Understanding the Behaviors of Toxic Accounts on Reddit

Deepak Kumar
Stanford University

Jeff Hancock
Stanford University

Kurt Thomas
Google, Inc.

Zakir Durumeric
Stanford University

ABSTRACT

Toxic comments are the top form of hate and harassment experienced online. While many studies have investigated the types of toxic comments posted online, the effects that such content has on people, and the impact of potential defenses, no study has captured the behaviors of the accounts that post toxic comments or how such attacks are operationalized. In this paper, we present a measurement study of 929K accounts that post toxic comments on Reddit over an 18 month period. Combined, these accounts posted over 14 million toxic comments that encompass insults, identity attacks, threats of violence, and sexual harassment. We explore the impact that these accounts have on Reddit, the targeting strategies that abusive accounts adopt, and the distinct patterns that distinguish classes of abusive accounts. Our analysis informs the nuanced interventions needed to curb unwanted toxic behaviors online.

ACM Reference Format:
Deepak Kumar, Jeff Hancock, Kurt Thomas, and Zakir Durumeric. 2023. Understanding the Behaviors of Toxic Accounts on Reddit. In *Proceedings of the ACM Web Conference 2023 (WWW '23), May 1–5, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3543507.3583522

1 INTRODUCTION

Content Warning: This paper studies toxic content online. When necessary for clarity, this paper directly quotes user content that contains offensive/hateful speech, profanity, and potentially triggering content related to sexual assault.

Toxic comments—such as insults, threats of violence, and sexual harassment—are the top form of hate and harassment experienced online [41]. Such toxic behaviors reduce the emotional safety of targets and audiences who view the content. This can lead users to self-censor to avoid further attacks, leave online platforms altogether, and in some tragic cases, inflict self-harm [11, 20]. Transparency reports from Meta estimate that 0.14–0.15% of all views on Facebook in 2021 were of toxic posts [12], while Twitter reports that it removed roughly two million accounts in the second half of 2020 due to hate and harassment [42].

Prior research into toxic comments has focused on a variety of themes including the experiences of targets [39, 40, 44], the characterizations of specific, large-scale events like #GamerGate [8], early warnings for how toxic conversations escalate [46, 48, 49], the off-platform coordination tactics for attacks against targets [1, 29], and the impact of intervention techniques such as suspending accounts or banning entire communities [5, 7, 36]. While these studies all paint a rich tapestry of online toxic behaviors, none capture the long-term activities of *abusive accounts* (i.e., accounts that post toxic comments), such as the frequency of their toxicity behaviors or their impact on the platform itself. Such analysis is crucial to understanding what interventions—such as nudges, warnings, and bans—might best reduce online toxicity.

In this work, we present a quantitative study of accounts on Reddit that post toxic comments. Over an 18 month period, we identified 929K abusive accounts that posted 14 million toxic comments, and use this perspective to study three research questions:

RQ1: What is the aggregate impact of abusive accounts on Reddit? Abusive accounts that post at least one toxic comment make up 3.1% of all accounts that posted to Reddit during our analysis window, with their toxic comments comprising 0.8% of all content on Reddit. Toxic comments are highly visible on Reddit: 55.2% of Reddit accounts post directly on a thread with a toxic comment. Unlike automated, fake accounts that solely post spam [15], abusive accounts readily engage in non-toxic conversations, contributing an astounding 33.3% of all comments to Reddit. As such, simply banning abusive accounts would have substantial additional consequences to the platform.

RQ2: What are the unique attack patterns (e.g., mob-like coordinated attacks on a single individual) that abusive accounts use when posting toxic comments online? In graphing the reply relationships between attackers and their 1.6M receivers[1] (i.e., accounts who received a toxic comment as a reply), we observe three classes of attacks. The majority of receivers (76.1%) experience spurious, one-off toxic interactions. These receivers of abuse rarely have existing network relationships with their attacker, suggesting that the majority of abuse on Reddit is contextual and not necessarily premeditated. However, the remaining attacks are more pernicious: 15.8% of receivers experience *repeated abuse*, where a single abusive account continuously attacks the target, often across subreddits. Another 8.1% of receivers experience *flooding*, whereby a cluster of abusive accounts simultaneously attack the target, akin to coordinated raids.

RQ3: What are the classes of abusive accounts, and how do they inform more nuanced defenses against toxic behaviors? Finally, we cluster the toxicity behaviors of abusive accounts based on their posting volume, toxicity levels, subreddit participation, and community norm violations. We identify three distinct classes of

WWW '23, May 1–5, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583522

[1] Similar to research in intimate partner violence, we intentionally avoid the term "victim" to not disempower people facing abuse.

attackers. *Occasional abusers*—accounts that post just a handful of toxic comments—make up 71% of abusive accounts and 71% of all toxic comments. This suggests that modest interventions, such as nudges or warnings, may be effective for more than two-thirds of the toxic behaviors on Reddit. Conversely, *moderate abusers*—accounts that post a substantial volume of toxic comments—make up another 24% of abusive accounts. *Serial abusers*—accounts that extensively post toxic comments—make up the remaining 4.3% of abusive accounts. These two latter classes pose a greater threat and require more stringent interventions, however, their volume of toxicity make them potentially easier to take action on.

Combined, our findings illustrate the need for nuanced interventions in tackling unique toxicity patterns and varied classes of toxic accounts. In addition, we highlight a variety of features, such as reply relationships between abusive accounts, account toxicity trends over time, and community norm violations, that might inform future work into contextual signals for toxic account detection. To this end, we plan to release anonymized datasets to researchers on request to reproduce our analyses, develop new detection mechanisms, and further explore how toxic behaviors are operationalized online.

2 BACKGROUND AND RELATED WORK

In this section, we provide the necessary background and describe prior work that we build on to conduct our analysis.

2.1 Accounts that exhibit anti-social behaviors

Our study primarily builds on a number of quantitative and qualitative studies of accounts that exhibit anti-social behaviors online, such as trolling, bullying, and toxicity. Early work demonstrated that "anyone can become a troll" depending on contextual factors like the time of day and users' moods [9]. Newer studies have focused on the accounts that post toxicity and hate speech and the adversarial nature of toxic interactions. Maity et al. study toxic conflicts on Twitter, demonstrating how context and a predisposition to toxic behaviors can cause accounts to become *repeat offenders* in terms of toxic interactions [28]. Most similar to our work, Mathew et al. studied behaviors of hateful accounts on Gab, a popular fringe social platform for "unregulated speech", and highlighted how hateful behaviors can grow over time [30].

Other lines of work have focused on the properties of abusive accounts. For example, Ribeiro et al. studied abuser properties, like follower-following ratios and account age, with the aim of detecting hateful users based on their previous comments and their place in the social graph [35]. Several studies have also looked at abuse targeted to high-profile populations. For example, Hua et al. identify specific properties of abusers that adversarially interact with political candidates on Twitter [17, 18]. Finally, our work leverages methods and techniques of prior work that has investigated fringe hate groups and online communities, including discourse on Gab [30, 47], Dissenter [37], the Manosphere [16], and 4chan's politically incorrect board [33]. To extend this work, our study contributes a characterization of the distinct attack patterns that abusive accounts exhibit on Reddit.

2.2 Toxicity on Reddit

We build on a number of Reddit-focused studies to inform our experimental design and analysis choices. Gilbert documented how a culture of masculinity on Reddit forms a toxic technoculture on `r/AskHistorians`, leaving both moderators and users subject to abuse anywhere from name-calling to "prolonged harassment, doxxing, death threats, and rape threats." [13] Other studies have focused on how toxicity plays a role in shaping norms of subcommunities on Reddit [6, 34] and identified that subreddits often exhibit unique macro, meso, or micro-norms, highlighting challenges in applying a broad definition of toxicity throughout the entire platform. Such subcommunity specific-norms also lead to varied experiences with the precursors and effects of toxic discussions. Xia et al [46] leverage the Perspective API to study how specific *antecedents* of a toxic interaction, such as an accounts' prior history posting toxic comments and the subreddit context, can play a significant correlative role in predicting new toxic behaviors. Finally, several studies have focused on shifts in toxicity both on-platform and off-platform due to cross community movement following notable bans [5, 36]. Our study contributes distinct classes of abusive accounts on Reddit while taking into account both global and subcommunity-specific toxicity norms.

2.3 Interventions and mitigation strategies

Our work is ultimately grounded in prior work in studying defenses against online toxicity. Such interventions include nudges [22], providing realtime feedback on toxicity [46], foregrounding norms [34], and outright permanent bans [23]. Other research has focused on predicting toxic behavior based on early warning signs in conversation [48, 49], and how such signs from conversation flow can aid in forecasting personal attacks [21]. Similarly, our work builds on recent research from the computer security community, which has recently drawn parallels between classic cybersecurity problems (e.g., for-profit cybercrime) and anti-social online abuse [41]. As such, our defensive recommendations and techniques also build on work on measuring and mitigating spam [15, 27] and prior work that study graph-based spam propagation [32]. Our study contributes additional context by outlining what interventions may be most effective for distinct classes of abusive accounts.

3 METHODOLOGY

In this section, we detail our methodology for collecting a corpus of 2.2 billion Reddit comments as well as our classification techniques and thresholds for identifying abusive accounts. Figure 1 shows each step in our data collection pipeline.

3.1 Collecting Reddit comments

Our dataset consists of 18 months of comments posted on Reddit between January 2020 and June 2021. We collected a total of 2.2 billion comments via Pushshift, a third-party API that aggregates Reddit comments and posts (Figure 1.A) [3]. Each comment includes a timestamp, the username of the author, the subreddit (i.e., community) where the comment appeared, and graph data that allows us to identify if the comment was a top-level thread (i.e., the author was the original poster), or a reply to an existing thread. From this,

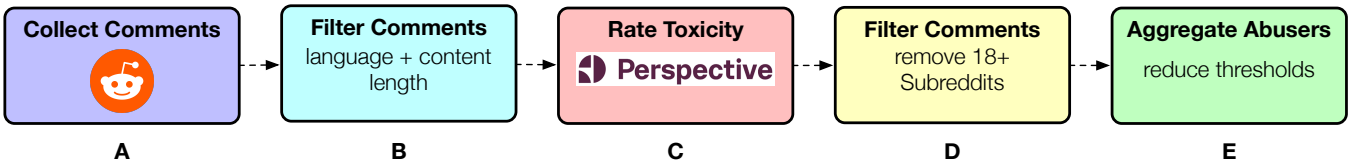


Figure 1: Reddit Processing Pipeline—We label Reddit comments sourced from Pushshift through the Perspective API. We explicitly filter out comments that are not in English or are from subcommunities tagged as 18+. We leverage these classifications to identify the 14M toxic comments and 929K abusive accounts we study.

we re-constructed accounts' posting histories and interactions with other accounts.

3.2 Filtering comments

We restricted our dataset to only English comments, in part to enable manual analysis by the researchers and due to the fact that existing toxicity classification models are trained primarily on English text.[2] We omitted comments that are less than 15 characters or more than 300 characters in length, which is aligned with prior research on the limitations of using existing toxicity models for short and very long text [24]. These filters reduced our corpus to 1.8B comments, 32.1M accounts, and 845K unique subreddits.

3.3 Identifying toxic comments

We classified the toxicity of each comment (Figure 1.C) using the Perspective API, a set of out-of-the-box toxicity classifiers from Google Jigsaw, which has been used extensively in prior research [17, 38, 46].[3] The Perspective API takes a comment as input and returns a score from 0–1 for several classifiers (e.g., profanity, threats, identity attacks, general toxicity). As the Perspective API is not explicitly trained on Reddit data, we needed to take an additional calibration step to identify the best classifier and classification threshold for our study.

To identify the best model and threshold for our context, we leveraged a public dataset that contains crowdsourced toxicity ratings for 16K Reddit comments [24]. Five-participants labeled the toxicity of each Reddit comment on a 5-point Likert scale from "not at all toxic" to "very toxic". We consider a comment to be toxic by raters if the median score across all raters was "moderately toxic" or higher. We swept over each Perspective API model and threshold value (e.g., 0.0, 0.01, 0.02, etc.) and compared results to participant labels (Table 1). Only the `SEVERE_TOXICITY` classifier achieved an acceptable precision of 0.75 at a threshold of 0.9. As such, when identifying toxic comments, we filter based on if a comment has a `SEVERE_TOXICITY` score > 0.9. We stress that this decision intentionally favors precision over recall, as our intention is not necessarily to study *all* toxic behaviors on Reddit, but rather, to study the toxic behaviors we have high confidence in.

3.4 Removing 18+ subreddits

As part of our manual validation of the pipeline, we observed that many comments flagged with high toxicity scores were sexually explicit and sourced from subcommunities that are tagged as 18+.

[2]We identify English comments using the `whatlanggo` Golang package.
[3]https://perspectiveapi.com

Classifier	Threshold	Precision	F1
`IDENTITY_ATTACK`	0.9	0.62	0.02
`INSULT`	0.9	0.53	0.11
`TOXICITY`	0.9	0.51	0.24
`SEVERE_TOXICITY`	**0.9**	**0.75**	0.02
`THREAT`	0.9	0.43	0.06

Table 1: Optimal Perspective API Thresholds for Ground Truth Toxic Comments—The thresholds that maximize precision for each Perspective API classifier on Reddit are all 0.9 or higher, with only one classifier, `SEVERE_TOXICITY`, achieving an acceptable precision for our study.

Threshold	Nonabuser Prec.	Abuser Prec.	Comment %
0.5	0.19	0.56	64M (3.9%)
0.7	0.22	0.68	27M (1.7%)
0.8	0.29	**0.72**	**14M (0.8%)**
0.9	0.0	0.79	1.7M (0.09%)

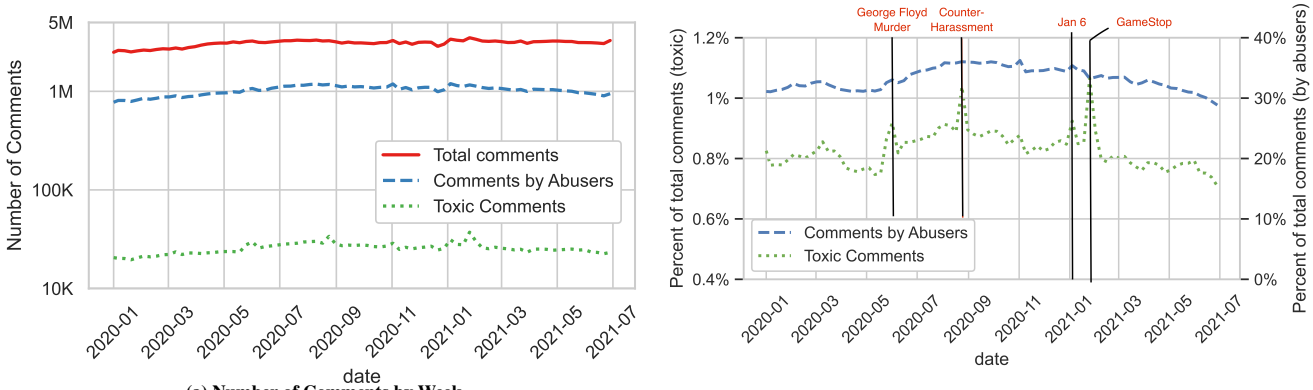
Table 2: Identifying Thresholds for Comments From Abusive Accounts—Reducing the toxicity threshold for comments posted by abusive accounts increases the volume of comments available to analyze while maintaining overall precision.

In many of these cases, the classifier's high toxicity score likely does not match the intention of these community norms, and including them as toxic comments would likely negatively tag benign accounts. As such, we chose to explicitly exclude subcommunities that are tagged as 18+ from our study. This filtering step removed 79M (3.6%) of total comments. We stress that while harassment may occur in these subcommunities, including them would cause unintended false positives and negatively affect the quality of our results, especially given our platform-wide approach to our measurements.

3.5 Aggregating abusive accounts

In the final stage of our methodology, we leveraged our corpus of labeled, filtered comments to identify "abusive" accounts (Figure 1.E). We categorized an account as abusive if it posted a comment with a `SEVERE_TOXICITY` score > 0.9, yielding our final dataset of 929K abusive accounts. However, by only considering comments from these abusive accounts that meet our strict threshold of 0.9, we significantly reduce the volume of comments available to analyze to just 1.7M (0.09%) comments.

Given this low comment volume, we next examined whether we could adopt a lower threshold for comments, conditioned on the



(a) Number of Comments by Week

(b) Fraction of Comments by Week

Figure 2: Toxic Behaviors on Reddit Over Time—Toxic comments are increasing slightly over time, and account for an average of 0.8% of all comments during our observation period. Toxic comments spike several times during our study period, typically in response to real-world events that spur significant discussion, like the murder of George Floyd in May 2020 or the January 6th insurrection of the US Capitol in January 2021.

account having posted at least one comment with high toxicity.[4] Our hypothesis was that if an account engages in toxic behavior, some of their other comments may also be toxic even if they do not meet our strict Perspective API threshold. To evaluate this hypothesis, we randomly sampled 200 comments from "abusive" accounts and 200 comments from "nonabusive" accounts at each of four severe toxicity thresholds: 0.5, 0.7, 0.8, and 0.9.[5] We then measured how well each threshold performed on our manual sample (Table 2).

For comments from accounts that did not post a toxic comment, all thresholds performed poorly, reaching only a maximum precision of 0.29 at a decision threshold of 0.8 (nonabusive accounts posted no comments higher than 0.9 by definition). For comments posted by known abusive accounts, performance was significantly higher, achieving a 0.72 precision at a threshold of 0.8, and a maximum precision of 0.79 at a threshold of 0.9. Based on this analysis, we expanded our corpus of toxic comments to include all comments—predicated on originating from an abusive account—that meet a threshold of `SEVERE_TOXICITY` score > 0.8. After this secondary threshold, our final dataset consists of 929K abusive accounts and 14M toxic comments that span 146K subreddits.

3.6 Limitations

We caution that our strategy for filtering and classification is not a perfect indicator of toxic behaviors on Reddit. We may omit some toxic behaviors due to false negatives, and given that our final per-comment precision is 0.72, we may also include some nonabusive comments in our final dataset. Furthermore, our analysis does not address fake accounts—such as sockpuppets—for which detection mechanisms are nascent and can also lead to errors [25, 45]. We stress that online hate and harassment is a novel problem, made additionally challenging by differing opinions on what constitutes toxic content [14, 24]. Our final precision numbers are consistent and at times stricter than prior research in this area [17, 18, 34, 38, 46], and, as we find, still provide significant signal for large-scale measurement analysis.

[4]We show how stricter thresholds (e.g., posting at least three highly toxic comments) has little effect on the results in Appendix A.

[5]For this experiment, two expert raters classified each comment as toxic or not using the definition provided by Google Jigsaw: "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." Raters agreed with a Cohen's Kappa value of 0.8. This avoided the stratified sampling bias present in the labeled Reddit corpus.

3.7 Ethical considerations

Without proper care, targets of abuse or the abusers themselves might be inadvertently harmed by our study. To mitigate these risks, we never interact with accounts, we never attempt to deanonymize receivers of abuse or the abusive accounts themselves, and we never report accounts to the platform due to the risk of unintended false positives. Furthermore, we note that our dataset is constructed off of an existing third-party source; we are only augmenting this existing dataset with toxicity labels using a standard approach (Perspective API) which is publicly available. We plan to release our labeled datasets to researchers by request, and we will remove any personally identifiable information (e.g., account names) before release.

4 RQ1: THE IMPACT OF ABUSIVE ACCOUNTS ON REDDIT

We begin with an aggregate analysis of the 929K abusive accounts and 14M toxic comments they post to Reddit. We study the volume of toxic content abusive accounts produce as well as their longitudinal toxicity behaviors.

4.1 Toxicity across the platform

Abusive accounts represent 3.1% of all accounts that post comments on Reddit. Despite their relatively small footprint, abusive accounts play an active and outsized role on the platform. Such accounts post 559M comments, which amounts to 33.3% of all comments posted to Reddit during our study period. Of these comments, 14M (2.9%) were toxic, and span a wide array of different attacks, like bullying, identity attacks, and threats. We provide a detailed breakdown of the types of toxic comments and the how they are distributed across subreddits in Appendix B and C. We note that while abusive accounts do post a large volume of comments, simply posting significant comments does not correlate with toxic behavior ($r = 0.01$, $p < 0.01$),

highlighting that the proclivity to post toxic comments is not simply a product of heavily using the platform. The harm caused by toxic comments is amplified when viewed by thousands of other users that engage with threads where the toxic comments appear. In total, 15M accounts (55.2%) participate directly in a thread where a toxic comment is posted, suggesting that toxic comments are not easily avoided.

4.2 Toxicity trends over time

Toxic comments are regularly posted to Reddit, accounting for between 0.75–1.05% of all comments in any given week (Figure 2). All abusive behaviors increased over time during our study period: in particular, the raw volume of toxic comments and their relative presence compared to non-toxic comments increased, which we confirm with a Mann-Kendall trend test ($p < 0.01$). The raw and relative volume of active abusive accounts also increased over time.

We note that there are four spikes in both raw and relative volumes of toxic behavior in our corpus. The first is between May 26 and June 5 2020, which we manually confirmed to be related to the murder of George Floyd [26]. The second spike of toxic comments occurred in August 2020, which was due to an automated counter-campaign against a bot that aimed to facilitate kinder language on Reddit. The third spike period occurred on January 6th, 2021, with toxic posts largely relating to the insurrection at the US Capitol.[6] The final spike in toxic comments came at the end of January 2021, and was directly related to the `r/WallStreetBets` takeover of the Gamestop stock on Reddit.[7] As a case study, we examine the first spike period in Appendix D.

5 RQ2: TOXICITY PATTERNS IN THE REPLY GRAPH

Toxic comments are rarely posted in isolation—the majority of comments (56%) are sent in reply to other comments, creating an underlying social structure that connects accounts to one another across the platform. Understanding these relationships can provide deeper insights into toxicity patterns, how toxic comments are operationalized, and ultimately how toxic comments are experienced by their receivers (i.e., those that receive a toxic reply in response to their own comment). We examine the structure of these relationships by studying the underlying reply-graph based on toxic interactions between abusive accounts and the account that receive abuse.

To study these latent relationships between accounts, we construct a toxic reply-graph which links participants if they interact directly with one another. Specifically, we build a weighted, directed graph $G = V, E, w$ where the vertices V are Reddit accounts and a directed edge $e \in E$ represents if an account posts a toxic response directly to another account. To capture interactivity between accounts, we also draw an edge if a receiver account replies to the toxic response. Edges are weighted based on repeated interactions between accounts.

[6]https://apnews.com/article/capitol-siege-police-riots-congress-c632472d5e11063611b4a902859d49fb

[7]https://www.theverge.com/22251427/reddit-gamestop-stock-short-wallstreetbets-robinhood-wall-street

5.1 Abusers and receivers

The toxic reply-graph G contains 1.8M vertices and 6.1M edges, of which 1.6M (89.3%) are receivers and 651K (36.3%) are abusers.[8] We focus our attention on the roles that accounts play and the behaviors they exhibit when engaging in toxic interactions.

Most toxic interactions are one-offs. The overwhelming majority of toxic interactions on Reddit are one-offs, meaning they occur one time between an abusive account and a receiver account and never occur again. Of the 4.1M abusive comments posted by an abusive account in response to receiver accounts, 89.6% are one-offs. Toxic interactions are thus often fleeting and one-off occurrences on the platform.

Abusive accounts play dual roles. The majority of toxic comments are sent towards accounts that do not post a toxic comment, who make up 71.3% of all receivers. However, we note that abusive accounts can play the role of both an abuser and as a receiver of abuse—28.7% of receivers are abusive accounts, and 460K (70.6%) of abusive accounts play both roles in G. This dual role also highlights an underlying challenge with defending against toxic content, as account-level interventions may inadvertently harm abusive accounts when they are also receivers of abuse.

Abusive accounts reply to toxic comments. Many abusive accounts engage in discussion with other accounts, especially when the interactions are toxic. As such, 47.4% of toxic edges have a *reciprocal* connection, which means the receiver of the toxic comment replied to the original comment. We count both toxic and non-toxic replies as a reciprocal edge. 421K (54%) of abusive accounts engage in a conversation where a recipient reciprocates. When abusive accounts respond to toxic comments, some (18.1%) will respond with a toxic comment (23.3% of interactions), potentially escalating the toxicity of the conversation with their reply. For example, when one abusive account posted:

> "I just partially agreed with you, can you not read? I just said if Trump knew about his acts then he should of spoken about it. There is no evidence that trump is a rapist and a pedo you fucking retard, present it me?"

The receiving abusive account replied:

> "He's literally admitted it multiple times. You pedo apologists are fucking sick. You're blocked. I don't talk to pedophiles or rapists or people fucked enough to try to lie for them."

Some abusive accounts have pre-existing relationships with abusive accounts. 40% of receiver accounts that posted toxic comments have a pre-existing, non-toxic relationship with their abusers. Excluding one-off abusive interactions, the majority (53%) of abuser receivers have a pre-existing relationship with their abusers, suggesting that toxic interactions between abusive accounts may be predicated by previous interactions on the platform. Such underlying network relationships may be useful in automatically *predicting* toxic encounters before they occur on the platform.

[8]We note this does not sum to 100% because some abusers are also receivers of abuse.

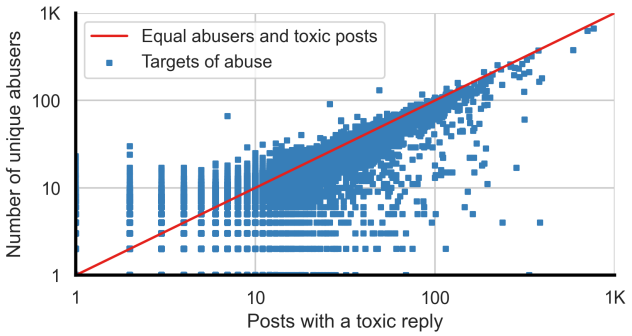


Figure 3: Receiver Experiences—Receivers experience three distinct types of toxicity: spurious abuse (e.g., a single abusive accounts replies to a single comment), repeated abuse (e.g., a small handful of accounts repeatedly harass a target through their posting history), or flooding (e.g., one comment triggers many toxic replies). The line at $y = x$ denotes an equal number of abuser and posts that trigger toxic replies. While 85.8% of receiver accounts fall on the line, 14.2% experience either flooding or repeated abuse.

5.2 Receiver experiences

Despite the fact that the majority of interactions between abusive accounts and receiver accounts are one-offs, many receivers experience many different types of toxicity. 59.5% of receivers experience just one abusive interaction from one abusive account in G, however, 40.5% of receiver accounts experience multiple abusive encounters during their time on the platform. To better explain these experiences from a receiver perspective, we measure receiver experiences by two criteria: the number of unique abusers that send toxic comments to the receiver and the number of posts that lead to a toxic reply (Figure 3). The line at $y = x$ indicates an equal number of abusers and toxic interactions. We identify three distinct toxicity patterns: spurious abuse, repeated abuse, and flooding. We detail each below:

Spurious Abuse. 76.1% of receiver accounts experience spurious abuse, meaning each toxic interaction that they encounter comes from a distinct abusive account. These accounts are those that fall on the line at $y = x$ in Figure 3. Such experiences have more to do with the content of the discussion rather than specific toxicity directed towards the receiver account. Protecting these receivers from unwanted toxicity is the most challenging, as they often have no prior relationships with their abusers and attacks may happen without any explicit warning signs.

Repeated Abuse. A significant number of receiver accounts (102K, 15.8%) experience *repeated* abuse, which are repeated toxic comments that come from the same abusive account (these accounts fall below the line at $y = x$). These are accounts whose comments regularly trigger a toxic reply, but are targeted by a smaller group of abusive accounts. Most alarmingly, abusive accounts seek out and repeatedly harass 5700 (0.5%) of receivers across *different* subreddits.

As an example of this type of abuse, we observe one account repeatedly harassing another account for their support of then US President Donald Trump across 20 distinct threads. The abusive account regularly antagonizes the receiver account for their beliefs and refers to the account in the second person, tacitly acknowledging the abuse:

> Yes, "news" must never be shared. Journalists must go out and find their own "news". Pffft, Are you fucking retarded? Do you know how stupid you sound? Of course, it's to be expected from you.

Repeated abuse has a distinct behavioral footprint from the majority of toxic interactions on Reddit. As such, it may be easier to defend against as it can be more readily identified on a platform level. Despite this, we are aware of no existing proactive protections for repeated abuse on Reddit.

Flooding. In contrast to those that experience repeated abuse, receiver accounts that fall above the line at $y = x$ in Figure 3 experience *flooding*, where a single comment may trigger many abusive interactions in reply. In our dataset, 53K (8.1%) receivers experience flooding, with the most flooded receiver experiencing 56 toxic replies to a single comment. In one such case, an account posted in the subreddit `r/wallstreetbets` expressing that those who were stymied by Robinhood (a popular trading platform) in a recent policy change should stop complaining. This comment was met with 52 distinct abusive accounts berating the author, insulting their intelligence, and in some cases, wishing for the author's death. These types of flooding attacks impact a significant number of receivers and happen almost in real-time, rendering post-hoc moderation limited in its effectiveness.

6 RQ3: CATEGORIZING ABUSIVE ACCOUNTS

Finally, we examine how distinct toxic behaviors of abusive accounts might be shared across accounts. We leverage these similarities to build *abuser personas*, which are groups of abusers that share behavioral traits. Such personas can help to inform more effective interventions to prevent abuse. We focus on three distinct toxicity behaviors that we then use to build personas: abusive accounts' toxicity behaviors on the platform in aggregate, their toxicity behaviors in the subcommunities they participate in, and finally, their behaviors in relation to subcommunity norms.

6.1 Toxicity behaviors in aggregate

Abusive accounts post a median six toxic comments during our study period and toxic comments make up a median 2.9% of all comments posted by abusive accounts. A small handful of accounts (1.3%) exclusively post toxic comments (100% of their comments are toxic), many of which are low-activity accounts, posting just a median five comments in total. When we consider highly active users (i.e., ones that post more than 100 comments), the top 5% of abusive accounts post only 10.7% toxic comments. In the most extreme case, 25K (4.2%) highly active abusive accounts post a single toxic-message during our observation period; the majority of their posting history is non-toxic. Conversely, 16K (2.6%) of abusive accounts post at least 100 toxic comments, and toxic comments account for at least 10% of comments for 18.7K accounts.

Given the rise in toxic comments posted to the platform in aggregate (Section 4), we also measure whether abusive accounts increase in their individual toxic behaviors over time. We compute a Mann-Kendall trend test for every abuser that posts at least a

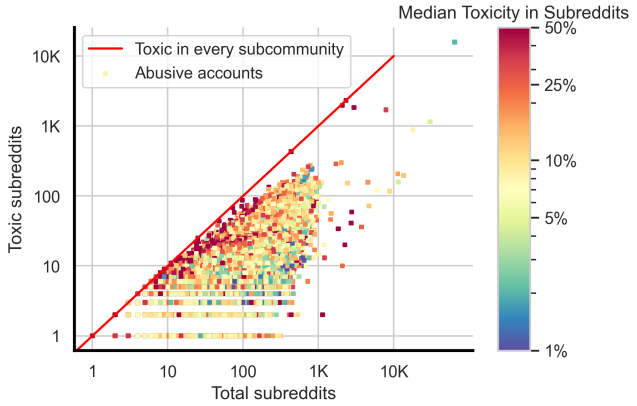


Figure 4: Abusers in Subreddits—Most abusive accounts restrict their toxic activity to a subset of their communities, however, 1% are highly toxic in nearly all the subcommunities they belong to. The gradient represents the median percent toxicity for each abusive account in their respective subreddits. Most abusive accounts fall into the yellow and blue portions of the graph, highlighting low overall toxicity in the majority of their communities.

single comment every week, and find that the majority of abusers (88.2%) exhibit no change in their toxicity behaviors over time. As such, toxicity behaviors are *stable* for most accounts and can form a foundation for building abuser personas.

6.2 Toxicity behaviors in subcommunities

Even if an abusive account is highly toxic during their lifetime on the platform, they are rarely abusive in *all* of the subcommunities they post in (Figure 4). Abusive accounts comment in a median 27 subreddits overall, but only post toxic comments in a median 13% of those subreddits. Abusive accounts may selectively choose when or where to be toxic on the platform due to a myriad of factors—for example, Cheng et al. found that seeing other trolling comments had an impact on trolling behavior [9].

However, this type of equivocation only accounts for a relatively small portion of toxic accounts' posting volume in each subreddit. Toxic comments make up a median of 12.5% of the comments they post in each of their subreddits. As such, abusive accounts may not only be selective in which subcommunities to post toxic comments in, but also in *how toxically* they behave in those communities. Figure 4 represents this idea as a gradient, where the intensity of the gradient (from blue to red) indicates an abusive account that posts significant toxic comments in *all* of their subcommunities. Most abusers fall between the blue and yellow portions of the graph (94%), which indicates they typically post only a small number of toxic comments in their subreddits. Yet, there are a small fraction (1%) of abusive accounts that choose to be highly toxic in the majority of their subcommunities (i.e., deep red points in the graph), again highlighting variance in abusive account behaviors.

Metric	Sub-Metric	Abuser Persona		
		Occasional	Moderate	Serial
Cluster	Size	350K (71%)	117K (24%)	21K (4.3%)
	Toxic Comments	71%	26%	3.1%
Activity	Comments	521	144	25
	Subreddits	63	22	3
	Social Homes	18	5	1
Toxicity	Agg. Toxicity	2.4%	7.2%	18%
	Tox Subreddits	11.4%	25%	50%
	Violat. Subreddits	12.5%	40%	100%

Table 3: Abuser Personas—Abusers fall into three distinct *personas* that capture their toxicity behaviors on the platform. Their distinct behaviors allude to nuanced interventions that may curb toxic behaviors on Reddit.

6.3 Violating subcommunity norms

Each subcommunity on Reddit is self-moderated and has its own set of unique norms about what type of discussion is allowed in the community. As such, posting toxic comments may not explicitly violate community norms, which a broad definition of toxicity (as we have been applying so far in this paper) may not appropriately capture. In order to study abusive behaviors in the context of community, we additionally consider how abusive accounts violate toxicity norms defined by Rajadesingan et al. [34].

We define a subreddit's toxicity norm as the fraction of toxic content posted in each subreddit. We restrict our analysis to subcommunities that have more than 50 comments in our dataset and to those that exhibit a "stable, distinctive" norm over our measurement period, which means the toxicity norm does not change beyond 2% from month to month and is distinct from the platform average (both thresholds were defined in prior work [34]). In total, we identify 35,840 subreddits with stable, distinctive toxicity norms for further analysis, of which 94.9% of subcommunities fall below and 5.1% fall above the platform-wide toxicity norm. We note that we only investigate a subreddit of an abusive account if they post a comment in that subreddit at least 5 times, which is denoted as a "social home" in previous work [10]. We do this to avoid counting single, spurious toxic comments in one-off communities.

We find that abusive accounts vary in terms of norm violations—abusive accounts violate the toxicity norms of a median 16.7% of their social homes. Despite this, a small handful of accounts (5.8%) violate the toxicity norms of every single subreddit they are a part of. When abusive accounts do violate subcommunity norms, they often do so to significant degrees—abusive accounts post a median 3.55x more toxic comments than the community standard. Such behaviors suggest that extreme norm-violations can serve as a broader signal to detect the most toxic accounts.

6.4 Classes of abusive accounts

We leverage our analysis of abusive behaviors to this point to ultimately build distinct abuser personas. We aggregate four key features of abuser toxicity behaviors:

(1) The fraction of comments that are toxic in aggregate.
(2) The fraction of subreddits that contain a toxic comment.
(3) The median fraction of toxic comments for each subreddit the abuser participates in.

(4) The fraction of subcommunities that each account violates a toxicity norm in.

We cluster abusive accounts using K-Means with Principal Component Analysis (PCA) for dimensionality reduction. We reduced to three components, as they capture the majority of the variance between each variable. We note that not every abusive account has explicit norm violation statistics (given only 35.8K communities have stable toxicity norms)—we exclude these abusive accounts and cluster the 489K (53%) remaining accounts. Abusers fall into three distinct personas, detailed in Table 3.

Persona 1: Occasional abusers 350K (71.7%) abusive accounts are occasional abusers, which are accounts that post relatively small fractions of toxic comments during their posting history (median 2.4%) and contribute 70% of all toxic comments posted to Reddit. These accounts post toxic content in a relatively small fraction of their subcommunities (11.4%) and tend to be regular, contributing members of the subcommunities they belong in. Indeed, these accounts are also the most active abusive accounts on the platform, posting a median 521 comments and participating in a median 18 social homes. While these accounts may have a proclivity towards toxic behaviors, they may also be the ones most amenable to nudges or other types of interventions, as toxic behaviors do not make up a significant portion of their overall platform behaviors.

Persona 2: Moderate abusers 117K (24.1%) abusive accounts are moderate abusers, which are accounts that post moderate amounts of toxic comments on the platform (median 7.2% toxic comments) and contribute 26% of all toxic comments posted to Reddit. Notably, these accounts are toxic in a larger fraction of their subcommunities, and violate the norms of a median 40% of communities they participate in. This is approximately 3.5 times that of occasional abusers but still a minority of their subreddits, highlighting that moderate abusers are selectively abusive in a handful of communities but not others. Curbing abuse from these types of accounts is challenging, as simple nudges are likely ineffective. Instead, more robust defenses that include subcommunity specific moderation practices may be most effective.

Persona 3: Serial abusers 21K (4.3%) of accounts are serial abusers, which is the least prevalent abuser persona and contribute just 3.1% of all toxic comments posted to Reddit. These accounts are *serial abusers*—they post a median 18.1% toxic comments, and post toxic comments in a median 50% of the subreddits they participate in. To make matters worse, they violate the toxicity norms of *every subcommunity they post toxic content in*, often entirely disregarding established toxicity norms. Encouragingly, these types of accounts are the least active type of abuser, limiting their existing impact on the platform. Still, given their proclivity to posting toxic content, simple interstitial defenses such as nudges may not be effective in curbing abuse from these accounts. Further abuse from these accounts can likely only be curbed through platform-level action (e.g., bans, suspensions).

7 DISCUSSION AND CONCLUSION

In this section, we synthesize our contributions into a set of open challenges and research directions for improving the automated detection of toxic behaviors and empowering community governance.

7.1 Abusive accounts contribute significantly to Reddit

Toxicity on Reddit accounts for a relatively small fraction of comments but are highly visible on the platform: 55.2% of Reddit accounts post directly on a thread with a toxic comment. Abusive accounts themselves make up just 3.1% of all accounts, but make up 33.3% of all comments to Reddit, which aligns with studies on other platforms; Hindman et al. describe the challenge on Facebook as a "superuser supremeacy problem [31]." The majority of these accounts engage in abusive infractions throughout their lifetime, but simultaneously contribute significant volumes of non-toxic content, ostensibly making them valuable contributors to the platform at large. As such, traditional strategies for dealing with abusive accounts (e.g., mass account bans) are likely untenable, as they would significantly reduce meaningful conversation on the platform and have potentially unforeseen consequences on platform health. Some platforms are attempting more nuanced actions. For example, Twitter deployed a strike system for handling accounts that post Covid-19 misinformation [4], and some subreddits have implemented similar strike systems for moderation at large [43]. Still, there is limited insight into how effective these strategies may be for handling online hate and harassment. The design of new defensive schemes and evaluating their efficacy for these types of attacks is a potential direction for future research.

7.2 Classes of abusive accounts can inform defensive design

Our identification of three distinct classes of abusive accounts (e.g., occasional abusers, serial abusers) and varied community norms suggests the need for targeted interventions, rather than a one-size-fits-all approach to actioning toxic behaviors. As we previously proposed, one-off attackers might benefit most from inline warnings or nudges. Chang et al. previously found that temporarily blocking Wikipedia contributors substantially reduced the rate of repeated abuse [7]. Likewise, Instagram now includes a feature that warns users before they post content that appears similar to previously reported hate and harassment [19]. However, the existence of moderate and serial attackers requires more serious interventions, up to and including suspension or permanent blocking. Chandrasekharan et al. previously showed that banning subcommunities can be highly effective [5]. However, the throw-away nature of accounts on Reddit may complicate applying such a strategy to individual attackers—though this limitation may not exist for all online social networks.

At the same time, some communities like `r/WallStreetBets` and `r/RoastMe` relish in offensive, profanity-laced discussions with other willing participants. Combined with varying personal definitions of what constitutes toxic content [14, 24], it is critical that platform designers consider empowering community-level moderators to best support conversational nuance online. However, it remains critical to enforce site-wide policies against hate and harassment, lest toxic subcommunities flourish that negatively impact other communities or users [5].

Ultimately, our measurements serve to better understand the dynamics of hate and harassment attacks in practice and inform nuanced interventions that may be most effective in curbing toxic content online.

8 ACKNOWLEDGEMENTS

The material is based upon work supported by the National Science Foundation under grant #2030859 to the Computing Research Association for the CIFellows Project. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers or the sponsors.

REFERENCES

[1] Max Aliapoulios, Kejsi Take, Prashanth Ramakrishna, Daniel Borkan, Beth Goldberg, Jeffrey Sorensen, Anna Turner, Rachel Greenstadt, Tobias Lauinger, and Damon McCoy. 2021. A large-scale characterization of online incitements to harassment across platforms. In *ACM Internet Measurement Conference*.

[2] Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Workshop on online abuse and harms*.

[3] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *International AAAI conference on web and social media*.

[4] Ian Carlos Campbell. 2021. Twitter will label COVID-19 vaccine misinformation and enforce a strike system. https://www.theverge.com/2021/3/1/22307919/twitter-covid-19-vaccine-labels-five-strike-system.

[5] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. In *ACM Conference on Computer-Supported Cooperative Work and Social Computing*.

[6] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. In *ACM Conference on Computer-Supported Cooperative Work and Social Computing*.

[7] Jonathan Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *The World Wide Web Conference*.

[8] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Measuring #gamergate: A tale of hate, sexism, and bullying. In *The World Wide Web Conference*.

[9] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *ACM Conference on Computer-Supported Cooperative work and Social Computing*.

[10] Srayan Datta and Eytan Adar. 2019. Extracting inter-community conflicts in reddit. In *International AAAI Conference on Web and Social Media*.

[11] Maeve Duggan. 2017. Online Harassment 2017. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/.

[12] Facebook. 2021. Transparency Center. https://transparency.fb.com/policies/community-standards/bullying-harassment.

[13] Sarah A Gilbert. 2020. " I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction* CSCW.

[14] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *ACM CHI Conference on Human Factors in Computing Systems*.

[15] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. @ spam: the underground on 140 characters or less. In *ACM conference on computer and communications security*.

[16] Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou. 2021. The Evolution of the Manosphere Across the Web. *International AAAI Conference on Web and Social Media*.

[17] Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020. Characterizing twitter users who engage in adversarial interactions against political candidates. In *ACM CHI Conference on Human Factors in Computing Systems*.

[18] Yiqing Hua, Thomas Ristenpart, and Mor Naaman. 2020. Towards measuring adversarial twitter interactions against candidates in the US midterm elections. In *International AAAI Conference on Web and Social Media*.

[19] Instagram. 2019. Our Progress on Leading the Fight Against Online Bullying. https://instagram-press.com/blog/2019/12/16/our-progress-on-leading-the-fight-against-online-bullying/.

[20] Ann John, Alexander Charles Glendenning, Amanda Marchant, Paul Montgomery, Anne Stewart, Sophie Wood, Keith Lloyd, and Keith Hawton. 2018. Self-harm, suicidal behaviours, and cyberbullying in children and young people: systematic review. *Journal of medical internet research*.

[21] Cristian Danescu-Niculescu-Mizil Jonathan P. Chang. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In *Empirical Methods in Natural Language Processing*.

[22] Matthew Katsaros, Kathy Yang, and Lauren Fratamico. 2022. Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content. In *International AAAI Conference on Web and Social Media*.

[23] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. *Proceedings of the ACM on Human-Computer Interaction* CSCW.

[24] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, and Michael Bailey. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. In *USENIX Symposium on Usable Privacy and Security*.

[25] Srijan Kumar, Justin Cheng, Jure Leskovec, and VS Subrahmanian. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*.

[26] David Leonhardt and Ian Prasad Philbrick. 2021. One Year Later. https://www.nytimes.com/2021/05/25/briefing/george-floyd-legacy-anniversary.html.

[27] Kirill Levchenko, Andreas Pitsillidis, Neha Chachra, Brandon Enright, Márk Félegyházi, Chris Grier, Tristan Halvorson, Chris Kanich, Christian Kreibich, He Liu, et al. 2011. Click trajectories: End-to-end analysis of the spam value chain. In *IEEE symposium on security and privacy*.

[28] Suman Kalyan Maity, Aishik Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2018. Opinion conflicts: An effective route to detect incivility in Twitter. *Proceedings of the ACM on Human-Computer Interaction* CSCW.

[29] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. In *ACM Conference on Computer-Supported Cooperative Work and Social Computing*.

[30] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction* CSCW.

[31] Nathaniel Lubin Matthew Hindman and Trevor Davis. 2022. Facebook Has a Superuser-Supremacy Problem. https://www.theatlantic.com/technology/archive/2022/02/facebook-hate-speech-misinformation-superusers/621617/.

[32] Shirin Nilizadeh, François Labrèche, Alireza Sedighian, Ali Zand, José Fernandez, Christopher Kruegel, Gianluca Stringhini, and Giovanni Vigna. 2017. Poised: Spotting twitter spam off the beaten paths. In *ACM Conference on Computer and Communications Security*.

[33] Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *International AAAI Conference on Web and Social Media*.

[34] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *International AAAI Conference on Web and Social Media*.

[35] Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *International AAAI Conference on Web and Social Media*.

[36] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Robert West. 2021. Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. In *ACM Conference on Computer-Supported Cooperative Work and Social Computing*.

[37] Erik Rye, Jeremy Blackburn, and Robert Beverly. 2020. Reading In-Between the Lines: An Analysis of Dissenter. In *ACM Internet Measurement Conference*.

[38] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In *The World Wide Web Conference*.

[39] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *International AAAI Conference on Web and Social Media*.

[40] Vivek K Singh, Marie L Radford, Qianjia Huang, and Susan Furrer. 2017. " They basically like destroyed the school one day" On Newer App Features and Cyberbullying in Schools. In *ACM Conference on Computer Supported Cooperative Work and Social Computing*.

[41] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. 2021. SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. In *IEEE Symposium on Security and Privacy*.

[42] Twitter. 2021. Rules enforcement. https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec.

[43] u/deliteplays. 2020. New Rule[0] and Strike System - please read before posting to avoid receiving bans. https://www.reddit.com/r/ProgrammerHumor/comments/bymrtt/new_rule0_and_strike_system_please_read_before/.

[44] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying women's experiences with and strategies for mitigating negative effects

of online harassment. In *ACM Conference on Computer Supported Cooperative Work and Social Computing*.

[45] Janith Weerasinghe, Rhia Singh, and Rachel Greenstadt. 2022. Using Authorship Verification to Mitigate Abuse in Online Communities. In *International AAAI Conference on Web and Social Media*.

[46] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. 2020. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction* CSCW.

[47] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *The World Wide Web Conference*.

[48] Justine Zhan, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of ACL*.

[49] Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020. Quantifying the Causal Effects of Conversational Tendencies. In *ACM Conference on Computer-Supported Cooperative Work and Social Computing*.

Comment Threshold	# Abusive Accounts	Precision
1	931K (2.9%)	0.71
2	332K (1%)	0.68
3	183K (0.5%)	0.68
4	119K (0.4%)	0.7
5	84K (0.3%)	0.66

Table 4: Abusive Account Thresholds—Increasing the number of toxic comments required to label an account as an abusive account does not demonstrably improve performance, while significantly reducing the account and comment volume available to study.

Category of Attack	% Comments	Std. Error
Insult	63.4%	2.2%
Identity Attack	14.2%	1.6%
Call to Leave	12.0%	1.5%
Threat	5.5%	1%
Sexual Aggression	2.8%	0.7%
Identity Misrepresentation	1.6%	0.6%
Doxxing	0%	
Targeted Toxicity	44.8%	2.2%
Generalized Toxicity	55.2%	2.2%

Table 5: Types of Attacks on Reddit—Attacks fall largely into two categories—attacks on authors or attacks on nonauthors. Attacks on authors are largely insults or calls for the participant to leave the community, whereas nonauthor attacks focus on larger identities (e.g., racial, political, etc.) but are not explicitly against the author of a post or comment.

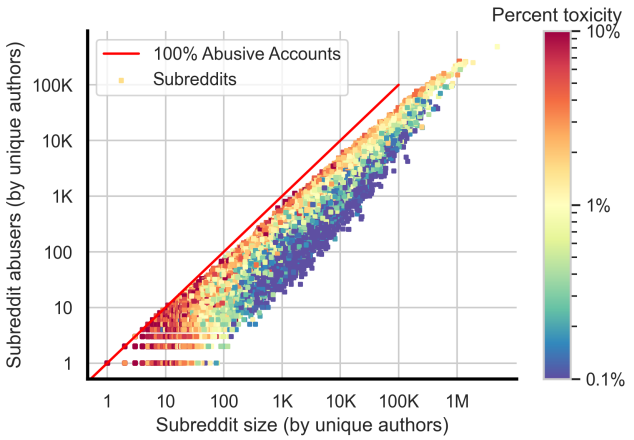


Figure 5: Subreddit Hotspots—We show subreddits by the number of abusive accounts as well as the size of the subreddit. Some subreddits (i.e., those in the red band) serve as hotspots for toxic behaviors, hosting both high fractions of abusive accounts and large amounts of toxic content.

APPENDIX

A EVALUATING ABUSER THRESHOLDS

For the scope of this study, we consider an account to be abusive if it posts just a *single* comment above the high precision threshold of `SEVERE_TOXICITY` > 0.9. However, we also evaluated whether increasing the number of high-precision toxic comments required to label an account as abusive could in turn increase our overall precision. To measure this, one expert rater manually sampled high-precision toxic comments from accounts that posted 1–5 toxic comments, and we evaluated the resultant precision to see if changing the threshold for abusive comments would increase our results. Table 4 shows the results. Ultimately, precision was stable for all samples at 0.7, suggesting that increasing the threshold would not result in higher quality data while also reducing the size of the abusive account population to a tenth of the size in the most extreme case.

B TOXIC COMMENT BREAKDOWN

We manually investigated a random sample of 500 toxic comments, which we coded into several categories of toxic behavior. We labeled each comment based on hate and harassment categories identified in prior work [2, 24]: doxxing, identity attacks, identity misrepresentation, insults, sexual aggression, threats of violence, and profanity. We added one additional category we find particularly prevalent on Reddit, a "call to leave conversation", which typically involves the attacker telling the target to leave the conversation, subreddit, or subcommunity. We excluded comments from our analysis that were not relevant (e.g., a false positive or general negative sentiment). Table 5 shows the breakdown of attacks per category.

The majority of attacks on Reddit are insults (63.4%), which are typically provided in response directly to a previous commenting account or a reply to the original poster themselves. For example, in the subreddit `r/ShitLiberalsSay`, a community designed to mock liberal opinions, one account wrote:

> "I don't know what's more salty. Your mouth or your asshole. Not like there is a big difference between them in your case, diarrhea and entitlement come out of both ends, and they're both just as pathetic."

Attacks also fall into a host of other categories, including identity attacks (14.2%), calls to leave the conversation (12%), and threats of violence (5.5%). One example of a call to leave is:

> "...Get the fuck off this subreddit. You clearly don't really care about how severe NVLD can be. There's literally no fucking help out there for us."

We did not observe any instances of doxxing in our analysis. However, this is likely due to our labeling and sampling mechanism, as well as our limited manual sampling (only 500 comments). More nuanced attacks like doxxing may need finer grained tools for identification [1].

C TOXICITY PER INDIVIDUAL SUBCOMMUNITY

A significant number of subreddits are affected by abusive accounts and toxic comments: 146,831 (63.4%) subreddits have participation by abusive accounts, of which 51K (22%) contain at least one toxic comment. These subreddits are typically the ones with the most activity, and 100% of highly active subreddits (i.e., more than

Metric	Control	George Floyd Incident
Toxicity Volume	68K (0.7%)	91K (0.9%)
Abusive Accounts*	332K (15.4%)	347K (15.7%)
Spurious Receivers	17.5K (95.1%)	23.2K (93.9%)
Repeat Receivers	575 (3.1%)	907 (3.7%)
Flooded Receivers	315 (1.7%)	584 (2.4%)

Table 6: Toxicity in response to the murder of George Floyd—The fraction of toxic comments increased by 30% on Reddit in response to the murder of George Floyd. This had downstream impacts on receiver experiences, which slightly skewed away from spurious toxic interactions and towards more impactful forms of abuse, like repeated attacks and flooding. Results are statistically significantly different between the two groups except when denoted with an asterisk.

100K comments) contain toxic comments. Figure 5 shows the fraction of abusive accounts in subreddits, with the gradient representing the fraction of comments in those subreddits that are toxic and the line at $y = x$ indicating the subreddit consists solely of abusive accounts. Subreddits that fall into the red band are *hotspots* of toxic activity, which tend to be smaller subreddits that contain tens or hundreds of unique accounts (i.e., the bottom left corner of the graph.) Such subcommunities tend to serve a niche user base (e.g., `r/FortniteBad`, which is a subcommunity where accounts share hateful memes targeted towards the video game Fortnite), and in 1% of cases, consist entirely of abusive accounts. However, even some highly active subreddits can consist of upwards of 30% toxic content, highlighting that even outside of hotspots of toxic activity, toxic comments are a potentially pervasive part of Reddit.

D CASE STUDY: THE MURDER OF GEORGE FLOYD

In our longitudinal analysis, we observed several spikes in toxic behaviors on Reddit. One such spike occurred between May 26th, 2020 and June 5th, 2020, which reached a peak during a three day period between May 29th and May 31st, 2020. This spike was primarily in response to the murder of George Floyd. In this section, we detail the impact of this real-world event on toxic interactions on the platform in aggregate, on abusive account behaviors, and on receiver experiences. To do this, we compare behaviors from this spike period against a 3 day sample of the dataset collected from May 1st, 2020 to May 4th, 2020 as a control.

D.1 Changes in abuser behaviors

During the peak of toxic behaviors, 0.9% of all comments on the platform were toxic, which marks a 30% increase in overall average toxicity from the control period (Table 6). Overall abuser activity (e.g., number of comments, number of subreddits) stayed consistent throughout the control period and spike period, suggesting that the increase in toxic comment volume was not directly related to a significant number of new abusive accounts becoming active during this period. Rather, we observe that 90.2% of abusive accounts posted either the same volume or more *toxic* comments during the spike period compared to the control period, which contributed overall to an increased period of toxicity throughout the platform.

We observe no changes in the *structure* of toxic interactions (e.g., those discussed in Section 5), suggesting that such behavioral patterns remained consistent even when abusive accounts increased their volume of toxicity. Despite this, receiver experiences shifted slightly during the spike period. While interactions largely remained spurious, the number of receivers with solely spurious interactions decreased from 95.1% to 93.9%, and instead shifted towards more impactful forms of abuse, like repeated abuse (3.7%) and flooding (2.4%). This skew towards repeated abuse and flooding was largely in discussions of the George Floyd incident. For example, one account posted in `r/PublicFreakout`:

> "Now those *same exact people* are defending what these fascist pigs are doing."

The comment was met with 6 different attacks berating them for their comment and insulting them. The slight shift in the types of attacks that receivers experienced during the spike of toxicity may anecdotally suggest that the types of toxic interactions may increase in intensity during heated discussion of real-world events.

D.2 Subcommunity spread

Despite abuser behaviors remaining relatively consistent, we observed that the subcommunities where toxic behaviors took place changed during the spike period. Many large subreddits that were closely discussing the incidents as well as the resulting protests boomed in posting volume. As an example, `r/PublicFreakout`, which is a community designed for discussing videos of "people freaking out, melting down, losing their cool, or being weird in public"[9], saw its comment volume increase by 620% and its toxic comment volume increase by 670% during the spike. This also impacted many smaller subcommunities—for example, the subreddit with the largest change in toxic comment volume (9566% increase in toxic comments) was `r/Minneapolis`, which is where the murder of George Floyd took place. Other communities with a stark increase in toxic comments were other cities where protests were taking place, for example, `r/philadelphia` (3720% increase), and `r/cincinnati` (3000% increase).

In all of these cases, we observe that many toxic comments are posted by accounts that *never post in the subcommunity prior to the event*. 569 (18.2%) of the accounts that posted toxic comments in `r/PublicFreakout` never posted in this subreddit prior. A similar result holds true for `r/Minneapolis` (26.2% new members), `r/philadelphia` (24.5%) and `r/cincinnati` (10.1%), suggesting that at least some fraction of an increase in toxic content in these subcommunities comes from *outsider* accounts, likely joining these subcommunities to discuss ongoing incidents and post inflammatory content. As an example, one new account that joined `r/Minneapolis` posted inflammatory responses to accounts talking about the ongoing protests. In one instance, they wrote:

> "hmm, lets ruin people's businesses, earnings they have worked for to feed their family and shit, or work that they have put years into being ruined. yeah you are a bunch of fucking retards, all of you need to be executed"

Such examples highlight that some abusive accounts may actively seek out contentious discussion and participate in a toxic manner during known real-world events on the platform.

[9] https://www.reddit.com/r/PublicFreakout/