Report on analysis of the data collected from 4chan and Reddit

Mukhil Venkataramanan
SUNY Binghamton
Binghamton, USA
mvenkatarama@binghamton.edu

Gurusaran Venkatachalam Rajarajacholan
SUNY Binghamton
Binghamton, USA
gvenkatachal@binghamton.edu

Sai Prakash
SUNY Binghamton
Binghamton, USA
snalubolu@binghamton.edu

ABSTRACT

This report explains the methodologies and experiments designed to analyze the data collected from various Reddit's subreddits and 4chan's boards. The primary focus is on measuring toxicity, analyzing sentiment trends, and understanding user engagement in discussions of various topics related to politics and sports. The analysis will utilize the Moderate Hatespeech API to assign toxicity scores in real-time, while sentiment analysis will be conducted using natural language processing (NLP) techniques. Key trends in posting frequency, sentiment evolution, and the correlation between engagement and toxicity will be examined. The proposed research aims to uncover valuable insights into the nature of online interactions in sports communities, provid ing a foundation for deeper exploration in future work. The analysis will also address three primary research questions concerning toxi city comparisons, the relationship between user engagement and toxicity, and sentiment trends across the platforms

KEYWORDS

Data Collection, Reddit API, 4chan API, Sports Subreddits, Report, Implementation, PostGreSQL, Faktory, Docker, Python

1 INTRODUCTION

The proliferation of data on social media platforms such as Reddit and 4chan provides a rich source of information for real-time analysis and insights. These are two of the most influential online platforms, playing a significant role in shaping public opinions, and trends, especially within niche communities. Collecting data from these platforms provides us a wealth of user-generated content that is highly unfiltered and diverse. The real-time, organic conversations that take place on these platforms offer invaluable insights into not only what people are discussing but also how ideas and sentiments spread across different demographics.

Motivation: The motivation for this project stems from the increasing influence of online platforms like Reddit and 4chan on public discourse. While Reddit represents a semi-moderated environment, 4chan offers a uniquely anonymous space, often linked to controversial discussions and behaviors. Understanding the dynamics of these platforms, particularly the nature of toxic and non-toxic content, can provide valuable insights into online community behavior, misinformation spread, and digital culture trends.

The broader impact of this work lies in its potential applications in content moderation, public policy, and social media research. By analyzing the prevalence and patterns of toxic content, this study contributes to creating safer online environments and equipping stakeholders with tools to address harmful behaviors. Additionally, this research enhances our understanding of user behavior in anonymous versus semi-anonymous platforms, aiding in the design of better communication and engagement strategies.

2 BACKGROUND AND RELATED WORK

2.1 Existing Literature Review:

There is a rich body of literature on sentiment analysis and toxicity detection in online platforms. Works like Davidson et al.'s "Automated Hate Speech Detection and the Problem of Offensive Language" have laid the foundation for toxicity classification. Similarly, Waseem and Hovy's research on abusive language detection provides insights into leveraging machine learning models for text-based analysis. Studies on Reddit, such as Chandrasekharan et al.'s "The Internet's Hidden Gems: Empirical Research on Reddit", explore its user dynamics and moderation impacts. However, literature on 4chan remains relatively sparse due to its highly anonymous and transient nature.

2.2 Existing Frameworks and Tools:

- Hugging Face Transformers for sentiment analysis using state-of-the-art models like BERT and RoBERTa.
- Tools like NLTK and Scikit-learn for natural language processing and feature extraction.
- APIs for structured data collection from Reddit (PRAW) and 4chan's JSON endpoints.

2.3 Comparable Works:

Several studies inspired this project, including Matsuba et al.'s research on 4chan's role in spreading misinformation and works analyzing political discourse on Reddit. This project extends their scope by incorporating a cross-platform analysis and focusing on sentiment and toxicity patterns.

By building on existing methods and adapting them to the unique challenges of 4chan and Reddit, this project contributes novel insights into the nature of online discourse.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 DATASET DESCRIPTION

3.1 Reddit

For the posts table:

- **Post ID:** Unique identifier for each post. This is essential for tracking and avoiding duplicate entries.
- **Title:**: The title gives a brief description of the post's content, making it crucial for text analysis. It also allows for easy post retrieval based on search queries or keywords.
- **Content (selftext)::** The main body of the post, which contains the detailed text that users write. It is important for analyzing the actual discussions happening in a subreddit. By storing the full content, we can extract insights like common topics, user interests, or text sentiment.
- **Author:** The username of the person who made the post. Collecting author information helps to analyze user behavior patterns, track the activity of specific individuals, and correlate authorship with post quality, engagement, or popularity.
- **Score:** The post score represents the total number of upvotes minus downvotes. This is a key metric for understanding community engagement and gauging the overall popularity or reception of a post.
- **Comments:** Number of comments on the post. The comment count indicates the engagement level of the community with a particular post, providing insight into the relevance and discussions generated by the content.
- **Created At:** This is the timestamp indicating when the post was created. This field is essential for time-series analysis, understanding patterns of activity over time, and examining trends within a subreddit.
- **Subreddit:** The specific subreddit where the post was made. This is important for categorizing data and allowing us to compare patterns across different communities.
- **URL:** The direct URL to the post, useful for retrieval, validation, or manual inspection. Having this stored allows for easy referencing of the original content on Reddit.
- **Toxic Class** Takes the value of 'flag' or 'normal'. Says whether the contents in the reddit posts are toxic or not
- **Toxic Score** Takes the values between 0 and 1. Says the numerical value for confidence level by which the text is classified as 'flag' or 'normal'
- **Toxic Flag** Takes in values as True/False. Marks True if the toxic class is 'flag' and the toxic score is more than 0.9, else False.

For the comments table:

- **Post ID:** The foreign key connecting the comment to its parent post. This establishes a relationship between posts and their comments, allowing us to perform queries that analyze comment activity on a per-post basis.
- **Comment ID:** Unique identifier for each comment. Similar to the post ID, this ensures that each comment is uniquely identifiable and prevents duplication.
- **Author:** The username of the commenter. Tracking who is making comments can reveal patterns in engagement and contribute to user-based analysis, such as identifying frequent contributors or potential influencers.
- **Body** The text of the comment. This allows for sentiment analysis and deeper textual analysis of discussions happening in response to posts. It's crucial for understanding the nature of conversations around posts.
- **Score:** The score of the comment, representing upvotes minus downvotes. Just like post score, this helps gauge the community's reaction to individual comments, allowing for an understanding of which comments contribute most to the discussion.
- **Created At:** Timestamp indicating when the comment was made. This field is vital for studying patterns over time, understanding how quickly discussions grow after a post is made, and performing time-based analysis.
- **Toxic Class** Takes the value of 'flag' or 'normal'. Says whether the contents in the reddit comemnts are toxic or not
- **Toxic Score** Takes the values between 0 and 1. Says the numerical value for confidence level by which the text is classified as 'flag' or 'normal'
- **Toxic Flag** Takes in values as True/False. Marks True if the toxic class is 'flag' and the toxic score is more than 0.9, else False.

3.2 4chan

For posts table:

- **Board:** This field indicates which specific 4chan board (such as /pol/, /b/, or /g/) the post belongs to. Each board has a distinct community, rules, and content focus. Collecting this data allows for filtering and analyzing posts by the board, which helps understand different discussion trends, content types, and user engagement across various topics. It's essential to identify patterns and differences in discourse depending on the context of the board.
- **Thread Number:** A thread on 4chan serves as a container for a series of related posts. The thread number is vital for grouping posts together, as it ensures that all related posts can be analyzed as a single conversation. This data can be used for analyzing the lifespan of a thread, how quickly discussions evolve, or how viral certain threads become. It's also crucial for identifying influential threads within a board.
- **Post Number:** The post number provides a unique identifier for each individual post within a thread. This field is necessary for maintaining the hierarchy of posts, especially when analyzing post reply chains or creating visualizations of user interactions within threads.
- **Data**: The content of the post (or "data") includes the text, images, or other media shared by users. This is the most important field for sentiment analysis, text mining, topic modeling, and other content analysis techniques. Capturing the post data enables a deep dive into what users are discussing, the tone of their interactions, and emerging themes or controversial topics within the boards.
- **Toxic Class** Takes the value of 'flag' or 'normal'. Says whether the contents in the 4chan posts are toxic or not
- **Toxic Score** Takes the values between 0 and 1. Says the numerical value for confidence level by which the text is classified as 'flag' or 'normal'

- **Toxic Flag** Takes in values as True/False. Marks True if the toxic class is 'flag' and the toxic score is more than 0.9, else False.

4 DATA ANALYSIS

4.1 4chan

S.No	Board	Count
1	pol	5185451
2	sp	3079416

Table 1: Posts per board in 4chan

The Figure 1 shows that more data is collected from /pol board than /sp. The board /pol is quite popular than /sp as it receives more posts.

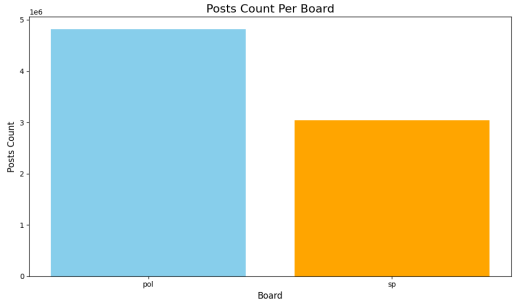

Figure 1: Posts per board

The Figure 2 clearly shows that there is increase in the uses activity and posting in the 4chan around the dates of election and its results.

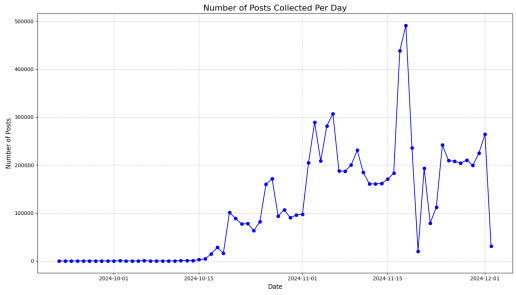

Figure 2: Posts in 4chan per day

The Figure 3 clearly shows that there is increase in the user activity and posting in the 4chan around the dates of election and its results.

The Figure 4 shows us the 4chan posts than are flagged as toxic and not flagged as toxic from moderate hate speech API. Clearly

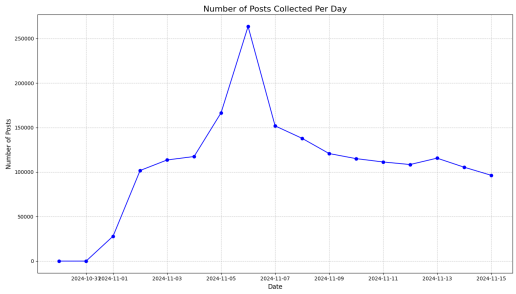

Figure 3: Posts in per day in pol board

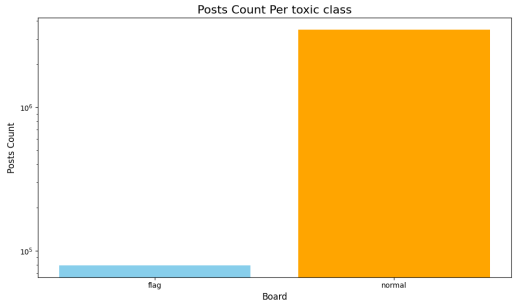

Figure 4: Flag vs Normal class

there are more normal posts than toxic posts.

The Figure 5 out of all the posts collected per day, how many are flagged as toxic and non toxic posts. We can see the increase in toxicity around the dates of the US presidential election.

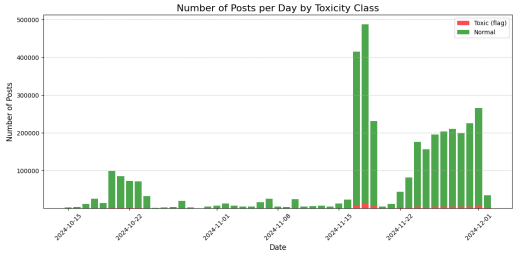

Figure 5: Posts per day by toxicity class

The Figure 6 shows the scatter plot of engagement vs toxicity in 4chan posts. We infer that if the engagement score is more, the post doesn't have to be toxic. Less toxic posts also get a higher engagement score.

The Figure 7 shows the comments collected from the 4chan pol board. More comments are seen during the dates of US presidential election.

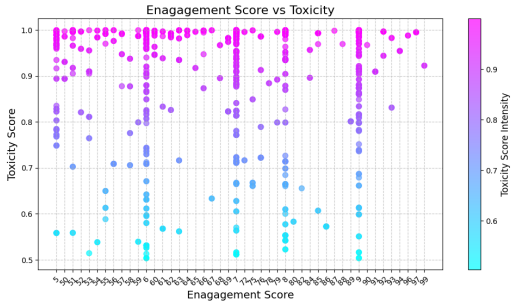

Figure 6: Engagement vs toxicity

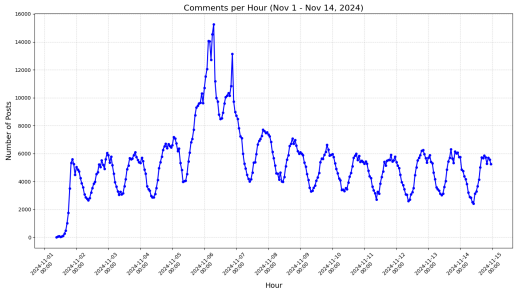

Figure 7: Comments per hour in /pol board

The Figure 8 shows average toxic and non-toxic scores of posts collected from board pol over the time.

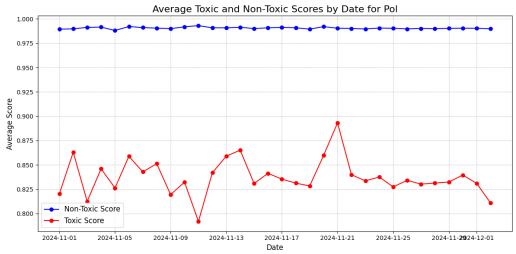

Figure 8: Average toxic vs non toxic scores in 4chan posts

The Figure 9 shows sentiment distribution of the top 100 most engaging posts collected in 4chan over the time

4.2 Reddit

The Figure 10 shows that the r/politics have most posts collected over the time.

The Figure 11 shows that the r/politics have most comments collected over the time.

The Figure 12 shows that there is peak user activity during the time of US presidential elections and during the Indian cricket test series.

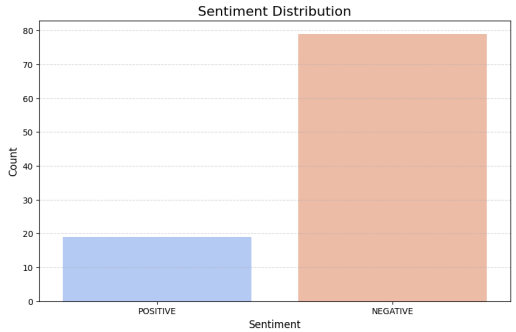

Figure 9: Sentiment distribution of trending posts

S.No	Subreddit	Posts Count	Comments Count
1	CFB	3150	91407
2	Cricket	1925	129743
3	football	657	9236
4	formula1	1731	46905
5	mlb	628	13432
6	nba	3895	69940
7	politics	8230	132543
8	soccer	6352	99216
9	sports	1058	20476
10	tennis	2104	36165

Table 2: Subreddit Posts and Comments Count

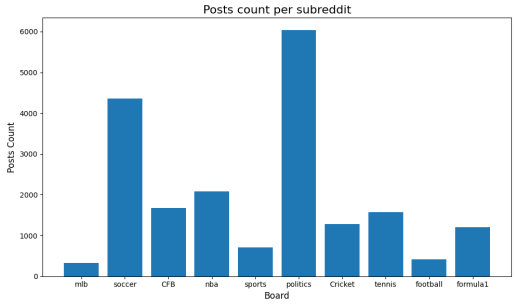

Figure 10: Posts per subreddit

The Figure 14 shows that more toxic comments are collected compared to toxic comments.

The Figure 16 shows that toxic posts collected per day is relatively small compared to nontoxic posts.

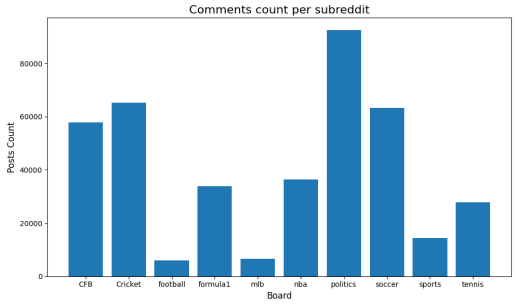

Figure 11: Comments per subreddit

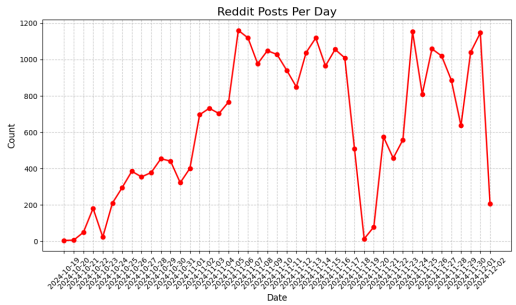

Figure 12: Posts per day

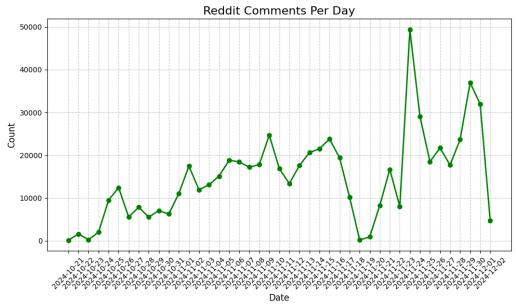

Figure 13: Comments per day

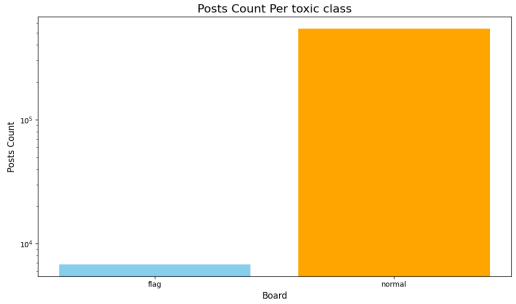

Figure 14: Comments - flag vs normal distribution

The Figure 16 shows that toxic posts collected per day is relatively small compared to nontoxic comments.

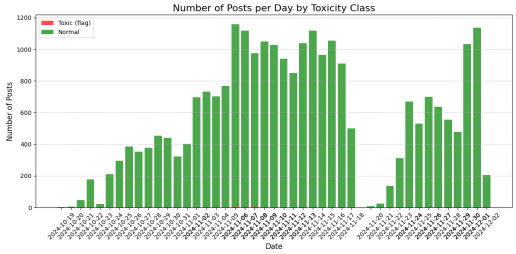

Figure 15: Posts per day flag vs normal

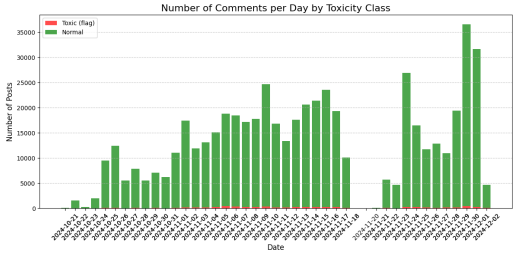

Figure 16: Comments per day flag vs normal

The Figure 17 shows that most of the collected posts are non-toxic irrespective of whether they have a higher engagement value or not.

Figure 17: Engagement vs toxicity

The Figure 18 shows the average scores of toxic and non toxic comments in Reddit.

The Figure 19 shows that there is more user activity on the days of US election

The Figure 20 shows that there is more user activity on the days of US election

The Figure 21 shows that there is more user activity on the days of US election

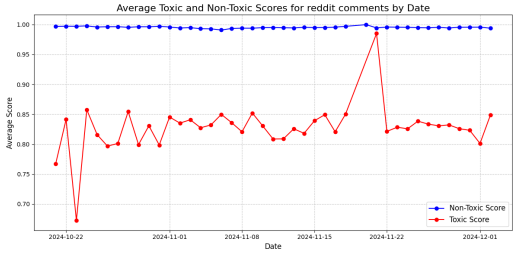

Figure 18: Average toxic vs non-toxic scores of reddit comments

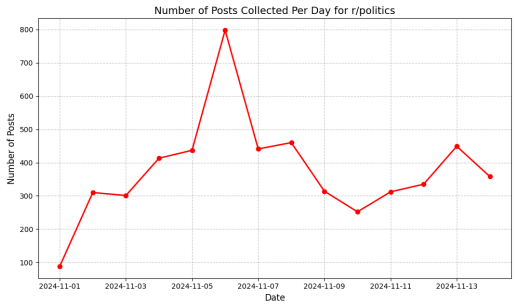

Figure 19: Posts per day in r/politics

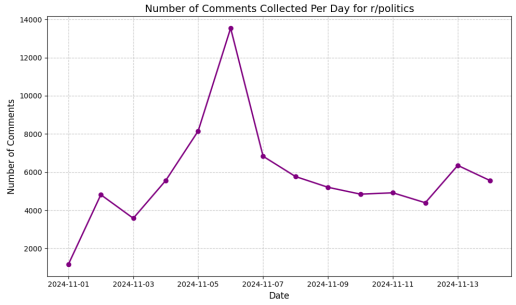

Figure 20: Comments per day in r/politics

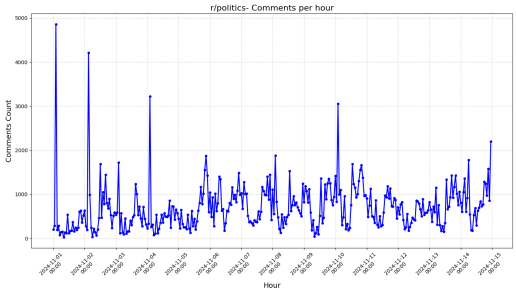

Figure 21: Comments per hr in r/politics

The Figure 22 shows the sentiment distribution of top 100 engaging posts from reddit

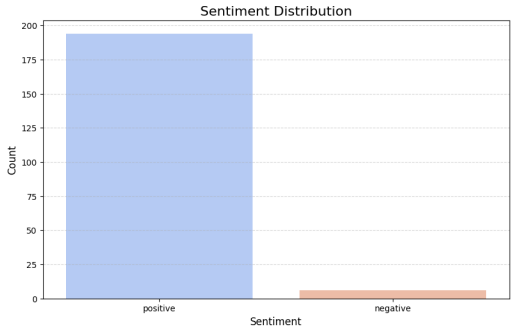

Figure 22: Sentiment distribution of trending posts

5 DISCUSSION

Progress so far includes the successful development and deployment of a data collection pipeline for both 4chan and Reddit. Using APIs and structured data storage techniques, a robust dataset has been curated. Preliminary sentiment analysis using state-of-the-art models has revealed distinct patterns in toxic and non-toxic content prevalence. Additionally, statistical comparisons between the platforms indicate differences in user engagement and behavior.

This work is significant as it addresses the growing need to understand toxicity and sentiment in online spaces, especially in anonymous platforms like 4chan, which are often under-researched. The insights gained from this study have implications for designing better content moderation systems, understanding community dynamics, and fostering healthier online environments.

5.1 Report summary

This project explores the collection and analysis of data from 4chan and Reddit, focusing on toxic content, engagement and sentiment trends. It leverages APIs and JSON endpoints to collect posts and applies sentiment analysis models to examine the differences between these platforms. The primary goals are to understand the volume and types of toxic content, compare sentiment trends, and evaluate the effectiveness of existing sentiment analysis tools on data from these platforms.

5.2 Implications of findings

- Preliminary analysis shows a significantly higher volume of user activity in both the platforms during the US presidential election and other major events in sports.
- Preliminary analysis shows a significantly higher volume of toxic content on 4chan compared to Reddit, likely due to differences in moderation practices.
- Sentiment trends on 4chan tend to skew towards negativity, while Reddit exhibits a wider range of sentiments, depending on the subreddit.
- The 4chan most engaging posts in 4chan are mostly being toxic, where as in case of reddit, it is not true. Reddit has more non-toxic engaging posts.

5.3 Limitations

- **Data Bias:** Data collection methods may not fully capture the diversity of posts on either platform as we focus only on parts of it. Moreover 4chan is not famous as reddit in some countries.
- **Computational Limitations:** The hugging face transformers, which is used to perform sentiment analysis takes a lot of time to analyse large amount of data.
- **Temporal Resolution:** The analysis is currently limited to a fixed time range and does not yet account for seasonal or event-based variations.
- **APIs:** While Reddit's API is structured, 4chan's JSON endpoints occasionally provide incomplete or inconsistent data.
- **System Failures:** Although the system is robust to well known failures, it is still a question that if it can handle unexpected failures and there are no data back stored for recovery.

5.4 Future work

- **Fine-tuning Models:** Custom training of sentiment and toxicity models using platform-specific data to improve accuracy.
- **Real-Time Analysis**: Developing a pipeline for real-time sentiment tracking to observe dynamic changes.
- **Context Analysis:** Incorporating contextual factors like post length, reply chains, and user engagement metrics for deeper insights.
- **Multilingual Support:** Extending the project to analyze posts in other languages for platforms with diverse user bases.
- **Multi-modal Support:** Extending the project to analyze posts in other formats like photos, videos and audios in platforms with diverse user bases.

5.5 Research questions

- How does toxicity in sports-related discussions compare across Reddit and 4chan?
- Is there a correlation between high engagement(e.g., number of comments or upvotes) and toxicity levels?
- What sentiment trends can be observed across sports discus sions, and how do they evolve over time

6 CONCLUSION

This project successfully demonstrated the feasibility of collecting and analyzing data from 4chan and Reddit to understand toxic content and sentiment trends across the two platforms. By leveraging API endpoints and sentiment analysis tools, we identified key differences in platform behavior, such as the higher prevalence of toxic content on 4chan compared to Reddit, and the broader sentiment diversity on Reddit due to better moderation and community structures.

Despite these successes, the project also revealed several limitations. The sentiment analysis models used were not fully optimized for the unique language and slang prevalent on 4chan, leading to potential inaccuracies. Additionally, data biases and platform-specific data inconsistencies highlight the need for more robust data collection strategies.

Looking ahead, there are several promising directions for extending this work. Developing custom sentiment analysis models tailored to the linguistic nuances of platforms like 4chan and Reddit would enhance accuracy. Expanding the analysis to include real-time tracking, multilingual data, and cross-platform comparisons could provide deeper insights into online discourse trends.

Ultimately, this project underscores the importance of understanding toxic behavior and sentiment dynamics on online platforms. Such insights have the potential to inform platform moderation strategies, combat the spread of harmful content, and contribute to broader discussions about digital ethics and community management.

7 GITHUB REPOSITORY

- **Project Implementation:** https://github.com/2024-Fall-CS-415-515/project-2-implementation-sgm
- **Final commit hash of the above repo:** 67bbb0cd863905e805ac05f984dd6d0b12ab490e

REFERENCES

[1] **4chan API documentation:** *https://github.com/4chan/4chan-API.*
[2] **Official Reddit API documentation:** *https://www.reddit.com/dev/api/.*
[3] **Faktory Official Github Page:** *https://github.com/contribsys/faktory.*
[4] Davidson, T., Warmsley, D., Macy, M., Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language.
[5] Waseem, Z., Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.
[6] Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., Gilbert, E. (2017). You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. Hugging Face documentation and model libraries for sentiment and toxicity analysis.