Hater Is Not the Only Source of Toxic Online Communication, But Also Fan

Akira Matsui[1]*, **Taichi Murayama**[1], **Mitsuo Yoshida**[2]

[1] Yokohama National University, Japan
[2] University of Tsukuba, Japan

Abstract

Online platforms often witness harmful communication. While the previous research has mainly focused on the linguistic aspect of such communication, the role of the authors' stance toward the discussed topic remains unclear. To address this gap, we first construct a unique dataset from an online platform containing self-disclosure annotations about authors' stances. Our analysis of the constructed dataset indicates that anyone can be a target of toxic communication, whether a real person or a virtual entity. We, then, demonstrate that toxic communication is not limited to haters, but also fans. Writing toxic comments does not necessarily indicate a negative attitude towards the subject or individual being discussed.

1 Introduction

Online platforms have significantly transformed our daily communication, but we often face abusive online communications. This prevalence of abusive behavior demands computational techniques to detect the patterns of such toxic online communication (Kowalski, Limber, and McCord 2019; Giumetti and Kowalski 2022; Avalle et al. 2024).

However, only linguistic features do not fully capture such complex form of communication because a highly toxic comment about a particular individual does not always guarantee the author's intention to target that individual; rather, it may be a defensive response (Edwards, Kontostathis, and Fisher 2016). For instance we often find partisans using highly offensive language to defend their preferred party's candidate in political discussion (Kalmoe 2014; Faizah, Mustika, and Prasetyo 2019; Lyu 2023). Despite this, most studies focus on the toxicity of comments within online discussions and pay little attention to the stances of authors toward topics.

Interactions among two groups with opposed views can increase their positional distance in discussion. Particularly, in political discussions, disagreements between opposing groups can escalate the intensity of discussions. Bail et al. (2018) show that exposure to opposing political views on Twitter amplifies polarization, highlighting the challenge in mitigating online political polarization. This escalation of political stances can be facilitated by interactions between the two with opposing views. However, it remains uncertain whether this phenomenon extends to all discussions, particularly those involving differing preferences or conflicts over various issues. Moreover, as Avalle et al. (2024) point out, such disagreements do not always result in escalating debates, and we little know what kind of groups or interactions exacerbate the situation.

While the literature on toxic behavior in online platforms often focuses on specific characteristics of individuals (Chandrasekharan et al. 2017; Rieger et al. 2021), any entity can become a subjects of highly toxic comments regardless of their traits. This suggests that personal distinctive traits alone do not fully elucidate why anyone can be a victim of toxic communication. For instance, terms like "idiot" or other aggressive synonyms are prevalent in abusive language, but such language can point to any individual or entity. Even fictitious entities, such as cartoon characters or virtual streamers, can become targets of hurtful comments in the online sphere (Boi 2015; Wikipedia 2024). An excessive focus on specific instances can veil toxic communication on online platforms.

To bridge the aforementioned research gaps, we conduct an extensive analysis of toxic discourse on an online platform that mandates users to disclose their stance towards the subject of discussion before commenting. These self-disclosure annotations enable us to categorize comments as either written by fans or haters of target individuals in forums. The website covers a wide range of genres, from famous actors to fictional characters, and user discussions often take the form of replies. To make the presented paper easy to follow, we introduce several terms to facilitate discussion, which are summarized in Table 1.

Leveraging this unique dataset, we address the following research questions:

- **RQ1**: Do toxic comments share a common pattern regardless of the target, or do they vary based on the target's characteristics?
- **RQ2**: Do haters predominantly drive toxic communication in online fandoms, or can fans be contributors as well?

By answering these questions, we provide the follow-

*Corresponding author. E-mail adress: matsui-akira-zr@ynu.ac.jp
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

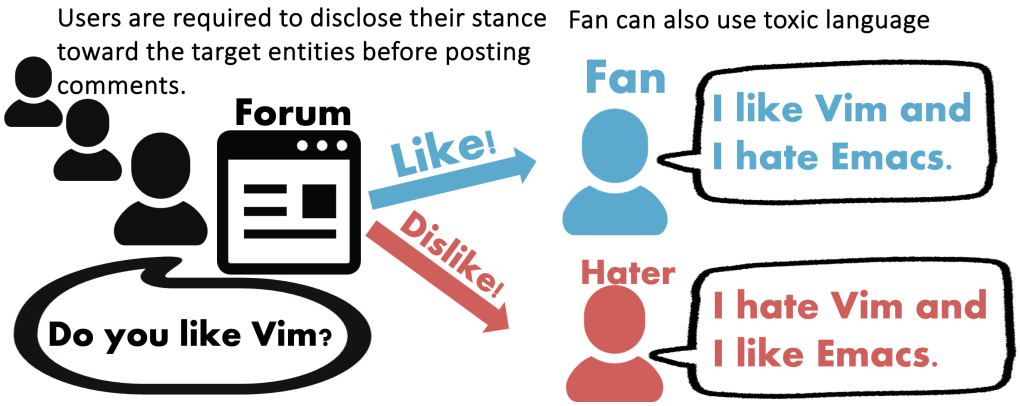

Figure 1: Schematic of the dataset with a mock example. This schematic is created to avoid presenting harmful content in the paper to be presented in public. In real scenarios, the forums on the platform are related to entities like celebrities, athletes, or virtual characters. The schematic serves as a conceptual illustration of user interactions and stance declarations within the platform's context.

Table 1: Glossary of Terms

Term	Description
Target	The subject of cyberbullying, encompassing tangible individuals, virtual entities, or fictional characters.
Fan	An individual who reveals supports the target on the platform.
Hater	An individual who reveals supports the target on the platform.

On the platform, users are required to state their stance towards the target. Through these self-declared annotations, we can clearly categorize users as either Fans or Haters.

ing insights into toxic communication in online platforms. First, RQ1 reveals that although the toxicity of comments varies across target categories, certain categories witness toxic comments from fans, not just haters. Our subsequent case study shows that virtual targets receive a comparable level of toxic comments to non-virtual ones, albeit with a distinct composition. Additionally, in RQ2, we discerned the scenarios in which fans write toxic remarks, a situation influenced by the friction between fans and haters.

2 Related Work

In this section, we review the related work of this study, mainly focusing on toxic communication on online platforms. The advent of online platforms has transformed our means of communication, but toxic communication has been a serious issue. To combat this dark side of the online community, many researchers have investigated efforts to develop robust methods to keep the community safe and positive (Almerekhi, Jansen, and Kwak 2020). Such interventions can improve user experience on online platforms, for example through safe ranking algorithms (Garcia-Pueyo et al. 2023). However, in terms of communication, online interactions take multifaceted forms, and there is a potential disconnect between authors' intent and readers' interpretation (Chang, Cheng, and Danescu-Niculescu-Mizil 2020).

Among many forms of toxic communication on online platforms, cyberbullying has been gathering a large amount of attention from not only researchers but also school officials, policymakers, and parents in general (UNICEF; Gradinger, Strohmeier, and Spiel 2010). Cyberbullying tends to take the form of a sequence of comments rather than a single message (Yao, Chelmis, and Zois 2019). On cyberbullying, various individuals can be victims of these malicious acts (Kowalski, Limber, and McCord 2019; Giumetti and Kowalski 2022), while specific individual traits are also being targeted (Smith et al. 2019; Li 2006). Therefore, it is crucial to capture the sequence of toxic communication about various topics to understand cyberbullying.

There has been research on hate speech on online platforms (Castaño-Pulgarín et al. 2021), and the literature has developed computational models to detect and mitigate hate speech using machine learning models (Nobata et al. 2016; MacAvaney et al. 2019; Bourgonje et al. 2018). Not only hate speech detection, such computational method is applicable to a wide range of offensive language communication (Pradhan et al. 2020; Bertaglia et al. 2021). The detection and mitigation of abusive language on online platforms also benefit to online communities such as Wikipedia (Cheriyan, Savarimuthu, and Cranefield 2021; Smirnov, Oprea, and Strohmaier 2023) or Stack Overflow (Kumar et al. 2023), Reddit (Kumar et al. 2023; Hebert, Golab, and Cohen 2022; Hiaeshutter-Rice and Hawkins 2022). Regarding this line of the literature, it is worth noting that Ashraf, Zubiaga, and Gelbukh (2021) indicate that understanding not only the textual but also the contextual elements is important for comprehending hate speech.

Therefore, to understand the mechanism of the toxic phenomenon on online platforms, it is crucial to comprehend the interactions among users because not only negative users play a pivotal role in such form of communication (Kayes et al. 2015; Avalle et al. 2024). Even if a very toxic communication occurs within a community or section of an online platform, it does not necessarily imply we find toxic com-

munication in other part on that platform (Caines et al. 2018; Mohan et al. 2017). In addition, the user interactions especially conflict between users plays a pivotal role in online communication (Levy et al.; Avalle et al. 2024).

In conclusion, the spread of toxic communication in online realms presents several challenges, particularly in understanding the role of users' stances toward topics and the sequence of communication. Additionally, the role of tensions among users have not been fully explored. Moreover, while most studies focus on specific factors or personal traits in understanding abusive language usage online, few investigate the banality nature of toxic communication and the reality that anyone can be a victim of such malicious acts.

3 Data

In this section, we explain the dataset employed in the presented study using the terminologies described in Table 1. This study constructs a distinct dataset comprising online fandom communities where authors openly reveal their stances toward the subjects of discussion.

3.1 Overview of the Platform

The dataset employed in this manuscript originates from an online forum[1]. The platform primarily centers on Japanese celebrities, enabling visitors to gauge the popularity of these celebrities. Notably, the platform includes a wide range of entities from different domains, such as actors, movie stars, and politicians. This diverse set of entities ensures the representation of the dataset without bias towards any specific genre. In this paper, we refer to the individuals receiving comments in the described format as *"target"*.

3.2 Like or Dislike, Fan or Hater

The crucial functionality of the platform is that it requests users to disclose their stance toward the target in a binary form, indicating whether a given user is a fan or a hater of the target. Users who have commented on the target are assigned labels that reflects their preference, either liking or disliking the target. Consequently, all users are aware of their own stance toward the target when writing posts, including responses to others' posts. We refer to the participant's vote "like" toward the target person are *fan* of them, and those who cast "dislike" vote for the target person are *hater* of them. We retrieve this self-reveal stance of each user and categorize whether each post is written by fan or hater.

The platform also has a function that displays the tension between fans and haters in each forum of a target, showing the number of fans and haters of that target. The ratio of fans to haters in a forum plays a pivotal role as it reflects the discourse's intensity on the forum. When the majority of users participating in a forum are fans, that forum is fan-driven. When dominated by haters, that forum is hater-driven. In cases where there is a well-balanced distribution of posts between these two groups, it signifies a forum where they are in competition. We will employ this information to explore whether the tension between these two groups is associated with the prevalence of toxic behavior on a forum platform.

[1]https://suki-kira.com

3.3 Tags and Categories of Targets

In addition to the comments and stances of users, the platform incorporates category tags associated with each target, thereby facilitating comparisons across the targets' spheres of influence. We present the top 30 frequent categories in Table 2. The table shows a wide range of categories of targets persons, and as discussed in the introduction we see some categories that represent non-human targets such as Character or V-tuber. After collecting data, we consequently have amassed approximately 83,004 comments pertaining to 4,700 users, strictly keeping users' anonymity.

4 Methods

In this section, we will outline the methods employed to analyze toxic communication on the platform. We first discuss the use of the Perspective API and provide a brief overview of the regression model utilized in this study.

To classify the toxicity of individual comments, we leverage the Perspective API [2]. This API offers an array of prebuilt toxicity classifiers, and its reliability and validity have been established in numerous studies (Hua, Naaman, and Ristenpart 2020; Kumar et al. 2023; Saveski, Roy, and Roy 2021; Xia et al. 2020; Rajadesingan, Resnick, and Budak 2020) [3]. Given the API's capability to process a broad range of languages, from Japanese to English, we believe our results possess a certain degree of generalizability.

The Perspective API analyze and score abusive language in text through six distinct metrics: Toxicity, Severe Toxicity, Identity Attack, Insult, Profanity, and Threat. This paper is interested in toxic behavior, therefore we conduct the analysis with a primary emphasis on the Toxicity Score, following by related studies (Hua, Naaman, and Ristenpart 2020; Saveski, Roy, and Roy 2021; Xia et al. 2020; Rajadesingan, Resnick, and Budak 2020; Avalle et al. 2024). We incorporate the other metrics to elucidate the distinctions between fans and haters. The comprehensive descriptions of these metrics are available on the Perspective API's documentation page[4].

We investigate the relationships between the scores provided by the Perspective API and attributes such as the categories of target persons. Also, we study whether each comment is written by a fan or a hater of the target. We employ a linear regression model to estimate the relationship between these attributes and the toxicity of the comments.

5 Results

This section discusses the results of the analysis to answer the research questions we proposed in the introduction.

[2]In the document page, the Perspective API defines Toxity as *"We define toxicity as a rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion"*(https://perspectiveapi.com). We follow this definition in this paper.

[3]https://perspectiveapi.com

[4]https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?

Category Tag Name	Description	Category Tag Name	Description
Female Model	Female Model.	Male Idol	Male performer in pop groups.
Stage Actor	Actor in theatrical plays.	Idol	Performer in pop groups.
Host/MC	Introduces guests and manages events.	Member of Parliament	Representative in the parliament.
Child Actor	A child, especially in films and plays.	Talent	Celebrity in entertainment industry.
Virtual YouTuber (VTuber)	A YouTuber with a CG avatar.	Male Voice Actor	Male who provides voice-overs.
TV Station Announcer	An announcer for a TV station.	Character	Person in a work of art.
Singer-Songwriter	Musician who writes/performs songs.	YouTuber	Manages content on YouTube.
Businessperson	Manages commercial activities.	General/Comprehensive	Relating to the whole.
Singer	Produces musical tones with voice.	Actor	Person who acts in films, stage shows.
Comedian	Makes people laugh by jokes or actions.	Female Idol	Female performer in pop groups.
Politician	In professional politics, elected office.	Idol Groups†	Idol groups under talent agency.
Infielder (baseball)	An inner field player in baseball.	A VTuber Group†	VTubers under Nijisanji company.
Figure Skating	Skating in prescribed patterns.	Actress	Female who acts in films, stage shows.
Lyricist	Writes the words to songs.	Game Commentator	Plays games with commentary.
Female Voice Actor	Female who provides voice-overs.	Baseball	Sport played by two teams of nine.

Table 2: Description of Entity Category Tag. The table that delineates the entity category labels for the analysis depicted in Figure 2. These entity category labels serve to categorize and identify entities within the dataset. In order to refrain from directly referencing specific individuals or persons, we have anonymized the names of such groups with the symbol "†".

5.1 RQ1: Patterns in Toxic Comments Based on Target Characteristics

We first answer the RQ1 "Do toxic comments exhibit common patterns regardless of the traits of targets, or do they vary depending on targets' attributes?" To explore this, we calculate the toxicity of comments across various categories, as outlined in Table 2. We then compare the differences in toxicity of comments among the categories. In this analysis, we take into account the users' stances toward the target, distinguishing between fans and haters.

Categorical Differences in Toxicity We calculate the differences in toxicity across the categories and plot them in Figure 2. The values presented in the figures represent deviations in toxicity from the average of the other categories. Therefore, the estimated value indicates the deviation of toxicity from the average, and those values demonstrate what categories receive more toxic comments by which type of users (i.e., fan or hater). A positive value implies that subjects within that category tend to attract more toxic posts compared to others, and vice versa.

The figure also indicates some variations in its toxicity levels with regard to the stance towards targets. As anticipated, haters tend to write more toxic comments compared to fans. However, some categories receive higher toxic comments from fans, such as in the case of Idol Groups. Additionally, we observed that there are categories where both fans and haters employ highly toxic language directed at the target, for example, the politician category. These observations imply that even fans can engage in toxic communication in specific contexts.

5.2 Case Study: Virtual vs. Non-Virtual Targets

We have discerned a link between target categories and the usage of toxic language and, we find that virtual entities can also receive toxic communication. To comprehend this interesting phenomenon, we conduct a deeper analysis to unveil why even virtual entities can receive toxic comments. For this aim, we conduct a case study contrasting extreme categories: virtual versus non-virtual targets. The rationale behind this case study is twofold. First, the recent work highlights the harassment in metaverses even if they are in their virtual avatar personas (Wiederhold 2022; Ortiz 2022). Second, even if targets are non-human, prevailing offensive language in online form can distort the safety of the online community. In addition, such heated phenomenon can be recognized as a social problem (St. Michel 2015)

We examine the targets classified as either Characters or VTuber categories. VTubers, or Virtual YouTubers, are online personalities who employ digital avatars, frequently powered by live motion-capture technology. They create content and engage with their audience on YouTube, gaining significant popularity in recent year (Lu et al. 2021; Bredikhina et al. 2020).

We begin this case study with comparing the distribution of toxic comments between these two categories and plot the results in Figure 3 where we observe subtle differences among the two. Although we find lower overall toxicity levels when considering various facets in aggregate, we still observe that virtual targets receive toxic comments as real persons do.

While both categories experience toxic comments, we investigate the contents of such comments and try to find the differences between them. To address this, we calculate the similarity among the toxic comments. We focus on the comments whose toxicity scores are in the top 75th percentile and create a set comprising the top 1000 words from these comments. We then calculate the Jaccard similarity among those vectors. Also, we construct vectors with normalized word frequencies and calculate the cosine similarity among them. We employ this analysis for both the Virtual and Non-virtual categories, as well as for Fans and Haters for comparison purposes. We present the reults in Figure 4, and observe that toxic comments directed at virtual targets consist of distinct words compared to the other groups (Figure 4(a)). This

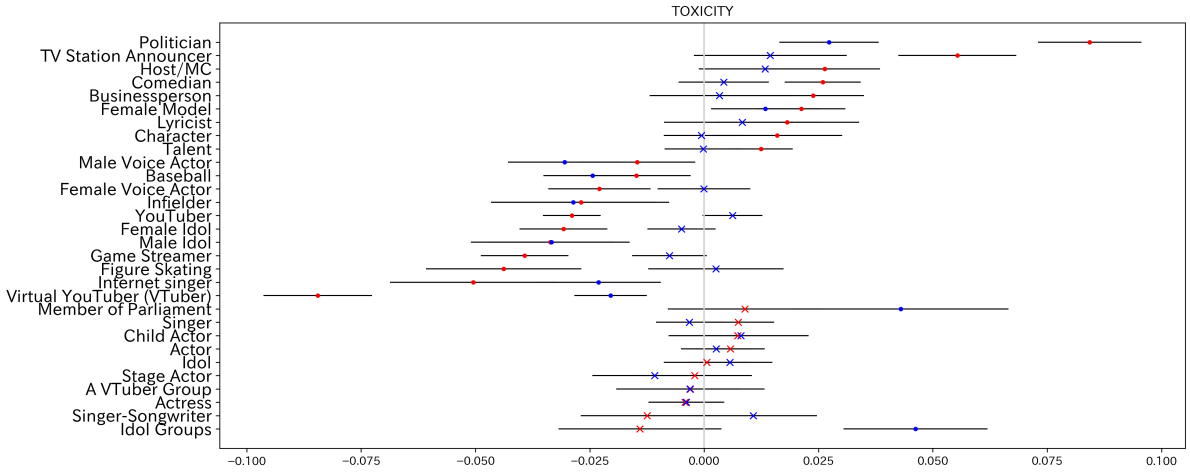

Figure 2: Toxicity Across Different Categories

Note: The estimated variations in toxicity across different categories for Hater in red color and "Fan" in blue color; • indicates p-val< 0.01, and otherwiwse $\times$; Bars represent 95% CIs. The figure illustrates that in most categories, Haters tend to post more toxic comments than Fans. However, there are certain categories in which Fans write comments with either higher or equivalent levels of toxicity as compared to Haters.

pattern is consistent when examining cosine similarity (Figure 4(b)). However, the consistent findings in the analysis of normalized vectors (cosine similarity) imply that toxic comments directed at virtual targets manifest differently than those directed at other subjects.

While we anticipate common words related to toxic language across all groups, the findings above suggest that even within toxic comments, there exists a degree of diversity in their contents. On the observation of the fan, we find relatively lower values(Figure 4), which suggests that fans also employ distinct words in their toxic comments. These disparities indicate the possibility of predicting attitudes toward the target (Fan/Hater) based on textual analysis.

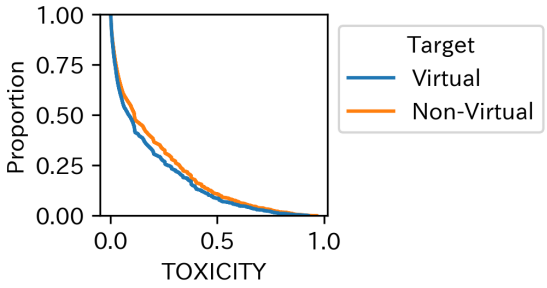

Figure 3: Toxicity Distribution: Virtual VS. Non-virtual

5.3 RQ2: Toxic Communication: Haters vs. Fans in Online Fandoms

Next, we aim to address the RQ2: "To what extent do individuals with negative attitudes (haters) dominate toxic communication within online fandoms, and can fans also contribute significantly?" Our investigations in RQ1 underscore

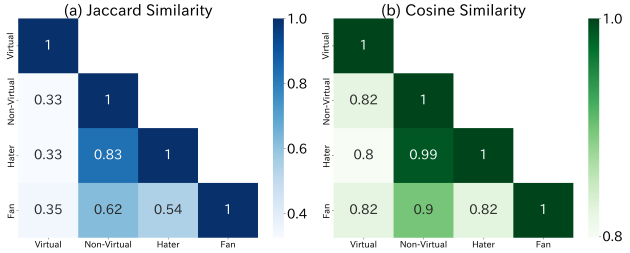

Figure 4: Comparative Analysis of Toxic Comment Content. The comparison involves comments authored by four distinct groups: comments on Virtual entities and Non-virtual entities, composed by Haters and Fans.

the role of user stance (fan or hater) towards targets and observed that fans can also write toxic comments. To convey a better understanding of this phenomenon, we direct our focus toward the context in which they employ toxic language.

Distribution of Toxic: Fans VS Hater To gain a comprehensive understanding of the distinctions between fans and haters, we have plotted the CCDF of toxicity scores in Figure 6. The figure illustrates that haters tend to compose more highly toxic content than fans, which aligns with our expectations. All four measurements from the API yield similar results.

While haters write more toxic, fans may also write at the same level as haters in specific contexts. To study this point, we focus on the distribution of scores more than 90 percentile of each score measurement, and we reveal different patterns as depicted in Figure 8. While the distribution of the toxicity scores still indicates that haters generate more toxic comments than fans, we observe that both fans and haters exhibit similar distributions in identity attacks and insults

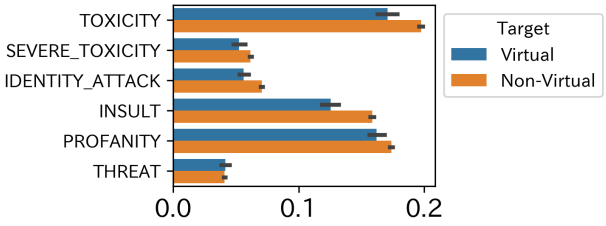

Figure 5: Detailed Comparison of Toxicity Measurement: Virtual VS. Non-virtual. The figure provides a detailed representation of toxicity measurements for each group (Virtual vs. Non-virtual) along with 95% confidence intervals (represented as bars).

scores (the bottom two categories in Figure 8). This finding supports the notion discussed earlier that fans can also contribute toxic comments under specific circumstances.

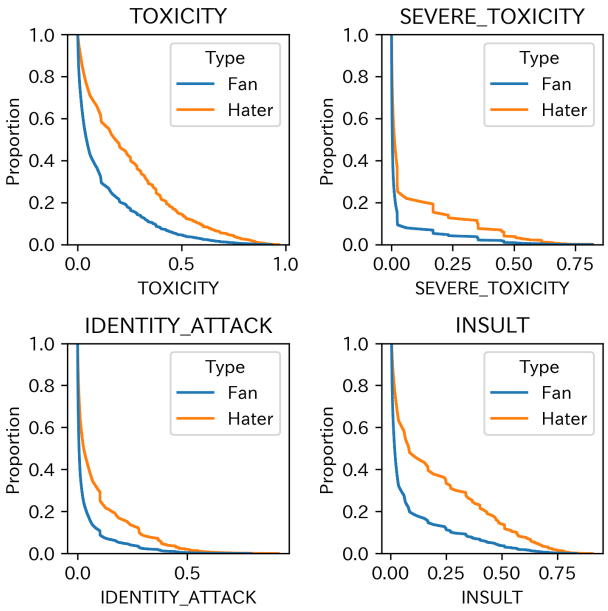

Figure 6: Overall CCDF of Perspective Scores: Fan VS Hater. The plot represents the distribution across the entire population

Toxic and Intensity of Fandom To examine the context of the intensity of fandom, we study if the tension between fans and haters can potentially affect the likelihood of generating toxic comments. As previously discussed in Section 3.1, each forum of target reveals the ratio of fans to haters. In fact, we find an association between the ratio of fans and toxicity score (Figure 9), and we depict the relationship between comment toxicity and the ratio of fans within each category of the target in Figure 7.

The figure demonstrates that as the ratio of fans decreases, which corresponds to an increase in the presence of haters,

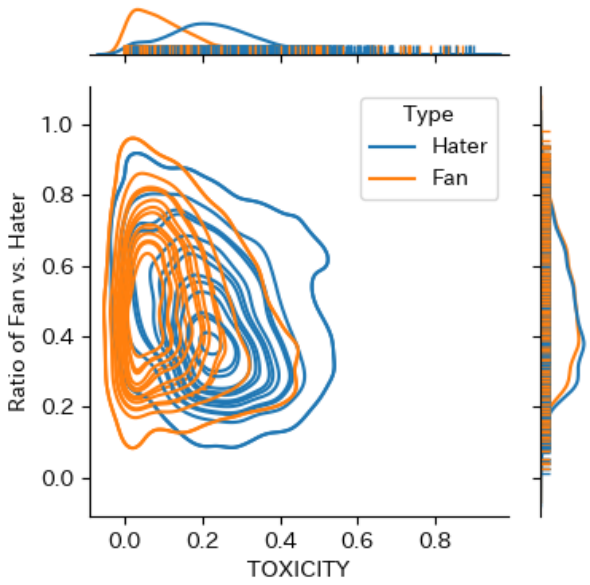

Figure 7: Joint Distribution: Fan/Hater Ratio vs. Toxicity

the fans employ toxic comments. The figure illustrates both the fans and haters as having a negative correlation between the dominance of fans and the toxicity of their comments. Conversely, in a fan-dominated forum, both users write fewer toxic comments. We also note that the distribution for fans (orange) broadens at its lower end, indicating a context where fans become toxic. This implies that fans can be inclined to employ toxic language when they are in forums dominated by haters. One possible interpretation of this observation is that fans employ toxic language as a means to defend their favored targets against haters because fans typically hold a positive attitude towards the target by definition. Therefore, examining the interactions between fans and haters within fandom communities holds substantial value.

Dependence on User Attitude: Fans vs. Haters We investigate whether fans and haters employ toxic language in their responses to individuals with opposing attitudes (haters and fans, respectively). Within the comment section of each forum, users can reply to other users' comments. To comprehend the contextual impact, we compute the differences in toxicity score between replies and non-replies using a linear regression model. Our regression model assesses the difference of toxicity score from the baseline (Intercept) that is the toxicity score of non-reply comments by haters. variations in toxicity compared to a baseline. In other words, we calculate the effect of attributes on the toxicity score as the differences from the average scores of haters' comments [5].

The results presented in Table 3 support the hypothesis

[5]In other wards, Intercept represents the average toxity scores of the comments by haters that are not Reply, Virtual, and not being replied.

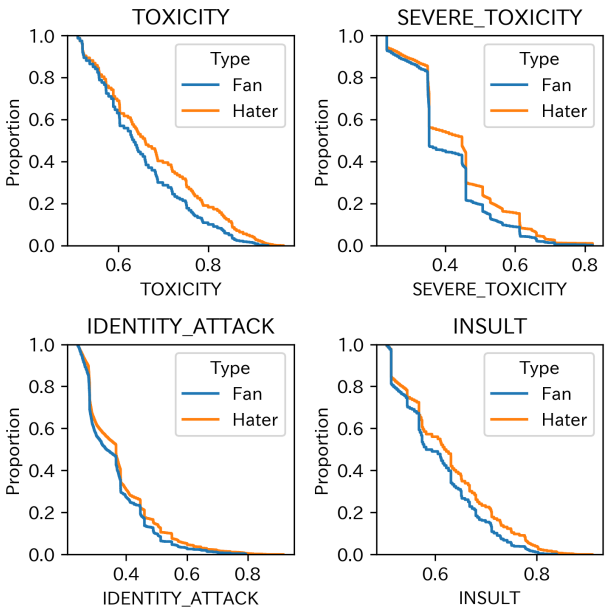

Figure 8: Top 10% CCDF of Perspective Scores: Fan VS Hater. The plot focuses on the distribution among the top 10% most toxic comments in each measurement.

introduced in the previous section. We first find that the absolute toxicity of haters' comments is much higher than fans. The first two rows of the table reveal that fans employ more toxic language in their replies. The estimations indicate that in their replies to haters, fans write comments with a toxicity score 0.088 higher than normal[6]. Interestingly, fans also exhibit an inclination to use more toxic language when replying to fans. In contrast, haters tend to compose less toxic comments in their replies. Hence, we observe a certain degree of homophily among haters. These effects surpass the impact of post popularity, as measured by the number of replies a post receives.

	TOXICITY
Reply: Fan to Hater	0.088*** (0.004)
Reply: Fan to Fan	0.033*** (0.004)
Reply: Hater to Hater	-0.036*** (0.003)
Reply: Hater to Fan	-0.007* (0.004)
Is Virtual	-0.036*** (0.002)
#of Being Replied	0.003*** (0.002)
Is Fan	-0.137*** (0.002)
Intercept (hater)	0.249*** (0.001)
R-squared	0.091

Table 3: The Toxicity and the Context of Reply. $*$ mark indicates the statistical significance (***p-val<0.01, **p-val<0.05, *p-val<0.1).

[6]Since the average score of non-reply comments by Fan is 0.112, the toxicity of fans reply to hater increases by about 78%.

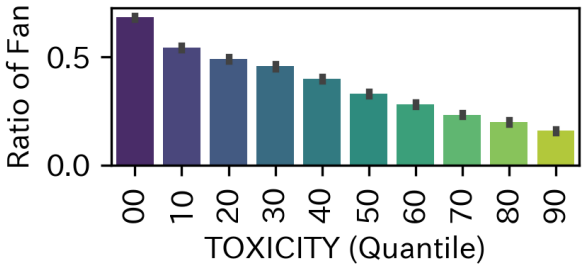

Figure 9: Average Toxicity Score in Each Quantile. The bars in the figure represent the mean values for each quantile group, along with 95% confidence intervals.

6 Discussion and Conclusions

We investigated the complexities of toxic discourse within an online platform, focusing on users' stances toward targets and the interactions between fans and haters. Our analysis first revealed that almost any entity can be a target of toxic discourse, regardless of whether it is a virtual or real person. However, they received comments composed of different contents, even though these comments were toxic. In addition, when focusing on highly toxic comments, we found that haters and fans can write almost equally toxic comments. We also observed that fans tend to employ abusive language when haters are predominant in a forum and when they reply to other users. These insights help us better understand the onset of toxic online communication and underscore the importance of understanding the authors' stance.

However, as with any study, ours too has its limitations. First, we analyzed the platform notorious for toxic content. While we ensured a robust dataset for our study, it might not be entirely representative of online platform. Our conclusions, particularly those centered around fan-hater dynamics, are rooted in specific fandom contexts. Our binary categorization of users into fans and haters might obscure the subtle shades of user attitudes because it is not always possible for us to decide for or against a topic. Therefore, our binary classification omit the discourse written by "neutral" authors. We also focused on textual interactions, neglecting other modes like emojis or images. We believe future research endeavors could expand our dataset and conduct a comprehensive analysis using full information of the forum and understand the dynamics of the communications on the platform.

Ethical Considerations

The data in this paper is derived from publicly-accessible user-generated content. We pay the utmost attention to the privacy of individuals in this study. We did not discuss the results regarding specific figures (described as "targets" in this paper), organization to keep the privacy of the individual studied in this paper.

Acknowledgements

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research Number JP22K20159, JP24K16359, JP23K16889, JP21H03557. Also, this work was supported by Research Institute of Science and Technology for Society, Japan, Grant Number JPMJRS23L4.

References

2015. Meet Cinnamon, the cute character who is viciously bullied every day in Japan.

Almerekhi, H.; Jansen, S. b. B. J.; and Kwak, c.-s. b. H. 2020. Investigating toxicity across multiple Reddit communities, users, and moderators. In *The Web Conference*.

Ashraf, N.; Zubiaga, A.; and Gelbukh, A. 2021. Abusive language detection in youtube comments leveraging replies as conversational context. *PeerJ Computer Science*.

Avalle, M.; Di Marco, N.; Etta, G.; Sangiorgio, E.; Alipour, S.; Bonetti, A.; Alvisi, L.; Scala, A.; Baronchelli, A.; Cinelli, M.; et al. 2024. Persistent interaction patterns across social media platforms and over time. *Nature*.

Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*.

Bertaglia, T.; Grigoriu, A.; Dumontier, M.; and van Dijck, G. 2021. Abusive Language on Social Media Through the Legal Looking Glass. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*.

Bourgonje, P.; Moreno-Schneider, J.; Srivastava, A.; and Rehm, G. 2018. *Automatic Classification of Abusive Language and Personal Attacks in Various Forms of Online Communication*.

Bredikhina, L.; Kameoka, T.; Shimbo, S.; and Shirai, A. 2020. Avatar driven VR society trends in Japan. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*.

Caines, A.; Pastrana, S.; Hutchings, A.; and Buttery, P. 2018. Aggressive language in an online hacking forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*.

Castaño-Pulgarín, S. A.; Suárez-Betancur, N.; Vega, L. M. T.; and López, H. M. H. 2021. Internet, social media and online hate speech. Systematic review. *Aggression and violent behavior*.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *CSCW*.

Chang, J. P.; Cheng, J.; and Danescu-Niculescu-Mizil, C. 2020. Don't let me be misunderstood: Comparing intentions and perceptions in online discussions. In *Web Conference*.

Cheriyan, J.; Savarimuthu, B. T. R.; and Cranefield, S. 2021. *Norm Violation in Online Communities – A Study of Stack Overflow Comments*. Springer International Publishing.

Edwards, L.; Kontostathis, A. E.; and Fisher, C. 2016. Cyberbullying, race/ethnicity and mental health outcomes: A review of the literature. *Media and Communication*.

Faizah, H.; Mustika, T. P.; and Prasetyo, S. E. 2019. Verbal Violence by Candidates Supporters Community in the Political Discourse on the 2019 Presidential Election on Instagram Social Media.

Garcia-Pueyo, L.; Kumar Sunkara, V.; Senthil Kumar, P.; Diwan, M.; Ge, Q.; Javaherian, B.; and Verroios, V. 2023. Detecting and Limiting Negative User Experiences in Social Media Platforms. In *The Web Conference*.

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

Gradinger, P.; Strohmeier, D.; and Spiel, C. 2010. Definition and measurement of cyberbullying. *Cyberpsychology*.

Hebert, L.; Golab, L.; and Cohen, R. 2022. Predicting Hateful Discussions on Reddit using Graph Transformer Networks and Communal Context. In *WI-IAT*.

Hiaeshutter-Rice, D.; and Hawkins, I. 2022. The Language of Extremism on Social Media: An Examination of Posts, Comments, and Themes on Reddit. *Frontiers in Political Science*.

Hua, Y.; Naaman, M.; and Ristenpart, T. 2020. Characterizing twitter users who engage in adversarial interactions against political candidates. In *CHI*.

Kalmoe, N. P. 2014. Fueling the Fire: Violent Metaphors, Trait Aggression, and Support for Political Violence. *Political Communication*.

Kayes, I.; Kourtellis, N.; Quercia, D.; Iamnitchi, A.; and Bonchi, F. 2015. The Social World of Content Abusers in Community Question Answering. In *WWW*.

Kowalski, R. M.; Limber, S. P.; and McCord, A. 2019. A developmental approach to cyberbullying: Prevalence and protective factors. *Aggression and Violent Behavior*.

Kumar, D.; Hancock, J.; Thomas, K.; and Durumeric, Z. 2023. Understanding the behaviors of toxic accounts on reddit. In *The Web Conference*.

Levy, S.; Kraut, R. E.; Yu, J. A.; Altenburger, K. M.; and Wang, Y.-C. ???? Understanding Conflicts in Online Conversations. In *The Web Conference*.

Li, Q. 2006. Cyberbullying in schools: A research of gender differences. *School psychology international*.

Lu, Z.; Shen, C.; Li, J.; Shen, H.; and Wigdor, D. 2021. More Kawaii than a Real-Person Live Streamer: Understanding How the Otaku Community Engages with and Perceives Virtual YouTubers. In *CHI*.

Lyu, Z. 2023. Cross-cutting interaction, inter-party hostility, and partisan identity: Analysis of offensive speech in social media. *New Media & Society*.

MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; and Frieder, O. 2019. Hate speech detection: Challenges and solutions. *PloS one*.

Mohan, S.; Guha, A.; Harris, M.; Popowich, F.; Schuster, A.; and Priebe, C. 2017. *The Impact of Toxic Language on the Health of Reddit Communities*.

Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive Language Detection in Online User Content. In *WWW*.

Ortiz, L. 2022. Risks of the Metaverse: A VRChat Study Case. *The Journal of Intelligence, Conflict, and Warfare*.

Pradhan, R.; Chaturvedi, A.; Tripathi, A.; and Sharma, D. K. 2020. *A Review on Offensive Language Detection*. Springer Singapore.

Rajadesingan, A.; Resnick, P.; and Budak, C. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *AAAI*.

Rieger, D.; Kümpel, A. S.; Wich, M.; Kiening, T.; and Groh, G. 2021. Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media+ Society*.

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

Smirnov, I.; Oprea, C.; and Strohmaier, M. 2023. Toxic comments are associated with reduced activity of volunteer editors on Wikipedia. *PNAS nexus*.

Smith, P. K.; López-Castro, L.; Robinson, S.; and Görzig, A. 2019. Consistency of gender differences in bullying in cross-cultural surveys. *Aggression and violent behavior*.

St. Michel, P. 2015. Meet Cinnamon, the cute character who is viciously bullied every day in Japan. Accessed: Oct-14-2023.

UNICEF. 2024 How to Stop Cyberbullying. https://www.unicef.org/end-violence/how-to-stop-cyberbullying. Accessed: May-03-2024.

Wiederhold, B. K. 2022. Sexual harassment in the Metaverse.

Wikipedia. 2024. Cinnamoroll.

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. *CSCW*.

Yao, M.; Chelmis, C.; and Zois, D.-S. 2019. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The Web Conference*.

7 Paper 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
   (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
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (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
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
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, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? NA
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (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 (see **?**)? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA