The Constant in HATE: Analyzing Toxicity in Reddit across Topics and Languages

Wondimagegnhue Tsegaye Tufa, Ilia Markov, Piek Vossen

Vrije Universiteit Amsterdam

De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

{w.t.tufa, i.markov, p.t.j.m.vossen}@vu.nl

Abstract

Toxic language remains an ongoing challenge on social media platforms, presenting significant issues for users and communities. This paper provides a cross-topic and cross-lingual analysis of toxicity in Reddit conversations. We collect 1.5 million comment threads from 481 communities in six languages: English, German, Spanish, Turkish, Arabic, and Dutch, covering 80 topics such as Culture, Politics, and News. We thoroughly analyze how toxicity spikes within different communities in relation to specific topics. We observe consistent patterns of increased toxicity across languages for certain topics, while also noting significant variations within specific language communities.

Keywords: Toxic Language, Reddit, Cross-Topic Analysis, Cross-Lingual Analysis

1. Introduction

Social media platforms have witnessed remarkable growth in their user base and significance as a means of communication. These platforms allow individuals to share whatever they wish, presenting diverse viewpoints that range from enlightening to objectionable and everything in between. As a side effect, platforms often provide a breeding ground for toxic content, such as instances of abuse and hate speech, resulting in adversities for online users (Fortuna and Nunes, 2018). Outside the confines of social media, this toxic content influences real-world dynamics. These are often manifested in instances of violence and crimes targeting minority groups (Mathew et al., 2020). The detection of toxic content has emerged as a progressively significant subject of investigation within the field of Natural Language Processing (NLP). Active research in this area focuses on creating datasets that cover different aspects of toxic content (Mathew et al., 2020; Vidgen et al., 2021; Sachdeva et al., 2022), or methods that rely on these datasets to analyze toxic content or train toxic language classification systems (van Aken et al., 2018; Radfar et al., 2020; Gevers et al., 2022; Markov et al., 2022).

While many existing studies focus on classifying whether a given text is toxic and why, the context in which such inappropriate content arises is less explored (Zhou et al., 2023; Sap et al., 2019).

In Reddit, a specific discussion often turns toxic when the topic of discussion is sensitive to a particular user. Participants of such discussions with opposing views engage in unhealthy debates, which can quickly escalate. A sensitive topic may evoke strong emotions, making participants use offensive remarks. This emotional intensity, combined with Reddit's anonymity, can lead to personal attacks
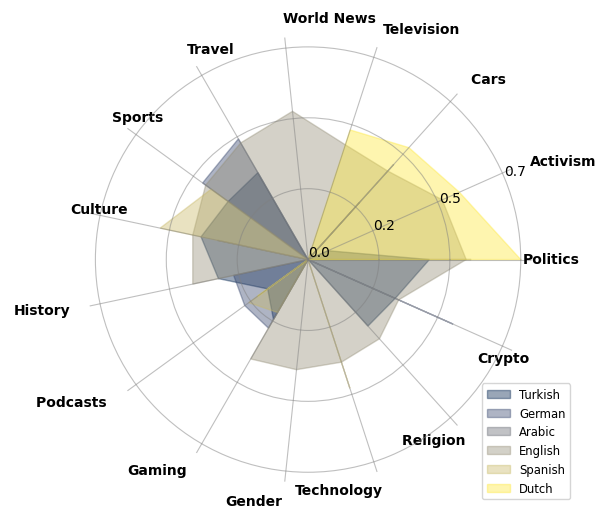

Figure 1: Comparison of toxicity levels in Reddit discussions across different topics and languages. The scores represent the toxicity density, the proportion of toxic comments within each topic. Each line illustrates the toxicity density for a specific language within a particular topic.

and offensive language use. Additionally, the platform's upvote and downvote system can reinforce popular opinions, creating echo chambers dominated by extreme viewpoints. Consequently, the type of topic being discussed might be a central factor for its potential descent into toxicity.

Reddit, as a social platform, has gained significant attention in the area of toxic language research (Baumgartner et al., 2020). The platform offers easy access to data collection in comparison to platforms such as Facebook and Twitter (Baumgartner et al., 2020). It is also reported that there is a significant inclination towards the use of language con-

sidered toxic and offensive (Demszky et al., 2020). This characteristic makes Reddit an ideal platform for studying how toxic language is manifested in various communities. Central to Reddit's structure are subreddits, proxies to communities comprising members who share mutual interests, such as political viewpoints or leisure pursuits. User interaction frequently occurs within these community boundaries around a particular topic.

The prevalence of toxic language on platforms like Reddit has been widely researched. These studies focused on aspects such as individual user comments and posts (Kumar et al., 2023; Hiaeshutter-Rice and Hawkins, 2022), community-level conversations (Farrell et al., 2019), or the behavior of the users (Urbaniak et al., 2022).

In this study, we view toxicity on Reddit from a broader contextual perspective encompassing topic, community, and language. We are specifically interested in how toxicity develops within communities in relation to topics. For this reason, we collect not only specific comments that are likely to be toxic but also the subthreads in which such comments occur, which may also exhibit more nuanced cases of stereotypical targeting, implicit hate speech, irony, and sarcasm within communities. We, therefore, collect conversation threads from Reddit spanning different languages, communities, and topics.[1]

Our analysis shows that toxicity in Reddit conversations strongly depends on the topic of discussion. As shown in Figure 1, certain topics show high toxicity in most of the target languages (e.g., Politics, Sports). In our monolingual analysis, we show that topics that would normally be considered neutral, such as History and Gaming, still have the potential to trigger toxicity. We also observe measurable differences in the toxicity of certain topics across languages. The result of our analysis can be used in several ways. Social media moderators can use the insights from our study for more effective content moderation. Since toxic content is more common in some topics than others, focusing on toxic-prone topics can be more efficient for filtering inappropriate content. It is also important to consider cultural differences. Our analysis shows that topics considered less toxic in one language are more prone to generate toxicity in another. In terms of training models for automatic content moderation, topic and language can be considered part of the context of a comment. This context information can be used in model training for more accurate detection of toxic content.

---

[1]The data and our analysis are available following the US and EU FAIR use principles and according to the license conditions of Reddit on source data. The GitHub repository can be accessed here: https://github.com/cltl/Reddit_topic/tree/main

In summary, our contributions are:

- We collect 1.5 million comment threads from 481 communities in six languages.

- We explore the relationship between toxicity and topics of conversation in mono-lingual and cross-lingual settings across different Reddit communities.

- We compare and contrast three distinct approaches to measure toxicity.

2. Related Work

The social media landscape has become a dynamic arena where users and groups interact, share their diverse viewpoints, and communicate. Within this theme, the occurrence and consequences of toxic language have garnered substantial attention from researchers across various disciplines, such as social sciences, political science, and NLP. Here, we use toxic language as an umbrella term similar to Sharma et al. (2022), broadly comprising hate speech, offensive language, abusive language, propaganda, cyberbullying, and cyber-aggression. In this section, we provide an overview of studies that analyze one or more aspects of toxic language in social media settings from user and community perspectives.

Comment and post analysis Kumar et al. (2023) provide an extensive study of the behavior of accounts on Reddit that post toxic content. The study shows that although accounts engaging in abusive behavior make up less than 4% of Reddit's total users, they are responsible for generating 33% of all comments posted on the platform. Mall et al. (2020) also explore similar user behavior analysis through a temporal analysis of user toxicity and show that the typical behavior of toxic users is switching between toxic and non-toxic commenting. Similar work by Hiaeshutter-Rice and Hawkins (2022) studies the relationship between major political events and hostility in a discussion using language analysis. The findings indicate that U.S. political events led to heightened hostility and increased negativity in Reddit discussions. Urbaniak et al. (2022) study correlation between username toxicity and toxic behavior of these users on Reddit. Users who have toxic usernames generate a greater amount of toxic content compared to those with neutral usernames.

Community analysis Farrell et al. (2019) constructed specific sets of lexicons to systematically study the changes in language use within Reddit communities known for misogynistic discussions. In the context of discussing negative interactions,

as highlighted by Urbaniak et al. (2022) in their work on "namespotting", Kumar et al. (2018) present findings that align with this observation, showing that a small percentage of Reddit communities are responsible for the majority of negative interactions on the platform. Radfar et al. (2020) explore toxicity in Twitter from the user relation perspective and show that tweet exchanges between users without any connection are three times more prone to toxicity than interactions involving mutual friends.

Toxic language resource There are various lexical resources for different languages that define offensive words. Such resources include HurtLex Bassignana et al. (2018), MOL Vargas et al. (2021), DALC Caselli et al. (2021), and Hatebase (hatebase, 2022). HurtLex is a lexicon that covers 50 languages and is divided into 17 categories, including ethnic slurs and derogatory terms, among others. MOL is a lexicon of abusive language annotated with contextual information. It covers English, Spanish, French, German, and Turkish. DALC is a Dutch lexicon of abusive words manually annotated from a Twitter corpus. Hatebase is a crowdsourced resource of hate speech lexicons. Though the Hatebase project was discontinued, the website can be accessed as a browsable archive. NRC lexicon is a manually annotated emotion lexicon for English (Mohammad and Turney, 2013). It includes basic emotions and sentiments, as well as their associated emotions. We specifically consider the NRC lexicon because our interest lies in understanding toxicity in a broader sense. This includes identifying negative sentiments, which are crucial for recognizing instances of implicit hate speech.

Measuring toxicity For quantifying the toxicity of a comment, a widely used approach is Google's Perspective API (Lees et al., 2022) and Detoxify (Hanu and Unitary team, 2020). Perspective is trained on comments to capture the toxicity of a text in various contexts (Salminen et al., 2020). It supports the detection of toxicity, insult, profanity, identity attacks, threats, and sexually explicit content. It covers multiple languages, including Arabic, English, German, Dutch, and Spanish. Detoxify is trained on the jigsaw challenges dataset for toxic comment classification (Hanu and Unitary team, 2020). It supports English, French, Spanish, Italian, Portuguese, Turkish, and Russian.

Topic and language analysis A study by Salminen et al. (2020) explores the relationship between toxicity and news topics. The results show that discussions related to racism, Israel-Palestine, and war exhibit higher toxicity in comments. It also shows instances of a typically less toxic topic that becomes more toxic when politics and religion are involved. A similar analysis by Hilte et al. (2023) analyzes profiles of users who post toxic content in different languages such as English, Dutch, Slovenian, and Croatian. Both of these works are similar to our work in using topics to analyze toxicity. In comparison, our work can be considered complementary as we include a broader range of topics and more languages in our analysis.

3. Methodology

3.1. Data source

We collected a total of 1.5 million comments in 80 topics and six languages. Each of the comments includes a timestamp, an anonymized username, the subreddit, the topics of the subreddit, the submission in which the comment was posted, the submission title, and the body. We also include graph data that enables us to reconstruct the thread structure. Ultimately, we are interested in analysing subthreads that have a high chance of exhibiting both implict and explicit toxic behaviour.

We anonymized the author's personal information according to GDPR regulations. We first identify user names from the author name attribute of our collected metadata. We then replaced each identified user name with a unique and non-descriptive identifier consisting of a random string and numerical code to remove any connection to the individual.

#Language	6
# Topics	80
# Communities	481
# Submissions	39,249
#Unique Users	511,464
# Comments	1,543,272

Table 1: Statistics for the collected data. Communities refer to the subreddit. Threads are all the comments under the same submissions or posts.

3.2. Data collection and preprocessing

We use PRAW[2], the Reddit Python package, to collect the data. We first extract lists of subreddits from the Reddit community ranking page.[3] The website contains subreddits ranked by the number of subscribers.

Language detection The Reddit API doesn't provide language information about the subreddit. To identify the language of a subreddit, we use Google

---

[2]https://praw.readthedocs.io/en/stable/index.html
[3]https://www.reddit.com/best/communities/1/

Translate API to automatically classify the description of the subreddit into target languages. Since Reddit is predominately used in the English speaking communities, the most popular subreddits are in English. To create a balanced list of subreddits for each of our target languages, we create a new list from the initial list by sampling an equal number of subreddits per language. We then collect posts and comments from each subreddit. We query 100 popular submissions for each subreddit based on the upvote count. We then collected all the comments under these popular submissions. This initial list contains 178K subreddits. Table 1 shows the main statistics of the collected data.

Topic identification In order to determine the topic of a specific subreddit, we employ a different approach. Since the Reddit API does not provide information about a subreddit's topic, we undertake a separate web crawl from the Reddit community ranking page. This allows us to associate each subreddit with its corresponding topic category.

Pre-processing We excluded comments that are either shorter than 15 characters or longer than 300 characters in length or comments which contains only emojis or punctuation. This decision aligns with previous research addressing the limitations of applying existing toxicity models to short, excessively long or noisy texts (Kumar et al., 2023).
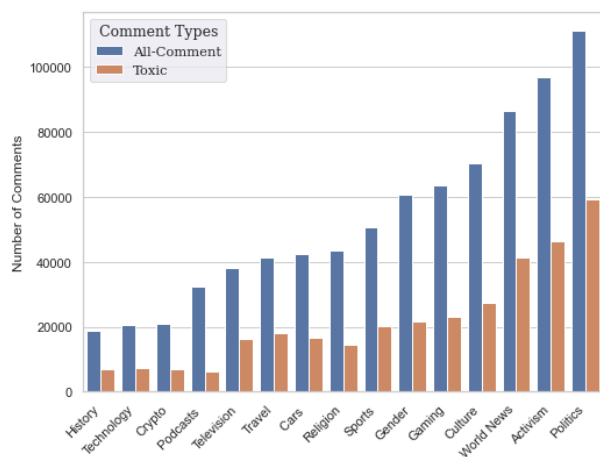

Figure 2: Distribution of toxic comments across topics based on the lexical-based approach. For visibility, we show the top 15 topics. Here, a comment is considered toxic if it contains at least one toxic word.

3.3. Toxicity scores

We stress that our ultimate goal is to create a dataset across communities in which we can find subreddit threads with a high probability of exposing toxic language and, specifically, hate speech. This should contain cases of not only explicit but also implicit hate speech. Because it is more difficult to find implicit hate speech, we are interested in a method that has high recall of finding toxicity comments so that we can further analyse the subthreads in which these occur. To decide on a high-recall method, we conduct a manual assessment of three methods to identify comment toxicity, focusing on those with the broadest applicability to our target languages. These methods include the Perspective API, a lexicon-based approach, and OpenAI's GPT-4.

3.3.1. Lexicon-based approach

For the lexicon-based approach, we combine HurtLex, MOL, DALC, Hatebase, and NRC and build a binary classifier to score the toxicity of a comment. For the NRC lexicon, we ignore the emotion layer and only used words that are associated with negative sentiment. If at least one toxic word is present in a comment, we consider it toxic. Lexicon-based approaches have shown to be robust when detecting toxic words in cross-domain settings (Schouten et al., 2023) and can easily be extended to other languages or adapted in the future. Our merged lexicon has 4,316 English, 7,041 Dutch, 1,831 Arabic, 2,782 Turkish, 2,903 Spanish, and 2,851 German words.

3.3.2. GPT-4

For GPT-4, we employ a simple zero-shot prompt to assign toxicity labels to a comment. We include a definition of what a toxic comment is in the prompt. We prompt GPT-4 to classify comments as toxic if it is hate speech, offensive language, abusive language, propaganda, cyberbullying, or cyber-aggression or non-toxic otherwise. Our prompt is "Review each comment and label it as toxic or non-toxic. To determine whether the comment is toxic if the comment falls into any of the following categories: hate speech, offensive language, abusive language, propaganda, cyberbullying, or cyber-aggression. If the comment aligns with any of these categories, label it as 'Toxic' in the label column. If the comment does not fit any of these categories, label it as non-toxic ".

3.3.3. Perspective API

Perspective API is an out-of-the-box toxicity classifier from Google. The API takes a comment as input and produces a score between 0 and 1 for different toxicity categories, such as threats, profanity, and identity attacks. Since we are interested in an aggregate score, we use the toxicity attribute

to get a single score. Based on a recommendation from the API documentation, we use a threshold value of 0.75, and we consider a comment toxic if its toxicity score is higher than this threshold value.

3.3.4. Expert annotation

We conduct an expert annotation to identify the most effective method for detecting toxic comments. Our goal is to evaluate the performance of the identified approaches, particularly focusing on high recall. We randomly sampled 500 comments from each language from our dataset. We prepared annotation guidelines with the definition of what kind of comment should be labeled as toxic. Our definition of toxic comment comprises hate speech, offensive language, abusive language, propaganda, cyberbullying, and cyber-aggression. We selected native speakers as subject matter experts. The annotators classified comments as toxic or not toxic based on the provided guidelines. We resolved questions and discrepancies through discussion. The languages covered in this paper include German, Turkish, Spanish, Dutch, Arabic, and English.

3.3.5. Thread toxicity

We use this metric to compute the toxicity of a thread (instead of single comment), where thread refers to all the comments that are part of a single submission. This analysis gives a more robust estimation of toxicity since a thread can have multiple comments from different users. To do this, we first reconstructed the thread structure of the comment from our dataset. We then filter threads with at least ten comments before computing the thread toxicity.

3.3.6. Topic Toxicity

We define topic toxicity as the proportion of toxic comments on a specific topic relative to the total number of comments on that topic. We computed topic toxicity for each topic in the target language.

4. Results and Analysis

In this section, we present the main findings. We divide our analysis into three parts. First, we compare the performance of the different methods in detecting toxic comments based on the test data we created. We then explore the relationship between toxicity and topics in aggregate and for each language separately. Finally, we analyze how consistent a topic toxicity is across languages by comparing the toxicity results across the six languages covered in this paper.

4.1. Evaluation of approaches

We present the result of the evaluation of the three approaches in Table 2. In the aggregate results, we observe a significant difference across the approaches. The lexical-based approach significantly outperforms both Perspective-API and GPT-4 in terms of recall of toxic comments (respectively .53, .08 and .08), whereas Perspective outperforms to the others in precision (.35 versus .17 lexical and .08 GPT-4). Similarly, the cross-lingual analysis shows that the lexical approach consistently has the highest recall in the toxic category, indicating that this approach is the most effective in identifying toxic content with high recall across languages. We do see some differences between languages, as the precision scores for Dutch and German using the lexical approach are significantly lower.

As we stated before, we prioritize recall over precision for our analysis because we want to maximize the probability that we find threads that exhibit explicit or implicit toxicity. Toxic comments are rare compared to non-toxic ones (Vidgen and Derczynski, 2020). We aim to flag potential toxicity broadly on this first pass to ensure that any potential toxic content is not missed, accepting the false positives.

4.2. Topic toxicity

In this section, we analyze topic toxicity in aggregate. We first identify the top 15 topics from the 80 topics based on the number of comments. Figure 2 shows the distribution of toxic and non-toxic comments for the top 15 topics. In the distribution, politics-related topics such as Politics, Activism and news-related topics like World News have a higher number of toxic comments. For a more accurate comparison of the toxicity of topics, we computed the topic toxicity for each topic as described in the methodology section. Since the topic toxicity is a normalized value, it is possible to directly compare this value across topics.

Similar to the distribution, we found topics related to Politics and World News to have the highest topic toxicity. This is partially consistent with the results reported by (Salminen et al., 2020), which shows that topics related to Politics and News are highly likely to generate toxic conversation. In contrast, we also observe high toxicity in less expected topics such as Travel and History.

4.3. Distribution of toxic threads

As described in the methodology section, we use thread toxicity for a more accurate estimation of toxicity. The thread toxicity provides an aggregate score rather than relying on the toxicity score of a single comment. In this analysis, we first group comments into different comment threads using the

		Lexical			Perspective			Gpt-4			Support
		P	R	F1	P	R	F1	P	R	F1	
	Non toxic	.90	.62	.74	.88	.98	.93	.87	.87	.87	1315
	Toxic	.17	.53	.25	.35	.08	.13	.08	.08	.08	190
	Macro avg	.53	.57	.49	.61	.53	.53	.48	.48	.48	1505
DE	Non toxic	.97	.46	.63	.96	.93	.95	.94	.94	.94	240
	Toxic	.07	.69	.12	.31	.24	.24	.31	.31	.31	13
	Macro avg	.52	.58	.37	.58	.62	.59	.47	.47	.47	253
ES	Non toxic	.88	.46	.61	.82	.99	.88	.86	.86	.86	178
	Toxic	.30	.79	.44	.82	.17	.28	.29	.29	.29	53
	Macro avg	.59	.63	.52	.81	.58	.58	.53	.53	.53	231
NL	Non toxic	.97	.38	.55	.94	1.00	.97	.96	.96	.96	252
	Toxic	.08	.81	.14	.00	.00	.00	.12	.12	.12	16
	Macro avg	.52	.60	.35	.47	.50	.48	.54	.54	.54	268
AR	Non toxic	.88	.94	.91	.86	1.00	.92	.86	.86	.86	457
	Toxic	.41	.24	.30	.00	.00	.00	.00	.00	.00	75
	Macro avg	.65	.59	.61	.43	.50	.46	.43	.43	.43	532
EN	Non toxic	.86	.51	.64	.85	.95	.90	.84	.84	.84	188
	Toxic	.16	.55	.25	.17	.06	.09	.06	.06	.06	33
	Macro avg	.51	.53	.45	.51	.50	.49	.47	.47	.47	221
TR	Non toxic	.69	.57	.62	-	-	-	.6	.87	.71	180
	Toxic	.49	.61	.54	-	-	-	.37	.12	.18	120
	Macro avg	.59	.59	.58	-	-	-	.48	.49	.44	300

Table 2: Evaluation of Lexical-based approach and Perspective API. The first three rows show the aggregate result for all languages, followed by a language-specific breakdown. Here, we put '-' since Perspective doesn't support Turkish. We also exclude Turkish in the aggregate result of the first three rows.
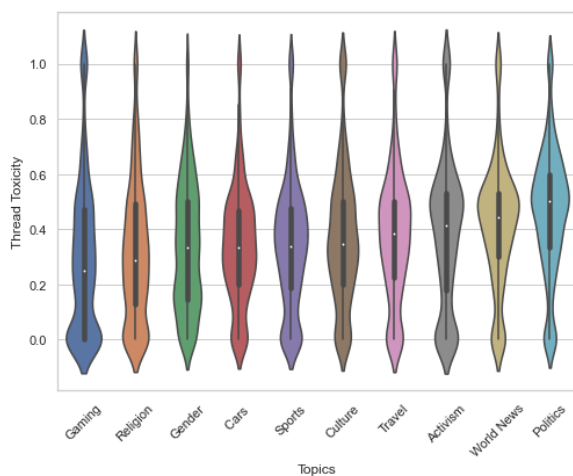

Figure 3: Distribution of thread toxicity across topics. For visibility, we only show the top 10 topics. Plots are sorted by the mean value.

parent-child relationship of comments and submissions. We then compute the thread toxicity for each of the threads. Comment threads with more toxic comments will have a score close to 1, and threads with less toxic comments will have a value close to 0, as shown in Figure 3.

The y-axis represents the toxicity level, ranging from 0 to 1, and the x-axis shows different Reddit topics. Each violin shape provides a density estimate of the data at different toxicity levels. The wider a section of the violin, the higher the density of threads at that toxicity level. Here, we notice that Politics, World News, and Activism have a higher mean toxicity score and a greater number of threads with high toxicity scores. A dense concentration of toxic thread for Activism shows a broad dispersion, with a high density in the upper quartile, indicative of the potential contentiousness of discussions on this topic. In World News, while there is a significant central tendency around the median, a non-negligible spread towards the upper toxicity range is evident. Lastly, Politics is characterized by its extensive variance and significant density at the toxicity scale's lower and upper bounds.

4.4. Monolingual topic-toxicity

We compute the topic toxicity per language to analyze which topics stand out as more toxic than others in each language. Table 3 shows each language's top five toxic topics based on topic toxicity. Since the topic toxicity is a normalized value, we can use it to compare topic toxicity within and across languages.

English comments have the highest toxicity in Politics and Activism. In terms of intensity, conversations related to politics and news have the highest toxicity. Similar to the aggregate result,

Arabic	Turkish	Spanish	German	Dutch	English
Politics	**Politics**	**Culture**	Crypto	**Politics**	**Politics**
Culture	**Culture**	Technology	Travel	**Activism**	News
Cars	Travel	**Sports**	**Sports**	Cars	**Activism**
Podcasts	**Sports**	Podcasts	Cars	Television	Travel
Activism	Crypto	Gaming	Gaming	Podcasts	**Sports**

Table 3: List of top five topics that have the highest topic toxicity score in each language. Topics that are toxic in more than two languages are shown in bold.
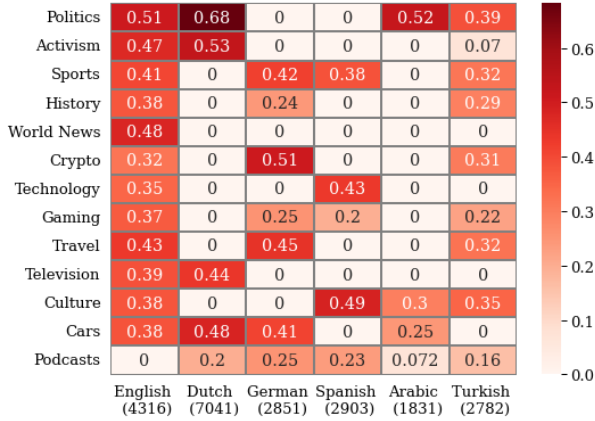

Figure 4: Toxicity scores using the lexicon-based approach. The number under each language shows the total number of lexicon entries in that language.

this result is partially in line with (Salminen et al., 2020). Contrary to (Salminen et al., 2020), discussions related to religion do not have high toxicity in our analysis. For Arabic conversations, partially similar to English, we observe high toxicity in discussions related to politics, such as Politics and Activism. We also observe high toxicity in discussions that involve Culture. In German, contrary to English and Arabic conversations, we observe high toxicity in more unexpected topics such as Crypto, Travel, and Cars. Similar to English and Arabic, Dutch conversation has the highest toxicity in political conversations. In Spanish, similar to German, the toxicity is concentrated in less expected topics such as Technology and Gaming.

In summary, conversations in Politics and Sport consistently show high toxicity in four out of the six target languages. We also observe high topical-toxicity patterns in Culture (Spanish, Arabic, and Turkish) and Gaming (Spanish and German). In the next section, we expand on a cross-lingual toxicity analysis for topics shared across the target languages.

4.5. Cross-lingual toxicity analysis

For cross-lingual analysis, we select topics that are shared by at least two languages. Figure 4

shows a Heatmap of toxicity across the selected topics and languages.

4.5.1. Consistent toxicity in politics

Politics, one of the cross-lingual topics shared by English, Dutch, Arabic, and Turkish, shows the most consistent toxicity in English, Dutch, and Arabic. In terms of intensity, we observe that it is more toxic in Dutch than in the other languages we analyzed. In general, we observe a similar pattern of toxicity with variation in intensity.

4.5.2. Diversity in toxic topics

While some languages like Dutch and Arabic show high toxicity in topics such as Politics and Activism, others like German demonstrate high toxicity in seemingly neutral topics like Crypto and Travel. The Spanish conversations tend to express stronger reactions when discussing culture and ethnicity. English and Turkish languages show a more diverse picture; comments in these languages display varied toxicity levels across multiple topics. This suggests that users in these languages have a broader range of subjects that elicit strong, potentially toxic responses. The results underscore the cultural and linguistic nuances in how different topics are perceived and discussed across languages.

5. Conclusion

Our findings support prior research emphasizing the relationship between topics and the toxicity of a comment. We broaden this correlation to encompass a broad range of topics and languages. In the aggregate analysis, we found conversations that involve politics and news to have the highest toxicity, which is partially consistent with the results reported by Salminen et al. (2020). In contrast, we also observe high toxicity in less-expected topics such as travel and history. In monolingual analysis, we demonstrate that conversations in Politics and Sports consistently show high toxicity in the majority of our target languages. We also observe such topical-toxicity patterns in Culture, Ethnicity,

and Gaming. Furthermore, we observe major differences across languages in relation to the topics. Whether these differences also correspond to variations in community dynamics cannot be determined from the current data. Further investigation is required to answer to what extent these language communities actually discuss the same things within the broader topic clusters. In future research, we want to analyze the topics of the subreddits in more detail using entity recognition and topic classification in comparison to similar time frames to further compare the content across languages. Similar entities and topics in similar periods could be used as an indication of parallel discussion across communities that potentially exhibit different toxicity. Furthermore, we want to analyze the build-up of toxicity within the thread and also focus on the targets of such language and implicit hate speech instances in our dataset.

5.1. Limitations

We identify some limitations in our work. First, using topics to categorize a subreddit can oversimplify the rich nuances of a conversation that may take place in a particular community. Many conversations may not clearly fit into one topic, often overlapping with multiple topics. These conversations are also dynamic in nature, with threads evolving and branching into subtopics. A static categorization might not capture the fluidity of these discussions. The level of detail within a topic is another factor to think about, as certain topics can be overly general while others are highly specific. Finding the right balance between granularity and generality in categorization is challenging. The lexicons we use for computing the toxicity also have a limitation. The variation in the quality and quantity of lexicon items for each language might lead to results that favor certain languages over others.

5.2. Ethical consideration

In this paper, we use information collected from the Reddit platform, a public online platform where users post content and take part in discussions. We recognize and emphasize the importance of ethical considerations when handling and analyzing such datasets. Firstly, all data used were publicly accessible and did not involve any private or confidential information. We take all the necessary steps according to GDPR regulations to anonymize any identifiable user information to ensure privacy. Furthermore, we use the collected data strictly for research purposes, and no attempt was made to exploit, manipulate, or otherwise use the data in a manner that could harm or prejudice any individual or group. Any insights drawn from this work are based only on patterns in the data and should

not be used to stereotype or make generalizations about specific groups or individuals.

6. Bibliographical References

Hind Almerekhi, Haewoon Kwak, Bernard J. Jansen, and Joni Salminen. 2019. Detecting toxicity triggers in online discussions. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, page 291–292, New York, NY, USA. Association for Computing Machinery.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1217–1230, New York, NY, USA. Association for Computing Machinery.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE*, 6.

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 87–96, New York, NY, USA. Association for Computing Machinery.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Ine Gevers, Ilia Markov, and Walter Daelemans. 2022. Linguistic analysis of toxic language on social media. *Computational linguistics in the Netherlands journal*, 12:33–48.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

hatebase. 2022. Hatebase.

Jack Hessel and Lillian Lee. 2019. Something's brewing! early prediction of controversy-causing posts from discussion features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1648–1659, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Hiaeshutter-Rice and Ian Hawkins. 2022. The language of extremism on social media: An examination of posts, comments, and themes on reddit. *Frontiers in Political Science*, 4:805008.

Lisa Hilte, Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2023. Who are the haters? A corpus-based demographic analysis of authors of hate speech. *Frontiers in Artificial Intelligence*, 6:986890.

Deepak Kumar, Jeff Hancock, Kurt Thomas, and Zakir Durumeric. 2023. Understanding the behaviors of toxic accounts on reddit. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2797–2807, New York, NY, USA. Association for Computing Machinery.

Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 933–943, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Bumsuk Lee. 2012. A temporal analysis of posting behavior in social media streams. *Proceedings of the International AAAI Conference on Web and Social Media*.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA. Association for Computing Machinery.

Raghvendra Mall, Mridul Nagpal, Joni Salminen, Hind Almerekhi, Soon-Gyo Jung, and Bernard J. Jansen. 2020. Four types of toxic people: Characterizing online users' toxicity over time. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, NordiCHI '20, New York, NY, USA. Association for Computing Machinery.

Ilia Markov, Ine Gevers, and Walter Daelemans. 2022. An ensemble approach for Dutch cross-domain hate speech detection. In *Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings*, page 3–15, Berlin, Heidelberg. Springer-Verlag.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI Conference on Artificial Intelligence*.

Alexey Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. 2019. *The Anatomy of Reddit: An Overview of Academic Research*, pages 183–204.

Amirhossein Nadiri and Frank W. Takes. 2022. A large-scale temporal analysis of user lifespan durability on the reddit social media platform. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 677–685, New York, NY, USA. Association for Computing Machinery.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:13.

Nicholas Proferes, Naiyan Jones, Sarah A. Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, 7.

Bahar Radfar, Karthik Shivaram, and Aron Culotta. 2020. Characterizing variation in toxic language by social context. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:959–963.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

9

Joni O. Salminen, Sercan Şengün, Juan Corporan, Soon-Gyo Jung, and Bernard Jim Jansen. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLoS ONE*, 15.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Martin Saveski, Brandon Roy, and Deb Roy. 2021. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, WWW '21, page 1086–1097, New York, NY, USA. Association for Computing Machinery.

Stefan F. Schouten, Baran Barbarestani, Wondimagegnhue Tufa, Piek Vossen, and Ilia Markov. 2023. Cross-domain toxic spans detection. In *Natural Language Processing and Information Systems: 28th International Conference on Applications of Natural Language to Information Systems, NLDB 2023, Derby, UK, June 21–23, 2023, Proceedings*, page 533–545, Berlin, Heidelberg. Springer-Verlag.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar I. Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Y. Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *International Joint Conference on Artificial Intelligence*.

Rafal Urbaniak, Patrycja Tempska, Maria Dowgiałło, Michał Ptaszyński, Marcin Fortuna, Michał Marcińczuk, Jan Piesiewicz, Gniewosz Leliwa, Kamil Soliwoda, Ida Dziublewska, Nataliya Sulzhytskaya, Aleksandra Karnicka, Paweł Skrzek, Paula Karbowska, Maciej Brochocki, and Michał Wroczyński. 2022. Namespotting: Username toxicity and actual toxic behavior on reddit. *Comput. Hum. Behav.*, 136(C).

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Hanh Vo, Hieu Tran, and Son Luu. 2021. Automatically detecting cyberbullying comments on online game forums. pages 1–5.

Steven Windisch, Susann Wiedlitzka, and Ajima Olaghere. 2021. Protocol: Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell Systematic Reviews*, 17(1).

Moran Yarchi, Christian Baden, and Neta Kligler-Vilenchik. 2020. Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, 38:1–42.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

7.   Language Resource References

Bassignana, Elisa and Basile, Valerio and Patti, Viviana and others. 2018. *Hurtlex: A multilingual lexicon of words to hurt*. CEUR-WS. PID https://github.com/valeriobasile/hurtlex.

Caselli, Tommaso and Schelhaas, Arjan and Weultjes, Marieke and Leistra, Folkert and van der Veen, Hylke and Timmerman, Gerben and Nissim, Malvina. 2021. *DALC: the Dutch abusive language corpus*. PID https://github.com/tommasoc80/DALC.

Mohammad, Saif and Turney, Peter. 2013. *Crowdsourcing a Word-Emotion Association Lexicon*. PID https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

Vargas, Francielle and Rodrigues de Góes, Fabiana and Carvalho, Isabelle and Benevenuto, Fabrício and Pardo, Thiago. 2021. *Contextual-Lexicon Approach for Abusive Language Detection*. INCOMA Ltd. PID https://github.com/franciellevargas/MOL.