Differences in the Toxic Language of Cross-Platform Communities

Ashwini Kumar Singh,[1,5] **Vahid Ghafouri,**[2, 3] **Jose Such,**[1, 4] **Guillermo Suarez-Tangil**[1]

[1] King's College London
[2] IMDEA Networks Institute
[3] Universidad Carlos III de Madrid
[4] VRAIN, Universitat Politècnica de València
[5] Graphic Era University, Dehradun, India
ashwini.cse@geu.ac.in, vahid.ghafouri@imdea.org, jose.such@kcl.ac.uk, guillermo.suarez-tangil@imdea.org

Abstract

Cross-platform communities are social media communities that have a presence on multiple online platforms. One active community on both Reddit and Discord is *dankmemes*.

Our study aims to examine differences in harmful language usage across different platforms in a community.

We scrape 15 communities that are active on both Reddit and Discord. We then identify and compare differences in type and level of toxicity, in the topics of the harmful discourse, in the temporal evolution of toxicity and its attribution to users, and in the moderation strategies communities across platforms.

Our results show that most communities exhibit differences in toxicity depending on the platform. We see that toxicity is rooted in the different subcultures as well as in the way in which the platforms operate and their administrators moderate content. However, we note that in general terms Discord is significantly more toxic than Reddit. We offer a detailed analysis of the topics and types of communities in which this happens and why, which will help moderators and policymakers shape their strategies to mitigate the harm on the Web. In particular, we propose practical and effective strategies that Discord can implement to improve their platform moderation.

1 Introduction

The ample amalgam of Web communities provides safe spaces for diverse cultures to express their opinions. Due to the idiosyncrasies of the Web, these cultures naturally scatter their views across disparate platforms. For instance, some users may opportunistically (e.g., while on their phones) prefer the dynamism of Discord over the asynchronous nature of Reddit. While it is well established that we adapt our language according to the audience and the medium to cope with social norms (Zhong et al. 2017), it is less clear to what extent individuals self-impose different norms around the use of *toxic language* according to the platform they are in. Also, different platforms such as Discord and Reddit have their own policies and guidelines, and moderators who may apply them differently.

Related work has established links in the spread of toxic content between different *loosely* connected communities like fringe communities (e.g., 4chan), mainstream (e.g., Reddit or Twitter) (Zannettou et al. 2018; Ribeiro et al. 2021), and chat-based platforms (Si et al. 2022). While there is a "need to have a multi-platform point-of-view when studying [problematic content] on the Web" (Tahmasbi et al. 2021), there have been limited attempts in measuring *strongly* connected communities.

In this paper, we collect a unique dataset of Web communities that are present simultaneously on different platforms. Our dataset opens up new opportunities for NLP researchers and Computational Social Scientists to compare the discourse across the two social media platforms. We then design a methodology to discover the differences in problematic content. At the core of our methodology, we use toxicity detection, and semantic analysis to identify nuanced contrasts in the usage of toxic language at the sentence level. We then identify which platforms have a larger number of toxic users and we show how toxicity has evolved differently over time across platforms and communities.

Through the use of our methodology to analyze 15 popular communities simultaneously present in Reddit and Discord, our paper makes the following findings:

- Overall, we see more toxicity in Discord than in Reddit. We see that communication takes different shapes on disparate platforms. Discord prompts users to communicate using more dynamic interactions, which could have an important effect on the amount and level of toxicity.
- The toxicity in Reddit is more fine-grained and oriented toward the main topic of the community (i.e., each individual subreddit) whereas the toxicity in Discord is more coarse-grained and scattered.
- We see that a handful of users account for most of the toxic content shared in most communities while the majority of users share no toxic content at all.
- There is a significant increase of toxicity across the time for most cases. This indicates that no significant change has occurred with respect to the moderation strategy during the time window of our analysis.
- There is a substantial difference in terms of moderation across platforms, but we observe that this difference does not completely explain the differences in toxicity we observe across platforms for the same community and other factors also seem to play an important role.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The paper is organized as follows. Section 2 briefly discusses the nature of the two platforms that we study (Reddit and Discord) and how they are connected. Section 3 presents our methodology. Section 4 explains the way we systemically select cross-platform communities and our dataset. Section 5 portrays the results of our methods applied to cross-platform communities. We discuss the limitations and takeaways of this in Section 6. Finally, we discuss related work in Section 7 and conclude in Section 8.

2 Problem Statement & Background

The number of controversial Web communities has grown significantly over the last few years judging by the uptake in the communities being suspended because of the use of toxic language.[1] As content moderation has an effect on the *deplatforming* of toxic communities, their users roam to those platforms that have laxer moderation as a side effect (Ali et al. 2021).

There are two factors that determine how a community is moderated. The first factor depends on the Terms and Conditions (T&C) of the platform, which may change over time, as we have recently witnessed with X (formerly known as Twitter), for example, the limitation set in July 2023 on the number of tweets each user can view.[2] The second one relates to the norms of the community and the way in which the moderators (Seering and Kairam 2023) enforce both these norms and the T&C. Moderators are generally appointed by the creator of the community (*administrator* in Discord or *top moderator* in Reddit) or by another moderator of the community, such as in Reddit. These moderators are volunteers, and their contribution is subject to their availability. In cross-platform communities with a high volume of posts, it is commonplace to have different moderators on each of the platforms. For instance, there are completely different sets of moderators (size 10) for the *music* community on Reddit and Discord.

Considering that every moderator is an individual with a unique personal perception of toxicity, their different restrictive standards may affect the level of toxicity across platforms. Testimony of this is the non-negligible number of moderated communities that had been running for a long time and have been eventually banned by the platform. The nature of two different platforms may propose disparate types of interventions, resulting in differences in terms of toxicity. Discord is structured like a group messenger which might encourage ping-pong dual dialogues whereas the design of Reddit initially encourages the users to react to a post (submission), yet with the possibility of replying to other users' comments. Meddling in a bidirectional dialogue as a moderator may have some different characteristics than meddling in the reaction to a post.

One **challenge** we face when we look for communities that coexist on more than one platform revolves around associating the coexistence of communities, i.e.: identifying how a community may be scattered across different platforms. We address this challenge by focusing on sub-communities that are *strongly* connected to each other. We say that there is a strong connection when one of the sub-communities self-declares the other one, typically through a link that reports the association. Figure 1 shows an example of such an association. In what follows we refer to cross-platform communities as sub-communities that are hosted in different platforms and they are *strongly* connected to each other.

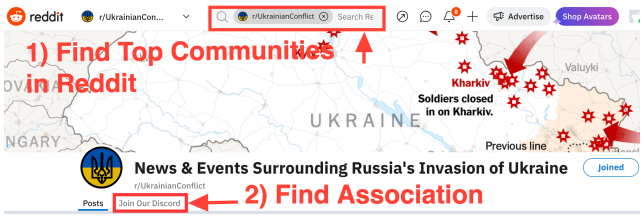

Figure 1: Sketch of the method used to find the association between communities that support multiple platforms.

We further explore the case we describe in Figure 1 and see that some subreddits set a pointer to the official Discord channel of the community. We leverage this vantage point to systematically collect associations between Reddit and Discord for the most popular communities as we explain next.

3 Methodology

Figure 2 shows the general pipeline used in this study. To observe the linguistic differences in cross-platform communities, we follow the next steps: First, we devise a systematic data collection method to find popular communities scattered across different platforms. We crawl, scrape, and process all textual comments posted in these communities. Then, we split the comments into sentences for further steps. Second, we use a machine learning classifier based on Bidirectional Encoder Representations from Transformers (BERT) to detect hateful sentences. We then perform a three-fold analysis of the differences between hateful sentences and toxic users at the platform level for every community, dubbed Differential Analysis. We next describe this approach in detail.

3.1 Data Gathering

To collect our dataset, we use the following three main steps.

Finding Cross-platform Associations As discussed in Section 2, we focus our analysis on *strongly* connected communities. To find the association between two generic communities (denoted as A and B), we start collecting data from the platform that sources the association. Let the expression $a \rightarrow b$ represent a community a containing a link to the community b.

We start a first crawling task over platform A. This crawling task is designed to query the index page of the platform and return as output the name of community $a \in A$, together with their link. We sort all communities by popularity, as given by the number of users in each community.

[1] Since 2020 Reddit banned several communities with hundreds of thousands of users, like r/TruFemcels, r/NoNewNormal, r/MGTOW, r/ChapoTrapHouse, r/GenderCritical, r/The_Donald.

[2] https://twitter.com/elonmusk/status/1675187969420828672

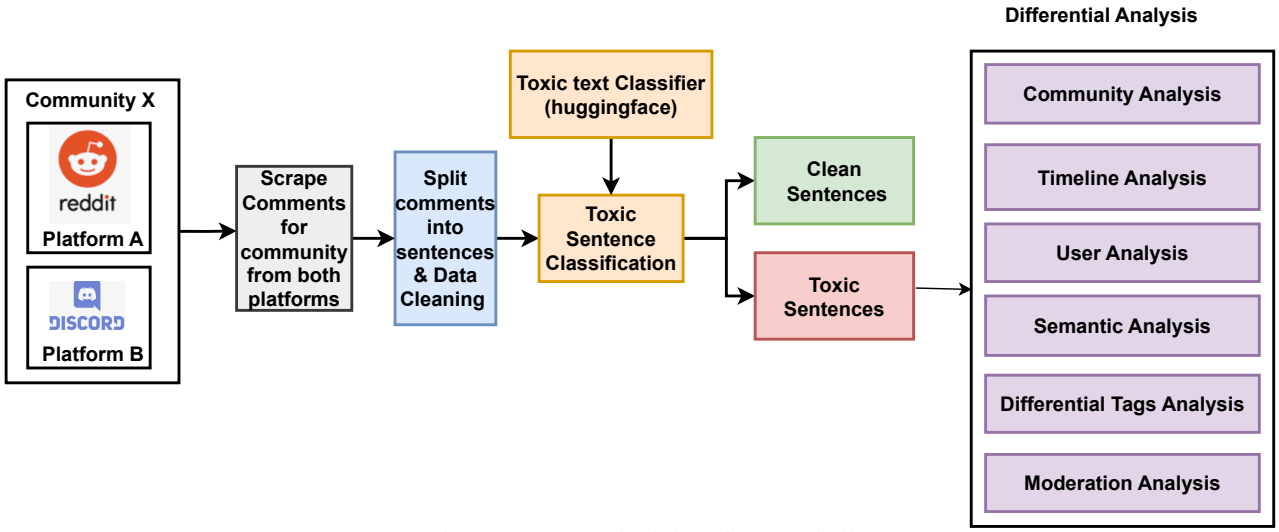

Figure 2: Our methodology in a nutshell.

We next inspect the top most popular communities in descending order and extract $a \rightarrow b$. Not all communities declare an association; therefore, we iterate through our association step until we obtain a significant set of associations. In this paper, we limit the scope of our data collection to 20 cross-platform communities to avoid indiscriminate data collection. In our implementation, our data gathering departs from Reddit and gets associations to Discord.

Scraping data for selected communities Once we have the list of associations, we continuously crawl the posts shared in both Reddit and Discord for all communities. We use the publicly available APIs for data collection from the Reddit and Discord platforms. The main attributes of the scraped data that we use in our study are "users", "posts", and "timestamps". We have anonymized the user names in the scraped data. Additionally, we have opted not to conduct any analysis at the individual user level. This approach ensures that our study mitigates any potential ethical implications associated with scraping user data. Details about data scraping are discussed in detail in Section 4.

3.2 Differential Analysis

We use Differential Analysis (Evans and Savoia 2007) to compare toxicity across platforms across three axes: the semantics of the topic, the users, and the time. Differential Analysis is a general method that compares two properties by subtracting the normalized value of the property itself. Next, we explain each dimension of our analysis in detail.

Community Analysis We examine the comments we scrape from each platform as a first step. Preliminary results show that comments on Reddit are significantly larger than in Discord. This is due to the dynamic nature of Discord proper of an instant messaging platform. For a fair comparison, we split the comments we collect from each community into sentences. Then, we use a pre-trained transformer-based model from Detoxify library[3] for toxicity detection to machine-annotate the sentences in terms of toxicity. The model is not only trained to tell whether a sentence is toxic or not but also to categorize the toxicity of sentences as *"Severe-Toxic"*, *"Obscene"*, *"Threat"*, *"Insult"*, and *"Identity-Hate"*. Finally, we compare and report all categories of toxicity for the same communities across platforms in terms of the rate and distribution of the toxicity.

Timeline Analysis In addition to the static analysis of the overall toxicity, we also compute the rate of toxic comments per day to capture the possible effects of real-world events on the temporal toxicity rate. We then study the chronological distribution of hate across time.

Users Analysis We are also interested in the distribution of toxicity across users of each community to assess the share of the most toxic users from the overall toxicity. Thus, we also aggregate sentences per user to obtain the toxicity rate for each user. To regularize the problem, we only consider the "Hateful" category when calculating the user toxicity rate. For instance, a user with 10 total comments, 2 of which are "Hateful" and 8 non-toxic, is considered 20% toxic. This is useful in order to see how skewed the share of toxic content is distributed among all users of a community. The moderation policy can change accordingly considering that banning a few top most toxic users in a more skewed community can moderate a higher proportion of the entire toxic content whereas, in a more uniformly distributed toxicity, the policy might need to be more effective when oriented toward the content rather than users.

Semantic Analysis Semantic tagging (Rayson et al. 2004) is the task of assigning semantic class categories (tags) to the smallest meaningful units in a sentence, and it is an application of Natural Language Processing. We apply the semantic tagging technique to investigate and understand the linguistic differences, and topics of discussion in communities across platforms. In our experiments, we used Python Multilingual Ucrel Semantic Analysis System (PyMUSAS)[4]

[3]https://pypi.org/project/detoxify/

[4]https://github.com/UCREL/pymusas

library. It assigns a semantic category tag or tags to every word in a given text. We use toxic sentences as input to the USAS tagger and get the output as a list of associated tags for each token from text and the total count of each tag.

This comparison gives a view of the similarities and differences in toxic sentences posted by communities across platforms. We report the top 10% of semantic tags (ignoring other tags because of relatively low values) in each community for Reddit and Discord. Then, we compute the percentage of each tag in the community for both platforms. We subsequently compute the absolute differences in the percentages of Reddit and Discord tags. We finally sort the list of tags in decreasing order of absolute differences and pick the top 2 (most dissimilar) and bottom 2 (most similar) tags to highlight the most distinctive and common features across platforms. To give a more holistic view of similarities and differences across all tags, we also compute a measure of cosine similarity between semantic tags. For this, we take two vectors having counts of each semantic tags on Reddit and Discord respectively for the same community, we normalise the vectors, and then compute the cosine similarity between them.

Differential Tags Analysis Diving deeper into the linguistic contrasts between platforms, we aim to highlight the most significantly contrastive semantic tags between the two platforms. We subtract the frequency percentile ranking of every tag in Discord, with respect to other tags in the same corpus, from its frequency percentile ranking in Reddit (and vice versa). We then use this margin to measure a contrastive significance for each tag. Let $CS_{T_{ij}}$ denote the contrastive significance of tag i in community c. Also, $F_{T_{icp}}$ denotes the frequency percentile ranking of tag i with respect to other tags in platform p of community c. Then, we compute $CS_{T_{ij}}$ for Reddit (R) over Discord (D) as in Equation 1. Next, to calculate the mean contrastive significance across every cross-platform community, we also measure the 95% confidence interval for the salience of every tag and exclude the tags with a lower bound below zero.

$$CS_{T_{ij_{(R|D)}}} = \frac{\sum_c F_{T_{ic(p=Reddit)}} - F_{T_{ic(p=Discord)}}}{|C|} \quad (1)$$

Moderation Analysis To explore the moderation differences across platforms for the same community, we examine the rate of deleted comments. While we do not have access to the actual content of the posts been deleted, we do see a label that describes when a message has been deleted by the moderator (including auto-moderators[5]). Thus, we start by looking at all the content deleted by moderators for the communities and platforms under consideration as follows.

First, we assume that deleted comments have been moderated due to toxicity. We weight every deleted comment by the average sentences per comment in the community. Then we add this to the count of toxic comments and recalculate the percentage of toxicity in the community per platform. This allows us to investigate any differences between the percentage of moderated content in Reddit and Discord. Note that we are estimating the level of toxicity as if a comment would had been removed by a moderator because of toxicity and we are assuming that all sentences in that comment are toxic. Thus, this analysis has to be seen as a high over-approximation. However, this is sufficient to compare platforms and to show, as detailed later throughout our results, that moderation plays a role but it is not the only reason for differences in toxicity across platforms for the same community.

Communities	Description
dankmemes	discuss memes that are unique or odd.
europe	community of peoples from fifty-six plus countries and two hundred thirty plus languages.
games	interesting gaming content and discussions.
history	discussions about history.
jokes	posts hundreds of jokes each day.
kpop	discuss k-pop (Korean popular music).
ksi	discuss KSI (an English YouTuber and rapper).
music	a platform to discuss about music.
nosleep	share scary personal experiences.
overwatch	related to the Overwatch game.
rainbow6	discuss things about Rainbow Six Siege game.
rickandmorty	discuss animated series, Rick and Morty.
sports	discuss sports news and highlights.
Ukrainian-conflict	shares news, analysis, discussion and investigative journalism about the conflict in Ukraine.
writingprompts	a platform for people who like prompts, they write a short story based on it, post and discuss them.

Table 1: Communities description.

4 Data Collection

We take the following steps to find *strongly* connected cross-platform communities. We first identify top subreddits[6] in terms of the number of subscribers and select the top 200 subreddits. When we visit the landing page of a subreddit, we search for a Discord invitation link. This link is set by the creator of the subreddit and, while it is optional, its presence signals the existence of a Discord server for the community. When present, we use the link to join the Discord server.

Out of the 200 subreddits, we find 32 communities in both Reddit and Discord. Several Discord servers are either inactive or very small in size members. Thus, we shortlist the 20 most active communities, all with more than 500 users.

Data scraping: To scrape the subreddits (Reddit communities), we use PushshiftAPI.[7] The subreddit data is publicly available. For scraping the data from Discord servers, we use the *Requests*library in Python. We set an authentication code using a valid Discord account. We capture the server ID and channel ID to perform the crawling, which we can access after joining the server. We collect the data from both platforms for considered communities for a duration of around 7 months (January 2022 to July 2022).

After a preliminary study, we further shortlist the communities to 14 (out of 20). The most important factor in excluding 6 communities is the imbalance across platforms. These cases have one platform with significantly less number of comments available compared to the other platform

[5]Automatic Reddit built-in system based on rules: https://www.reddit.com/wiki/automoderator/

[6]http://redditlist.com

[7]https://github.com/pushshift/api

Communities	Size of communities		Duration (date)		Number of sentences		Avg sentence length	
	Reddit	Discord	from	to	Reddit	Discord	Reddit	Discord
dankmemes	5.8M	9.9K	3/1/2022	5/8/2022	3226022	502800	9.95	5.07
europe	3.4M	3.5K	2/1/2022	6/8/2022	5040172	245035	13.74	6.32
games	3.1M	4.2K	3/1/2022	5/8/2022	2457484	355211	15.6	6.98
history	17M	3.5K	2/1/2022	5/8/2022	170278	20142	16.98	12.24
jokes	23.8M	20K	2/1/2022	3/8/2022	861786	10583	9.35	4.14
kpop	1.7M	4.7K	2/1/2022	5/8/2022	675898	432422	12.14	6.33
ksi	2.6M	72.4K	2/1/2022	5/8/2022	1736469	502640	13.97	4.29
music	30.3M	22.9K	2/1/2022	3/8/2022	2761324	725668	12.65	6.16
nosleep	16.3M	2.2K	2/1/2022	6/8/2022	260787	9043	10.40	10.22
overwatch	3.9M	268K	3/1/2022	7/8/2022	1562967	2151877	13.13	4.73
rainbow6	1.5M	583.9K	2/1/2022	1/8/2022	828649	1880389	12.93	5.52
rickandmorty	2.6M	24.9K	2/1/2022	5/8/2022	256230	191391	10.15	5.72
sports	20.4M	7.9K	2/1/2022	5/8/2022	723473	10360	12.13	7.07
Ukrainian-conflict	0.361M	5K	3/1/2022	4/8/2022	4905343	388236	12.11	8.80
writingprompts	16.1M	1.8K	2/1/2022	6/8/2022	2164661	337422	11.16	6.25

Table 2: Dataset Statistics.

of the same community. After we started our data collection in January 2022, we added to our study a community called "*Ukrainian-conflict*" as the Ukraine war started in February 2022. Our rationale was to capture a freshly created yet active community. Overall, we have included a total of 15 communities in our study. Table 1 presents the description of each community.

Dataset Anonymization: We use *anonymizedf*[8] Python library to anonymize usernames and other sensitive data.

Dataset Statistics: Table 2 represents the statistics of the dataset used in the study. The size of the communities shows the total number of subscribers present in the communities. The average sentence length is given as the number of *words* per sentence. The average sentence length for Reddit and Discord is 12.43 and 6.67 respectively.

5 Results

We apply our Differential Analysis methods in Section 3.2 to measure differences in terms of toxicity across cross-platform (Reddit/Discord) communities.

5.1 Community Analysis

We compare the toxicity of Reddit and Discord as discussed in Section 3.2. We first measure the overall toxicity and we then break it down per community.

Overall Toxicity We study five categories of toxicity ranging from general hate ("Hateful" category) and toxicity (general and severe) to obscenities and insults. Figure 3 aggregates the average toxicity for all communities. We see a significantly higher toxicity rate for Discord in all categories. We observe how the communication over Discord is more dynamic and *chatty*, while on Reddit comments are argumentative. This has an impact on the type of language used, which reflects the toxicity used. Linguistic and semantic differences are further explored later on in Sections 5.4 and 5.5. Next, we take a look at toxicity per community, then in Sections 5.2 and 5.3, we look at toxicity across time and users, respectively. Finally, in Section 5.6 we look at differences in moderation and their potential relationship with the observed differences in the toxicity across platforms.

Takeaway: Toxicity seems way higher in Discord than Reddit for all categories. Interestingly, the frequency of "*Severe-Toxic*" is negligible on Reddit and more moderate on Discord, suggesting that Reddit has an uncompromising moderation policy and diligent moderators/processes towards "*Severe-Toxic*" toxicity while Discord appears more lenient.

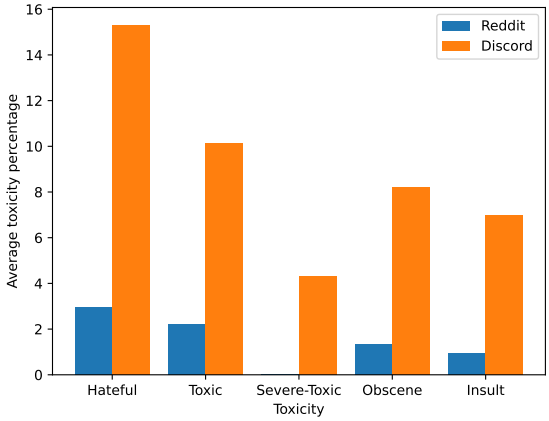

Figure 3: Average toxic sentences on Reddit and Discord platforms for communities under study.

Communities & Toxicity Table 3 shows the proportion of toxicity we see in each of the communities. Looking at the overall amount of hate ("*Hateful*" column) suggests that the most controversial community in Reddit is *rickandmorty* and in Discord is *overwatch* with 9.75% and 25.80% of hateful sentences respectively. Looking at other categories like "*Toxic*", "*Obscene*" and "*Insult*", we find *rickandmorty* as the most controversial community in Reddit and *ksi* in Discord. The "*Severe-Toxic*" is very low in Reddit communities,

[8]https://pypi.org/project/anonymizedf/

	Hateful		Toxic		Severe-Toxic		Obscene		Insult	
Communities	**Reddit**	**Discord**	**Reddit**	**Discord**	**Reddit** ($\times 10^{-4}$)	**Discord**	**Reddit**	**Discord**	**Reddit**	**Discord**
dankmemes	3.50%	14.85%	2.42%	16.89%	3.12%	5.50%	1.60%	12.24%	1.31%	11.10%
europe	1.02%	19.41%	0.89%	10.98%	3.98%	3.42%	0.44%	8.52%	0.41%	7.14%
games	1.10%	14.17%	1.03%	9.10%	0.00%	3.32%	0.71%	8.10%	0.32%	7.00%
history	0.62%	6.38%	0.64%	3.24%	0.00%	1.96%	0.13%	2.49%	0.32%	1.99%
jokes	1.93%	15.74%	1.35%	11.10%	1.24%	5.20%	0.66%	8.84%	0.51%	7.32%
kpop	1.94%	10.82%	1.43%	7.98%	0.00%	5.10%	0.92%	6.89%	0.46%	6.11%
ksi	4.53%	24.71%	2.24%	**20.10%**	11.68%	**6.34%**	1.52%	**17.22%**	0.96%	**15.21%**
music	0.85%	21.03%	0.63%	11.32%	0.43%	4.00%	0.39%	9.92%	0.26%	8.29%
nosleep	4.22%	11.31%	3.57%	8.23%	0.00%	3.12%	2.16%	7.77%	1.2%	6.28%
overwatch	1.23%	**25.80%**	0.87%	7.89%	7.69%	5.45%	0.48%	5.77%	0.31%	4.89%
rainbow6	2.66%	15.64%	1.68%	11.84%	0.00%	5.87%	1.04%	8.82%	0.65%	6.90%
rickandmorty	**9.75%**	17.15%	**6.81%**	15.45%	**46.66%**	6.22%	**4.29%**	11.32%	**2.78%**	10.33%
sports	7.14%	10.71%	5.72%	6.23%	13.65%	2.31%	3.63%	5.83%	2.73%	3.46%
Ukrainian-conflict	2.48%	14.62%	2.08%	6.87%	6.08%	3.88%	1.09%	5.10%	1.01%	4.87%
writingprompts	1.66%	7.28%	1.78%	4.76%	6.86%	2.66%	0.80%	4.44%	0.67%	4.00%

Table 3: Percentage of different types of toxicity across the two platforms per community. *(Note: We highlight in bold the highest value in a column and we underline the second highest.)*

with the exception of *ksi*, *sports* and *rickandmorty*. In Discord, the "*Severe-Toxic*" toxicity is better distributed across communities with *ksi* again standing out.

To offer a point of comparison, Table 4 aggregates the values in the *Hateful* column into three tiers of toxicity (*Low*, *Medium*, and *High*). In Reddit, we observe that all communities are in the low-toxicity tier. For Discord, most communities lie in the *Medium* and *High* level of toxic, while *history* and *writingprompts* communities lie in the *Low* level.

Toxic levels	**Reddit**	**Discord**
Low (Toxicity $<$ 10%)	All	history, writingprompts
Medium (10% $<$ Toxicity $<$ 20%)		europe, games, jokes, kpop, nosleep, sports, Ukrainian-conflict, dankmemes, rainbow6, rickandmorty
High (Toxicity $>$ 20%)		ksi, music, overwatch

Table 4: Toxicity level-wise communities.

Notably, we see that the most controversial communities across the different categories relate to the entertainment industry, including the music industry (with the KSI rap community leading the ranking), the gaming industry (led by the Overwatch gaming community), the community around Rick and Morty TV comedy show for adults, and the sports industry. Out of these categories, communities discussing the geo-political context (discussions around Europe and the Ukrainian conflict) are comparably the ones that show a larger drift in the level of hate between Reddit and Discord.

Takeaway: Overall, we see nuanced differences in toxicity across communities and we determine that the "*Hateful*" category offers a consistent summary of the different types of toxic comments. Hereafter, we focus into this category.

5.2 Temporal Toxicity

Figure 4 illustrates the Cumulative Distribution Frequency (CDF) of toxicity during our study (i.e., from January 3rd to August 3rd, 2022). We represent the average CDF values

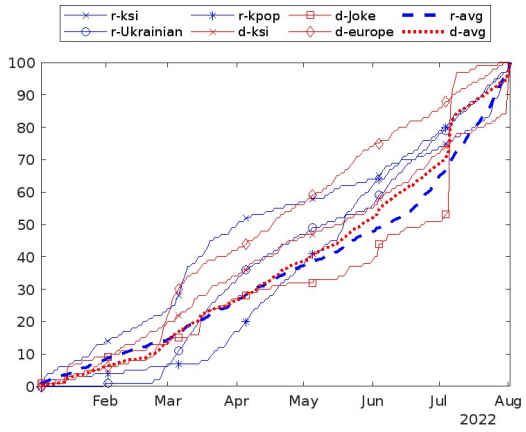

Figure 4: Toxicity Timelines.

of toxicity for the different Reddit (blue) and Discord (red) communities. Toxicity levels vary over time and can be seen through deviations from the average values (dashed blue and red lines). Some communities show a sharp increase in toxicity over time, including *Ukrainian-conflict* in Reddit and *kpop*, *joke*, and *ksi* in Discord. We attribute these spikes to various contemporary events as we discuss next.

Ukraine War In Reddit, *Ukrainian-conflict* has the highest deviation in CDF values. This is due to a drastic increase in toxicity after Russia started a full-scale invasion of Ukraine at the end of February 2022. Discord *europe* has a big jump in toxicity after the end of February which we attribute also to the effect of the war on Ukraine.

International Kissing Day The *joke* community in Discord has a significant jump of over 30% in toxicity on July 6th which is the international kissing day,[9] causing several inappropriate conversations around the topic.

[9]https://en.wikipedia.org/wiki/International_Kissing_Day

KSI vs Alex Wassabi The *ksi* on Reddit as well as on Discord shows a significant increase in toxicity starting from the end of July. We attribute this to the announcement of the fight in an exhibition boxing match between the British YouTuber KSI and American YouTuber Alex Wassabi.[10]

Takeaway: On both platforms, many communities do not show a significant variation in toxicity over time. Yet, one thing stands out: we have an increasing trend in toxicity rate on average, showing that existing moderation strategies can not scale. We also see how spikes in toxicity are contextual, mostly fostered by the existing socio-political landscape.

5.3 Toxicity analysis per user

We use the distribution of user toxicity rates in each community to provide insight into the skewness of toxicity production. Table 5 shows a summary of the results.

We consider a user to be toxic if we see a toxic statement in any of the sentences in their posts. We then look at the top 5% most toxic users and the prevalence of users with 100% toxic sentences.

Communities	Toxic Users		Top 5% Toxic		100% Toxicity	
	Red.	Disc.	Red.	Disc.	Red.	Disc.
dankmemes	47.9%	47.7%	42%	73%		11.2%
europe	37.8%	26.1%	37%	61%		6.9%
games	34.0%	15.9%	29%	52%		6.8%
history	12.2%	7.7%	15%	54%		2.2%
jokes	33.4%	37.8%	22%	38%		8.4%
kpop	36.7%	35.1%	29%	54%		4.7%
ksi	44.2%	43.9%	27%	77%	0%	14.6%
music	24.9%	35.9%	16%	69%		7.9%
nosleep	32.0%	21.0%	39%	48%		6.1%
overwatch	40.8%	34.1%	27%	85%		4.6%
rainbow6	39.8%	41.2%	33%	82%		7.0%
rickandmorty	38.8%	17.6%	34%	75%		2.4%
sports	37.1%	21.7%	25%	55%		4.1%
Ukrainian-conf.	47.8%	15.1%	44%	60%		1.7%
writingprompts	30.7%	18.2%	39%	60%		7.6%

Table 5: Toxic users for Reddit (Red.) and Discord (Disc.)

Rate of Toxic Users We see that *dankmemes* hosts the largest toxic user base, with 48% of their users posting toxic comments on both Reddit and Discord (see "Toxic Users" column of Table 5). Recall that *dankmemes* is a community that produces a relatively low or moderate level of toxicity overall (cf. Table 4 in Section 5.1). In context, this means that many of the toxic users in this community do not frequently produce toxic content.

On the contrary, we see that *history* has the lowest number of users who engage in toxic behavior with 12.2% and 7.7% of the users in Reddit and Discord using toxic language eventually. Interestingly, we observe that the number of toxic comments overall posted is 0.6% and 6.4% respectively (cf. Table 3 in Section 5.1). This shows that while *history* has more toxic users in Reddit than in Discord, Discord is overall more toxic than Reddit due to a highly skewed production of toxicity by a few top toxic users.

Takeaway: This common pattern suggests significant moderation differences between the two platforms for the same community. We come back to this point later in Section 5.6.

We further investigate the presence of the same set of users across platforms for the same community and find that some users coexist on both platforms. For instance, we see that around 13% and 8% of the Discord users in *writingprompts* and *nosleep* respectively are also present in Reddit. Note that the overlap is just based on an exact username match during the time span in our dataset, but we studied the user names and observed they were significantly unique.

To further study the nuanced differences between users in different communities (including *dankmemes* and *history*) we focus next on the top most toxic users.

Most Toxic Users We first look at the share of toxicity among the top 5% users in each community as shown in the middle column of Table 5. The numbers suggest that the share of toxicity among users is far more skewed in Discord, meaning that a few extremely toxic users account for most of the toxic content in this platform.

This finding is also consistent when we examine the proportion of users who *always* use toxic language (see "100% Toxicity" column in Table 5). As shown in the table, none of the Reddit communities have any individual who consistently generate toxic content, while all communities in Discord have a few of them. In particular, 15% of the users in *ksi* display toxicity in 100% of their posts. This figure ranges all the way to 2% in the case of *Ukrainian-conflict*.

Takeaway: While we have seen that toxicity in Discord is concentrated in a few accounts, the toxicity in Reddit is scattered across a wider range of users.

We next seek to understand if this toxicity is generally directed towards certain topics through the analysis of linguistics and semantic differences.

5.4 Semantic Categories Analysis

Aiming to compare the linguistic differences in toxic sentences across Reddit and Discord platforms, we compare the communities using their respective semantic tag values evaluated by the USAS semantic tagging model. To compute the cross-platform similarity in semantic tag values, we take two vectors for semantic tags, one for a community on Reddit and another for the same community on Discord. Then, we compute the cosine similarity between the two vectors and get a similarity score.

Table 6 shows the cosine similarity scores across platforms for all communities. Here, we can observe that the *nosleep*, *writingprompts*, *ukrainian-conflict* and *history* communities are more similar in topics, whereas *overwatch* and *dankmemes* communities have substantial differences in topics across platforms. Overall, the cosine similarity scores of semantic tags (topics) are high for all the communities, which indicates that the topics discussed in a particular community on different platforms are very similar in general.

Table 6 also shows the two most similar and the two most dissimilar topics in communities across platforms. Most of these topics are compatible with the basic theme of the communities, which validates the significance of semantic tags

[10] Announcement made July 17, 2022, https://en.wikipedia.org/wiki/2022_in_Misfits_Boxing.

Communities	Cosine Similarity	Most Similar Tags	Most Dissimilar Tags
dankmemes	0.92	K5, W4	S3, B1
europe	0.98	G3, I1	G1, S3
games	0.93	M7, L2	K5, B1
history	0.99	S9, M7	S3, G2
jokes	0.97	S9, K2	S3, S1
kpop	0.98	G3, K5	K2, S3
ksi	0.97	G1, G3	S1, S3
music	0.98	P1, X2	S1, K2
nosleep	0.99	Y1, B2	L1, S1
overwatch	0.91	S9, I3	K1, K5
rainbow6	0.95	C1, S7	S3, K5
rickandmorty	0.98	L1, B2	X2, S3
sports	0.98	B4, Y2	S1, S3
Ukrainian-conflict	0.99	G3, H3	X9, Z2
writingprompts	0.99	C1, B2	F1, L1

Table 6: Cross-platform cosine similarity for semantic tags with most similar and dissimilar tags in toxic sentences.

analysis used in our study. These topics are determined by using the method mentioned in Section 3.2. Table 10 represents the names of the tags mentioned in this paper. Intuitively, due to the escalation of disputes between Russia and Ukraine, *europe* community is talking about warfare, defense, and the army — i.e., weapons (G3) topics on both platforms. The *ukrainian-conflict* is using the terms related to G3 and areas; boundaries (H2) in a similar size on both platforms. The *history* community is also discusses topics related to places (M7) more evenly.

Interestingly, we see that *music* and *kpop* communities are dissimilar when talking about music and related activities (K2) across Reddit and Discord. Also, *games*, and *overwatch* are the most dissimilar in Sports and Games related semantic tags (K5). Both tags are directly referring to the topics of their community. Further analysis shows that the source of this dissimilarity is their extremely higher abundance on Reddit than on Discord. This means that the discourse in Reddit content is closer to the theme of the community (for example in sports community they talks about sports activities), whereas Discord content does not completely stick to the related topic of the community and can also drift to other topics. This may be related to the nature of Reddit, where comments include reactions to the submissions related to the main topic of the community. In contrast, Discord servers are structured as group messengers, which may favor back-and-forth conversations between users, including the toxic ones, who may then diverge from the main topic of the community.

Takeaway: We see that the topics and semantic similarity are very high for all communities across platforms, suggesting very similar topics being discussed most often aligned with the main theme(s) of the communities. Interestingly, we also observe some differences between platforms, where Reddit discussions are more often bounded to the main theme(s) of the community, while Discord discussions seem to more easily diverge from the main theme(s) of the community, while still being the main theme(s) discussed.

5.5 Linguistic Differences

We measure linguistic differences in toxic language by looking at differences in frequency percentile rankings of the USAS tags discovered in Section 5.4. In particular, we portray the contrastive nature of semantic tags as word clouds, where the sizes of words correspond to the measure of salience described in Section 3.2. The keywords we display in the word clouds correspond to the description of each tag.

Figure 5 shows tags of toxic content that are more present in Reddit when compared to Discord. Figure 6 shows the reverse, that is the tags of toxic content that are more present in Discord rather than Reddit. To provide more context to interpret the results, we further offer details and provide sample words and sentences associated with each tag in Table 7. We see that tags corresponding to gender ("People: Female"), drugs ("Plants"), and other general-purpose topics describing "Dislike", and "Sensory: Smell" are more frequent in Discord than in Reddit. Here, we observe more explicit toxicity associated with these popular tags (see column "Top words" in Table 7) in Discord. This suggests that, for the same community, the toxicity in Discord seems to be more explicit, particularly for some topics such as drugs and the female gender. This could be explained by the more semi-private nature of Discord as opposed to Reddit, where some users, even if anonymous, may be more reluctant to make some comments explicit *in public*. This could also be related to differences in moderation policies and processes as we explore in the next section, where Reddit policies and moderators may be harsher for explicit language.

Takeaway: Toxicity in Discord tends to be more explicit, particularly in reference to topics such as drugs and the female gender, when compared to Reddit.

5.6 Moderation Differences

We also study differences when it comes to moderation.

Attribution Table 8 shows the percentage of comments deleted or removed by moderators. We see that moderators on the Reddit platform are more active and strict than on Discord. In Reddit, "*nosleep*", "*sports*", and "*history*" maintain the highest number of deleted comments. In Discord, moderators seem much more lenient as deleted/removed comments are exceptional. Note that moderation policies in Reddit[11] and Discord[12] seem very similar when it comes to how moderators should handle toxic content. However, these differences we observe seem to be attributed to the way moderators apply policies in practice. We also see evidence of Reddit using automated systems to moderate comments (auto-moderation). We see in Table 8, column Reddit (AM), the percentage of those comments deleted because they matched the automated rules moderators set in Reddit. We note there are some communities where automated moderation is barely applied, but we do not see a connection

[11] https://www.redditinc.com/policies/moderator-code-of-conduct

[12] https://discord.com/community/your-responsibilities-as-a-discord-moderator-discord

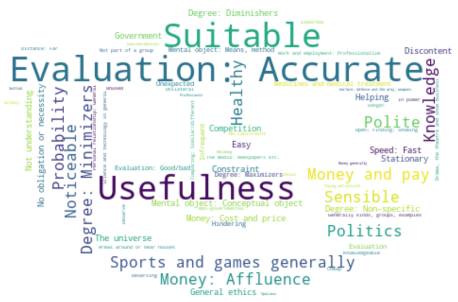

Figure 5: Salient USAS tags in Reddit toxic content.

Figure 6: Salient USAS tags in Discord toxic content.

USAS Tag	Description	Dominant Platform	Saliency	Top words	Example Sentences
S1.2	People: Female	Discord	0.12 ± 0.087	Bitch, Girl, Mom, Women, Whore, Cow	bitches come and go bruh, you little bitch
X3.5	Sensory: Smell	Discord	0.10 ± 0.08	Smell, Stink, Smelly	smells like shit though, my plushies stink, when your opinion smells of stupid
E2-	Dislike	Discord	0.097 ± 0.038	Damn, Hate, Bitches, Fuck	damn slowchat, lil whiney bitch
L3	Plants	Discord	0.083 ± 0.046	Weed, Smoke	polish cow weed, chat is too green and stupid
E6+	Confident	Discord	0.093 ± 0.079	Fuck, Hot, Shit, Cool	fuck indeed, fuck you shut up and go buy gold

Table 7: Tags description with sample sentences.

with the total amount of moderation (e.g., "*sports*" vs "*europe*" when comparing the two columns in Table 8) or the overall toxicity (e.g., "rickandmorty" vs "*sports*" or when comparing Table 8 and 3).

Explanation Next, we focus on whether these differences in moderation could explain the differences in toxicity observed in Section 5.1. That is, whether all communities are less toxic in Reddit simply because the moderation in Reddit is more strict when it comes to toxic content. Table 9 shows a substantial increase in toxicity percentage in Reddit communities when considering our estimate based on the moderated content. Still Discord exhibits a higher toxicity rate as we see in the majority of the communities, such as "*europe*", "*kpop*", "*ksi*", "*music*", "*overwatch*", "*rainbow6*", and "*Ukrainian-conflict*" when comparing the *estimated* (upper-bound) Reddit toxicity in Table 9 with the *actual* Discord toxicity back in Section 5.1 (Table 3).

Takeaway: Our analysis reveals that there are important differences in handling toxic content across platforms. Reddit has more proactive moderation strategies than Discord, with some of them driven by automated mechanisms. When we factor moderation in, we continue to see that Discord is more toxic than Reddit. This shows that there are other reasons beyond moderation to explain the difference in toxicity for the same community across Reddit and Discord. As these differences are substantial and the communities we study are *strongly* connected, meaning that administrators of the community may either be the same or cooperate, we partially attribute the drift in toxicity to the other differences observed across platforms beyond moderation, including the type of users there are or the nature of the conversations they have as we saw in Sections 5.3 (users), 5.4 (semantic differences) and 5.5 (linguistic differences).

6 Discussion

We discuss the main takeaways and limitations of our study.

Communities	Reddit	Reddit (AM)	Discord $\times 10^{(-4)}$
dankmemes	7.6%	2.1%	-
europe	4.4%	0.22%	-
Games	15.0%	0.75%	-
history	17.0%	5.9%	-
Jokes	11.8%	0.02%	-
kpop	2.9%	1.1%	9.3%
ksi	5.9%	2.2%	4.0%
music	2.8%	1.7%	-
nosleep	25.4%	5.2%	-
Overwatch	1.0%	1.1%	2.6%
Rainbow6	1.4%	2.1%	5.9%
rickandmorty	2.4%	1.1%	-
sports	21.2%	0.02%	-
Ukrainian-conflict	2.8%	1.2%	3.1%
Writingprompts	7.5%	5.6%	-

Table 8: Percentage of deleted comments per community and platform by moderators. AM: Auto-moderation.

6.1 Main Takeaways

Our paper offers a unique comparison of cross-platform communities that yields the following findings:

Discord is more toxic than Reddit Comparing the rate of toxicity across Reddit and Discord shows a clearly generalizable pattern. For all considered communities, the content of that community in the Discord platform is substantially more toxic in all categories of toxicity in comparison to the Reddit platform of the same community. Notably, the prevalence of the "*Severe-Toxic*" category is almost negligible on Reddit while clearly existing in Discord. Moreover, the toxicity is found to be more explicit (i.e., containing predefined toxic words) on Discord than on Reddit. We studied the root cause and made the observations that follow next.

Moderating toxic users may work for Discord We observe that the distribution of toxic behavior between users is not consistent when comparing Discord and Reddit. On Discord, a small number of users are accountable for the majority of negative content, whereas on Reddit, the toxic-

Communities	Reddit baseline	Reddit estimate
dankmemes	3.5%	10.33%
europe	1.0%	5.24%
Games	1.10%	15.7%
history	0.62%	14.58%
Jokes	1.93%	13.88%
kpop	1.94%	4.7%
ksi	4.53%	12.7%
music	0.85%	3.43%
nosleep	4.22%	26.63%
Overwatch	1.23%	2.21%
Rainbow6	2.66%	3.9%
rickandmorty	9.75%	12.1%
sports	7.14%	27.34%
Ukrainian-conflict	2.48%	5.1%
Writingprompts	1.66%	7.9%

Table 9: Percentage of toxicity before and after including deleted comments as toxic comments.

ity is spread more uniformly among the users. Consequently, on Discord, implementing fundamental moderation tactics, such as banning the primary toxic users, can be a successful strategy, while on Reddit, a more effective approach would be to target toxic comments than toxic users.

Increased tendency over time We see that the cumulative distribution of toxicity over *time* increases linearly (uniform distribution), with Reddit leading the way to Discord users. Interestingly, we see more spikes of toxicity over time in Discord than in Reddit where toxicity is scattered across time more homogeneously. While we see evidence of content moderation, we also see that the increase in toxicity rarely plateaus over time. This means that there is a baseline of toxicity that always permeates through. Observing the timeline of toxicity in communities such as *Ukrainian-conflict*, *europe*, and *ksi*, we can infer that the toxicity on platforms may also be related to specific events associated with the respective online communities.

Semantic and linguistic differences We observe that the use of toxic language can be attributed to different topics depending on the platform. This may mean the same community is represented by a different subculture, each attracted to the idiosyncrasies of the platform. For instance, semantic tag dissimilarities for communities such as *music*, *kpop*, *sports*, *games*, and *overwatch* suggests that content and toxicity are more fine-grained and focused in Reddit than in Discord. This refers to the nature of Reddit, where comments are reactions to the submissions that are directed toward the subreddit's topic, yet, in Discord servers, which are structured as group messengers, back-and-forth conversations between a few users, including the toxic ones, may easily diverge from the main topic of the community.

And without moderation, Discord is still more toxic We also see that moderation plays a significant role in explaining variations in toxicity levels, with instances where it independently influences outcomes. Nevertheless, even after estimating the level of toxicity that one would encounter in Reddit if moderation was not present, more toxicity would still be found in Discord across most of the communities. This observation prompts further exploration of additional contributing factors, such as differences in platform-specific language, in the type of communication, including topics, toxicity explicitness, and/or the level of (in)formality proper of a more/less public and direct channel. Regardless of the differences, we see Reddit using auto-moderation systems. It is unclear whether Discord also uses automated systems to help moderators but in either case, we see how the deployment of cutting-edge methods — e.g., (He et al. 2023) or Detoxify — is an open problem in practice most likely due to the implications of blocking content automatically under the presence of false positives.

Connection to Social Science Theories This study's findings resonate with several well-established social science theories that illuminate the dynamics of online toxicity and group behavior. Firstly, the concentration of toxic behavior within a small subset of Discord users aligns with the "*bad apple effect*" (Myatt and Wallace 2008). This theory posits that a few disruptive individuals can exert a disproportionate negative influence on the overall climate of a community. This suggests that targeted interventions aimed at these high-impact users could be a particularly effective strategy for reducing toxicity on platforms like Discord.

Secondly, the theory of deindividuation (Watson 1973) offers insights into the higher levels of toxicity observed on Discord. The anonymity and reduced personal accountability fostered by Discord's real-time chat format may lead to greater disinhibition and a willingness to engage in toxic behaviors. In contrast, Reddit's forum-like structure and comment voting system can promote greater self-awareness and a degree of social regulation.

Finally, the observed differences in semantic focus between platforms point to the potential role of social identity theory (Tajfel and Turner 2004). This theory suggests that individuals may gravitate towards platforms that reinforce their sense of group belonging, leading to the emergence of platform-specific subcultures with varying norms regarding acceptable discourse. The distinct linguistic patterns on Discord and Reddit could reflect these social identity processes and how they contribute to variations in online toxicity.

6.2 Limitations

Our method provides a holistic view of cross-platform similarity rates with a granularity that explains what the similarities and differences are. However, our granularity when it comes to linguistic and semantic differences is limited to semantic tags, which only provide an overall notion of the concepts mentioned in a text, rather than identifying the unique context in which the tags appeared. It is also worth noting that rule-based semantic taggers may have limitations in capturing non-defined or new tags and topics. However, finding a precise mechanism for understanding semantics is a daunting NLP task that is out of the scope of our contribution. Despite the tools we have used for semantic analysis having limitations, their use has led us to the identification of nuanced differences that advance our understanding of the use of toxicity in cross-platform communities beyond prior work which focuses on the use of sentiment analysis.

To examine the moderation differences, we used an upper bound that all the sentences in deleted comments are toxic. This assumption leads to an overestimation of the toxicity, but this limitation does not affect our findings since the toxicity in Discord is still higher than in Reddit before and after factoring in moderation. If we were to have access to the deleted comments and the amount of toxicity in moderated comments were to be accurate, we would find a smaller increment and we would reach the same conclusion.

7 Related Work

Related work has focused on differences in sentiment analysis of content generated across platforms. For instance, while examining the posts posted by the same group of users on Instagram and Twitter, (Manikonda, Meduri, and Kambhampati 2016) saw that posts on Twitter contain more negative expressions than posts on Instagram. (Ali et al. 2023) also argued that meta-data features (e.g., conversation length) were better predictors of risky conversations on Instagram. (Lin and Qiu 2013) found that Twitter posts are more causal, while posts on Facebook are more emotional. In addition, a case study by (Ruan et al. 2022) on the 2019 Ridgecrest earthquake showed that Reddit users' responses to the event were much less emotionally negative and covered more diverse topics than the same discussion on Twitter. Moreover, the responses to the event are more active and faster on Twitter than on Reddit.

More relevant to our research question, several works have attempted to compare the mechanism of harmful content and behavior across platforms. (Van Raemdonck 2019) studies how different platforms (Facebook and Reddit) allow for the spread of anti-vaccine conspiracy theory. Looking into news consumption during the Italian referendum, (Vicario et al. 2017) discuss that users on Facebook and Twitter are equally likely to restrict their attention to a certain group of pages/accounts. (Yang et al. 2021) have also looked into Facebook and Twitter's role in spreading COVID-19 misinformation and figured out that on both platforms, low-credibility content is generally much more prevalent than content from high-credibility sources. However, the ratio of low- to high-credibility information on Facebook is lower than on Twitter, suggesting that Facebook's misinformation moderation strategy is more effective.

Although many works have been studying linguistic differences on multiple platforms, no work has explored the linguistic differences for harmful content posted by communities across multiple platforms, which is a gap our work fills. Moreover, existing tools for cross-platform comparison are limited to sentiment analysis and conventional topic modeling next to temporal frequency counts (e.g., the number of comments with negative sentiment (Manikonda, Meduri, and Kambhampati 2016; Lin and Qiu 2013; Ruan et al. 2022), or the number of links to deleted YouTube videos (Yang et al. 2021)). Our study goes beyond sentiment analysis and makes nuanced comparisons across several axes.

8 Conclusion

In this paper, we make a novel analysis and collect a unique dataset of cross-platform communities. Our work is the first to study *strongly* connected communities that are simultaneously present on Reddit and Discord, focusing on the analysis of the differences in the use of toxicity and in moderation. We observed a substantially higher *overall* toxicity in Discord than in Reddit and we offered a nuanced analysis of root causes, including differences we attribute to the user base, to opportunistic events that happen over time, and to the semantic differences in the nature of the conversations.

While our work focuses on toxicity, our methods and dataset can be leveraged for a wide range of studies. In particular, the metrics we use (e.g., semantic analysis) are generalizable for measuring the similarity of any two corpora in the future. To foster future work in the space, we make our code and anonymized dataset available to the research community on GitHub.[13]

Acknowledgements

This research was supported by UKRI through REPHRAIN (EP/V011189/1), the UK's Research centre on Privacy, Harm Reduction and Adversarial Influence online, as part of its PROM project, MCIN/AEI/10.13039/501100011033 and the European Union-NextGenerationEU under grant TED2021-132900A-I00. G. Suarez-Tangil has been appointed as 2019 Ramon y Cajal fellow (RYC-2020-029401-I) funded by MCIN/AEI/10.13039/501100011033 and ESF Investing in your future.

References

Ali, S.; Razi, A.; Kim, S.; Alsoubai, A.; Ling, C.; De Choudhury, M.; Wisniewski, P. J.; and Stringhini, G. 2023. Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1).

Ali, S.; Saeed, M. H.; Aldreabi, E.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*, 187–195.

Evans, R. B.; and Savoia, A. 2007. Differential testing: a new approach to change detection. In *The 6th Joint Meeting on European software engineering conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering: Companion Papers*, 549–552.

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

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

He, X.; Zannettou, S.; Shen, Y.; and Zhang, Y. 2023. You Only Prompt Once: On the Capabilities of Prompt Learning on Large Language Models to Tackle Toxic Content. arXiv:2308.05596.

[13] https://github.com/aksiitbhu/cross-platform-analysis

Lin, H.; and Qiu, L. 2013. Two sites, two voices: Linguistic differences between Facebook status updates and tweets. In *International Conference on Cross-Cultural Design*, 432–440. Springer.

Manikonda, L.; Meduri, V. V.; and Kambhampati, S. 2016. Tweeting the mind and instagramming the heart: Exploring differentiated content sharing on social media. In *Tenth international AAAI conference on web and social media*.

Myatt, D. P.; and Wallace, C. 2008. When Does One Bad Apple Spoil the Barrel? An Evolutionary Analysis of Collective Action. *The Review of Economic Studies*, 75(2): 499–527.

Rayson, P.; Archer, D.; Piao, S.; and McEnery, A. M. 2004. The UCREL semantic analysis system.

Ribeiro, M. H.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; Long, S.; Greenberg, S.; and Zannettou, S. 2021. The evolution of the manosphere across the Web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 196–207.

Ruan, T.; Kong, Q.; McBride, S.; Sethjiwala, A.; and Lv, Q. 2022. Cross-platform analysis of public responses to the 2019 Ridgecrest earthquake sequence on Twitter and Reddit. *Scientific Reports*, 12.

Seering, J.; and Kairam, S. R. 2023. Who Moderates on Twitch and What Do They Do? Quantifying Practices in Community Moderation on Twitch. *Proceedings of the ACM on Human-Computer Interaction*, 7(GROUP): 1–18.

Si, W. M.; Backes, M.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; Zannettou, S.; and Zhang, Y. 2022. Why So Toxic? Measuring and Triggering Toxic Behavior in Open-Domain Chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, 2659–2673. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394505.

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

Tajfel, H.; and Turner, J. C. 2004. The social identity theory of intergroup behavior. In *Political psychology*, 276–293. Psychology Press.

Van Raemdonck, N. 2019. The echo chamber of anti-vaccination conspiracies: mechanisms of radicalization on Facebook and Reddit. *Institute for Policy, Advocacy and Governance (IPAG) Knowledge Series, Forthcoming*.

Vicario, M. D.; Gaito, S.; Quattrociocchi, W.; Zignani, M.; and Zollo, F. 2017. News Consumption during the Italian Referendum: A Cross-Platform Analysis on Facebook and Twitter. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 648–657.

Watson, R. I. 1973. Investigation into deindividuation using a cross-cultural survey technique.

Yang, K.-C.; Pierri, F.; Hui, P.-M.; Axelrod, D.; Torres-Lugo, C.; Bryden, J.; and Menczer, F. 2021. The COVID-19 Infodemic: Twitter versus Facebook. *Big Data & Society*, 8(1): 20539517211013861.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, 188–202.

Zhong, C.; Chang, H.-w.; Karamshuk, D.; Lee, D.; and Sastry, N. 2017. Wearing many (social) hats: How different are your different social network personae? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 397–406.

Ethics Checklist

Individual Ethics description

While Reddit data is publicly available, the data we gather from some Discord account is accessible through an invitation link we scrape from a public source. We have not informed the creators or moderators of the Discord servers, as doing so would risk our account being suspended, thus jeopardizing the feasibility of our study. While we use automated tools to anonymize our dataset in our study, we appreciate these tools are not perfect. We also study toxic language posted by users who may feel they are interacting with a semi-private space. All this together has important Ethical implications, which we discuss next.

First, we refrain from deanonymizing users, nor focus on analyzing toxicity by specific individuals. Our study does not focus on the real-life identities of individuals.

Second, the dataset we study contains a number of textual content (comments) depicting toxic or violent scripts. The dataset is intended as an academic resource and has been collected to extend the understanding of toxic language behavior in various communities across platforms. We use automated tools to identify toxic language, but some authors of this paper have been inevitably exposed to such toxicity while validating our method or while analyzing the case studies we present. We take rigorous precautions to ensure the well-being of the research team through regular meetings that are open to discuss the potential emotional toll of exposure to such content. These regular meetings serve as a forum for emotional support, allowing team members to express their concerns and feelings without fear of judgment.

We assessed the risks and benefits of our study and obtained approval from the Institutional Review Boards of our institution and the funding body that supports this work.

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, we have used the data anonymization technique mentioned in Section 4

(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? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, by the camera ready.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? No, because we have used state-of-the-art trained models available online (references/url links are provided in the paper.)

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, because we have used state-of-the-art trained models available online and already evaluated.

(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)? No

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, please refer Section 6.2.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? Yes, we have used the open source codes in our experiments, references to the source codes are mentioned in the paper.

(c) Did you include any new assets in the supplemental material or as a URL? No

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

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

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

A Appendix

Tag	Tag Name	Tag	Tag Name
B1	Anatomy and physiology	L1	Life and living things
B2	Health and disease	L2	Living creatures generally
B4	Cleaning and personal care	M7	Places
C1	Arts and crafts	P1	Education in general
F1	Food	S1	Social actions, states & processes
G1	Government, Politics & elections	S3	Relationship
G2	Crime, law and order	S7	Power relationship
G3	Warfare, defence and the army; Weapons	S9	Religion and the supernatural
H3	Areas around or near houses	W4	Weather
I1	Money generally	X2	Mental actions and processes
I3	Work and employment	X9	Ability
K1	Entertainment generally	Y1	Science and technology in general
K2	Music and related activities	Y2	Information technology and computing
K4	Drama, the theatre & show business	Z2	Geographical names
K5	Sports and games generally		

Table 10: Semantic tags used in this paper. Full list of tags https://ucrel.lancs.ac.uk/usas/semtags_subcategories.txt.