that the data contains toxic conversations.**
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see ?)? **NA**
- (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset **NA**
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 de-identified? **NA**