From Data to Insights: Interactive Exploration of Social Media Research

Rugvedh Vaidya
Binghamton University
Binghamton, New York, United States
rvaidya@binghamton.edu

Rohith Vardhan Siliveri
Binghamton University
Binghamton, New York, United States
rsiliveri@binghamton.edu

Prathyush Sirimalle
Binghamton University
Binghamton, New York, United States
psirimalle@binghamton.edu

1 Abstract

This report explains our plan to use social media data, collected through a live data gathering system, to address a key research question. Social media platforms offer a constant stream of discussions and trends, and by analyzing this data, we aim to discover patterns, identify connections, and gain meaningful insights relevant to our research. To make the analysis accessible and interactive, we built a web-based tool that allows users to explore the collected data through three types of analysis. Users can adjust settings and view results in real time, making it easier to understand trends and relationships in the data.

By combining powerful analysis methods with clear visualizations, this project aims to turn complex data into useful insights. In addition to answering our current research question, this application will serve as a reusable platform for future social media data analysis, enabling continued exploration and discovery.

Keywords: Sentiment Analysis, Keyword Frequency, Event Reaction, Social Media, Data Collection, Reddit, 4chan, Analysis.

2 Introduction

Social media has become a central part of modern communication, offering a huge volume of real-time data on public opinions, behaviors, and trends. Platforms like Reddit and 4chan provide a window into diverse perspectives, making them valuable resources for exploring complex questions about societal and technological trends. In this project, we aim to leverage the data collected through our live data collection system to address one of the research questions identified in our previous work. This effort builds on the foundation established in Project 2, where we developed a pipeline to gather and preprocess data. Our goal is not only to analyze the data but to present the findings in a way that enables users to explore the data. To achieve this, we built an web based interactive tool that will allow users to explore three of the analyses we previously conducted. By offering dynamic functionality, such as the ability to adjust parameters and filter data, we made the insights more accessible and actionable for users. Through this project, we demonstrated the power of combining robust analysis with interactive visualization. The ability to dynamically adjust analysis parameters enables a deeper understanding of the data, encouraging users to discover patterns and insights that static reports might overlook. Ultimately, this report is about transforming raw social media data into meaningful narratives that can inform decisions and spark further research.

3 Related Work

We build upon research investigating extremist online communities, such as the work by Papasavva et al. (2020) [04]that examined the Politically Incorrect (/pol/) board on 4chan, and the work by Balcı et al. (2023) [02]that focused on left-wing extremists on Reddit. These studies highlight the importance of understanding the unique characteristics and dynamics of such communities and their potential impact on online discourse and offline events. For example, Papasavva et al. (2020) [04]found that the /pol/ board is characterized by a high degree of hate speech and is influential in spreading disinformation and coordinating harassment campaigns. Similarly, Balcı et al. (2023) [02]showed that left-wing extremists on Reddit also exhibit some of the same worrying behaviors as right-wing extremists, such as relatively high toxicity and an organized response to deplatforming events.

Our research also draws upon studies that focus on characterizing online discussions and interactions, such as the work by Hamilton et al. (2014) [05]that investigated the emergence of communities on Twitch, and the work by Blackburn et al. (2012) [01]that examined the social dynamics of online gaming communities. These studies provide insights into the factors that shape online interactions, the formation of online communities, and the impact of social ties on individual behavior. Our research also incorporates insights from studies on detecting and measuring toxicity in online conversations, such as the work by Kwak et al. (2015) [03]that explored cyberbullying and toxic behavior in online games, and the work by Pavlopoulos et al. (2020) [06]that investigated the role of context in toxicity detection. These studies provide insights into the linguistic and behavioral patterns associated with toxicity, the factors that contribute to toxic behavior, and the challenges of detecting and mitigating toxicity in online environments.

Our sentiment analysis component draws upon research that has explored sentiment in various online communities. Papasavva et al. (2020)[1] analyzed sentiment in 4chan's Politically Incorrect board, using toxicity levels as a proxy for

sentiment and highlighting the challenges of sentiment analysis in toxic online environments. These studies inform our approach to sentiment analysis by highlighting the importance of considering the unique characteristics of different online communities.

4 Implementation:

4.1 Data Collection:

Data from Reddit and 4chan is collected and stored in a TimescaleDB database to support analysis and trend identification. For Reddit, posts and comments are gathered using Reddit's API, capturing key details such as timestamps, upvote counts, subreddit names, post IDs, and content. This data is stored in JSONB format, enabling efficient searching and analysis of trends across multiple data points within a structured and moderated environment. Similarly, for 4chan, each post includes essential details like the board name, thread ID, post number, timestamp, and content, all of which are stored in JSONB format. This approach facilitates the analysis of unstructured data and supports detailed time-based and sentiment analysis for both platforms.

4.2 Modern Hate Speech API

In our data collection system, we utilized the Modern Hate Speech API to measure the toxicity of user-generated content. After preprocessing, the cleaned text was sent to the Modern Hate Speech API via HTTP requests. The API provided a classification for each text as either normal (non-toxic) or flag (toxic), along with a confidence score indicating the certainty of the classification. For Reddit, toxicity analysis was performed on the posts and comments tables. The relevant user text in these tables resides in the title attribute for posts and the body attribute for comments. Similarly, in the 4chan database, toxicity analysis was conducted on the posts table. Here, the thread_number is analogous to posts or submissions on Reddit, while the post_number represents comments.

4.3 Data Analysis:

We aim to take the collected data from the database and use it for analysis.

4.3.1 Sentiment Analysis :

- In this analysis, after the data is collected and toxicity is measured using Hate Speech API. The sentiment of the data is classified using the two parameters in the data, toxicity and confidence score. I classified the data into 3 categories positive , negative and neutral. The classification is done by dividing the data by the toxicity variable. If the toxicity is flag then the sentiment is negative. If the toxicity is normal and confidence score is above 0.95 then that data is positive else it is neutral.

Platform	Positive	Neutral	Negative
4chan	392595	43419	83545
Reddit posts	9203	637	12
Reddit comments	25226	7911	809

Table 1. Sentimental Analysis Distribution of the data

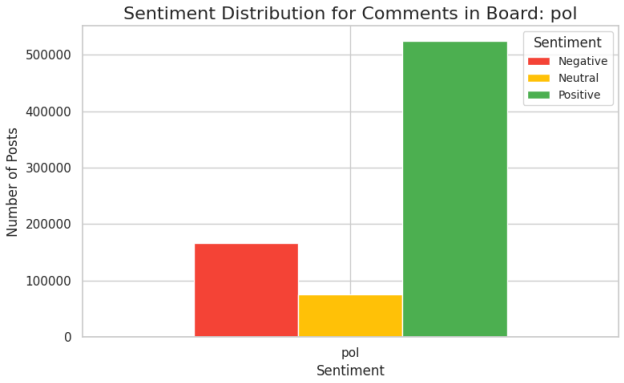

Figure 1. Sentiment distribution of /pol of 4chan

Figure 1 shows the distribution of the sentiment of /pol of 4chan data. The negative sentiment is higher compared to Reddit, since the platform has anonimity.

4.3.2 Keyword Frequency Analysis :

- In this analysis, we preprocessed the text content of the posts and comments by removing HTML tags, metadata, and stopwords. For example, the data extracted from 4chan required significant pre-processing to ensure the content was suitable for toxicity analysis.
  For instance, a sample record in the posts table looks like this:

```
{
    "id": 4824,
    "post_number": 102933491,
    "data": {
        "com": "<a href=\"#p102933441\" class=\"
            ↪ quotelink\">&gt;&gt;102933441</a><br>
            ↪ Is that real?",
        "resto": 102930087,
        "filename": "2024-10-22_00001_"
    },
    "thread_number": 102930087,
    "board": "g"
}
```

  Here, the com field contains user text but also includes extraneous elements such as HTML tags and metadata. We cleaned the data to retain only the readable content from com. Additionally, for records where com was empty, we checked the filename field for meaningful text. If the filename contained readable

information, it was used for analysis. However, if it consisted of random digits, the toxicity score for that thread submission was marked as NULL.

- Then a key word is taken and it is used to analyze by finding the number of occurrences of that keyword in the text field of the posts or comments data. Then a graph with the number of occurrences of that keyword against time is plotted. By looking at the graph we can say what people or the public are talking about during those days. After analysing we got the top 10 words we got that are related to the US election. The most used words were Trump and Harris. Similarly in the technology field we got the word 'apple' as the most used word. This was because Apple released new products.

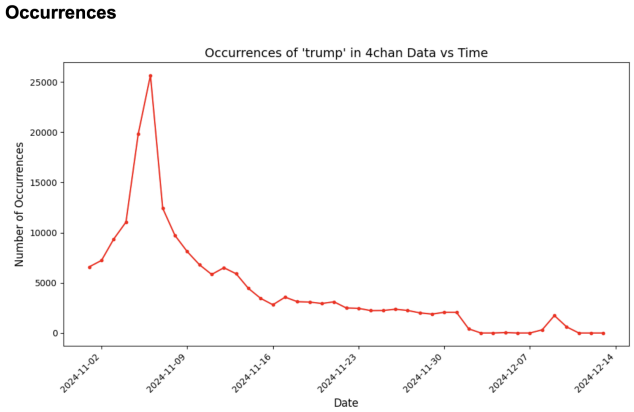

Figure 2. Occurence of Trump word on 4chan data

From Figure 2 we can see that the word Trump was discussed more during the election days.

4.3.3 Event Reaction Analysis: In this analysis we are trying to compare the reaction time of the platform to a particular event. Here we have used the word frequency as our main tool to get the occurrence of words related to a particular event. Since, there were only two major events(i.e. US election and Apple event) happened during the data collection period. Here we have made a list of words for each event and calculated the occurrences of those words in the text content of the data. Then we have scaled the occurrence of those words to 1 using min-max scaling, to make it easily understandable for event reaction analysis. Finally we have plotted the graph of the occurrences of those words against time. From the graph we can observe that there is a spike in the reddit plot before the 4chan has that spike, so we can say that the user on Reddit reacted to that event faster compared to the user on the 4chan..

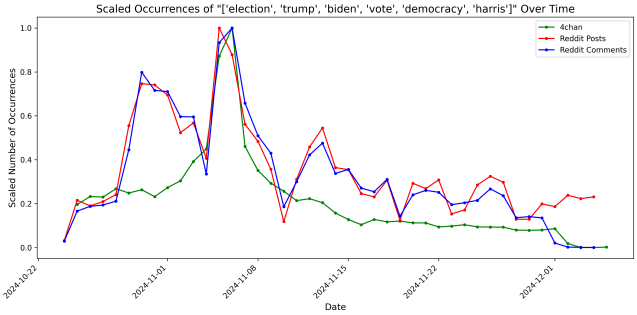

Figure 3. Reaction of users to US election on different platforms

The Figure 3 shows how the users of the platforms react to the event. Here, we can see that the reddit platform users reacted a bit early to the US elections event.

4.4 Live Web Analysis:

- The web application was built using Flask to provide an interactive and easy-to-use platform for analyzing social media data. It allows users to explore trends, sentiment, and keyword frequencies through a simple web interface. Some of the analysis, like keyword and event reaction, relies on the NLTK module for natural language processing and text analysis.
- To access the application, app.py script is run, which starts a local server. Once the server is running, the application can be viewed by opening http://127.0.0.1:5000/ in a web browser. The home page acts as a starting point, giving users access to key analysis tools like Sentiment Analysis, Keyword Frequency Analysis, and Event Reaction Analysis.
- The web application makes it easy for users to interact with the data in real time. Users can apply filters, set date ranges, and visualize the results instantly. This tool provides a clear way to explore social media trends, understand shifts in sentiment, and track keyword usage, all in one convenient interface.

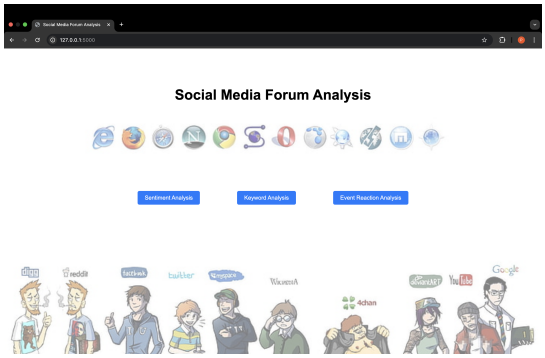

Figure 4. Web application home page

4.4.1 Sentimental Analysis:

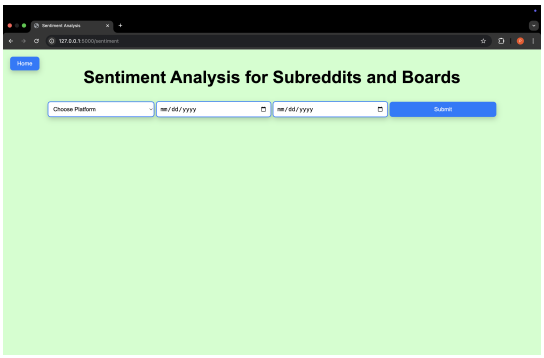

Figure 5. Sentiment Analysis page

- For the Sentiment Analysis, users begin by selecting a platform, either Reddit or 4chan. Based on the selected platform, they then choose a specific subreddit or board to analyze. After this, users specify a start date and end date to define the time period for the analysis.
- By clicking the Submit button, user input is taken, and bar plots for sentiment analysis of posts and comments from the selected platform will be displayed. The X-axis represents the sentiment categories, while the Y-axis shows the number of observed posts and comments corresponding to each sentiment.
- This process enables a focused examination of user sentiment in posts and comments within the chosen timeframe, providing insights into emotional trends and shifts in discussions.

Below are the images of how the plots will appear after clicking the Submit button.

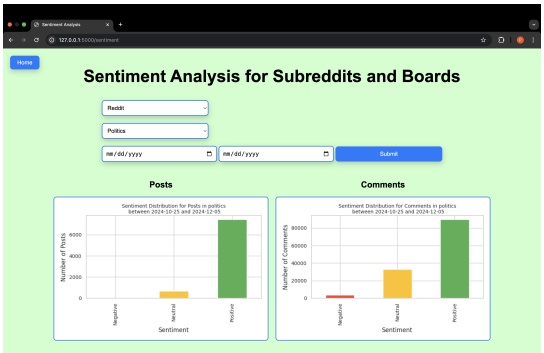

Figure 6. Sentiment Analysis for Reddit Politics posts and comments between October 25,2024 and December 5, 2024

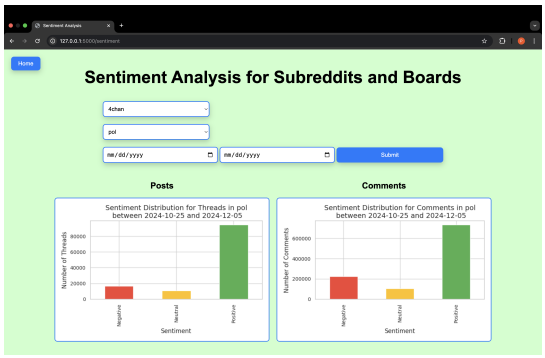

Figure 7. Sentiment Analysis for 4chan's pol posts and comments between October 25,2024 and December 5, 2024

The sentiment analysis for Reddit's Politics subreddit reveals a strong dominance of positive sentiment in both posts and comments, with only a small presence of negative and neutral sentiments. In contrast, on 4chan's pol board, while positive sentiment is still the most prominent, there is a noticeably higher share of negative posts and comments compared to Reddit. This suggests a more polarized discussion environment on 4chan, reflecting the differences in tone and engagement between the two platforms.

4.4.2 Keyword Frequency Analysis:

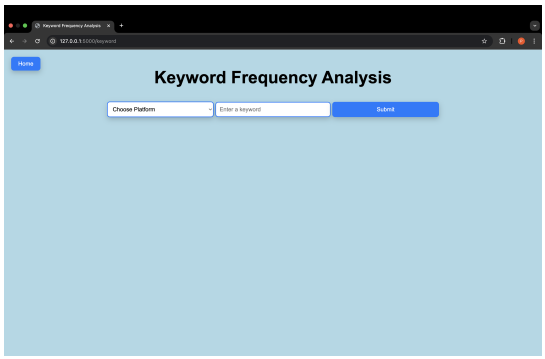

Figure 8. Keyword Frequency Analysis page

- The Keyword Frequency Analysis starts with the user selecting a platform, either Reddit or 4chan, as the source of the analysis. After choosing the platform, the user enters a keyword to track its frequency over time.
- On clicking the Submit button, the user input is processed and the results are displayed as time series plots. For Reddit, the plots show the number of keyword occurrences in posts and comments over time, with the X-axis representing the date and the Y-axis representing the frequency. For 4chan, the plot illustrates the total number of keyword occurrences in the entire dataset, with the X-axis showing the date and the Y-axis indicating the total count.

- This process allows for a focused examination of how often the chosen keyword appears in posts and discussions, providing valuable insights into its relevance and impact on user conversations.

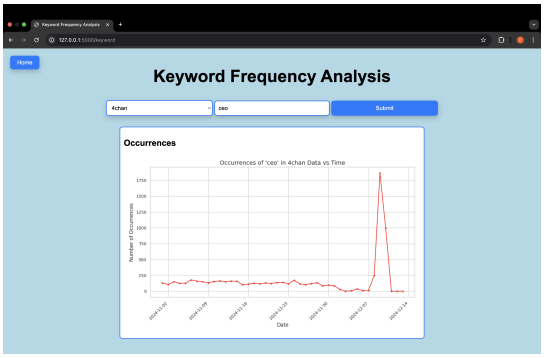

Figure 9. Keyword Frequency Analysis Plot for keyword 'ceo'

In the above image, we can see the occurrence of keyword ceo showing a huge spike after 7 December, 2024. This is possibly due to a recent incident involving the UHC CEO.

4.4.3 Event Reaction Analysis:

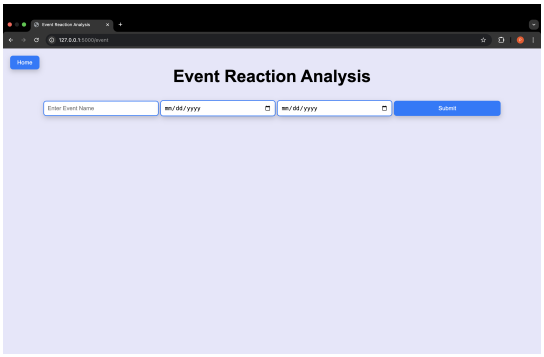

Figure 10. Event reaction analysis page

- The Event Reaction Analysis begins with the user entering an event name to define the focus of the analysis. After specifying the event, the user selects a start date and end date to set the time range for the analysis.
- Upon clicking the Submit button, the user input is processed to generate a time series plot. The plot displays the Date on the X-axis and the scaled number of keyword occurrences on the Y-axis, providing a clear visualization of how the event-related activity changes over the selected time period.
- This allows for a detailed examination of the event's impact and the associated keyword's activity within the selected period, offering insights into how discussions around the event evolve over time.

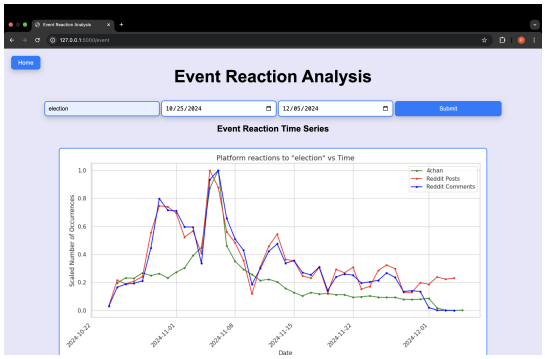

Figure 11. Event Reaction analysis plot for 'election'

In the above image, For the event 'election,' we can observe interaction trends related to the keyword among users on Reddit and 4chan.

5 Research Questions

After collecting the data continuously and performing some meaningful analysis for the data we collected, we raise three research questions.

RQ1: How does the level of toxicity vary between the user base of Reddit and 4chan forums?

- Our analysis, leveraging the Modern Hate Speech API, has illuminated a distinct difference in toxicity levels between Reddit and 4chan by looking at the Figure-6 and Figure-7. **4chan exhibits a higher level of negative sentiment compared to Reddit.** This difference is likely due to several factors, including 4chan's inherent anonymity and less strict moderation policies. Anonymity can often lead to a sense of disinhibition, emboldening users to express themselves more freely, sometimes resulting in a higher prevalence of toxic language.
- Additionally, the moderation policies on the two platforms play a crucial role in shaping the online discourse. Reddit, with its more active moderation, tends to curb the spread of toxic content more effectively. 4chan, on the other hand, is known for its laissez-faire approach to moderation, which can contribute to a more unfiltered and potentially toxic environment.

RQ2: What events are majorly discussed or inferred from the collected data?

- We analyzed the data to see what people were talking about. We found that major events were driving a lot of interesting discussions. Two major events stand out: the US election and Apple's product releases.
- The US election, particularly during the election days, generated significant buzz on both platforms, with

'Trump' and 'Harris' being the most used words as we can see it in the Figure-2. This highlights the intense public interest and engagement with the political landscape during this period.

- In the technology topic, discussions were dominated by Apple's new product releases, with 'apple' emerging as the most used word. This underscores the significant influence of tech giants like Apple in shaping online conversations and consumer interest.
- Interestingly, there was also a spike in discussions about the UHC CEO after December 7, 2024 as we can see in the Figure-9. While the specific incident prompting this surge in discussions remains unclear, it suggests that real-world events, particularly those involving prominent figures or organizations, can quickly become focal points of online discourse.

RQ3: How do users from Reddit and 4chan react to offline events and real-world technological or political events?

- Our investigation into user reactions to offline events has uncovered an interesting dynamic between Reddit and 4chan by looking at the Figure-3. Reddit users tend to react faster to real-world events compared to 4chan users. This difference in reaction time could be attributed to several factors, including the platforms' user demographics, moderation policies, and the overall speed of information dissemination on each platform.
- Reddit, with its diverse user base and active moderation, tends to be a hub for quick information sharing and discussions. 4chan, on the other hand, while also responsive to real-world events, might exhibit a slightly delayed reaction due to its less structured environment and less active moderation.
- Overall, our analysis has shed light on the distinct characteristics of Reddit and 4chan, highlighting their unique user dynamics, content trends, and responsiveness to real-world events.

6 Challenges Faced

- At one point, the virtual machine (VM) was unexpectedly rebooted, which caused all the running containers to stop. We faced some difficulties restarting the 4chan container, but after some troubleshooting, we were able to resolve the issue and get everything running smoothly again.
- We also ran into a 'full memory' problem on the VM, which locked us out and prevented us from logging in. After investigating, we found the root cause to be a massive PostgreSQL log file, which had grown to around 60 GB. To fix this, we truncated the log file without deleting it, freeing up enough memory to regain access to the system and continue working on the project.

7 Conclusion

This project highlights how social media data from platforms like Reddit and 4chan can be used to gain valuable insights into trends, sentiments, and user interactions. By storing the data in a structured format using TimescaleDB, we enable efficient and scalable analysis. The user-friendly web interface built with Flask allows users to explore Sentiment Analysis, Keyword Frequency Analysis, and Event Reaction Analysis in an interactive way. With real-time visualizations, the platform makes it easy to identify patterns and understand complex relationships within the data.

By combining advanced text analysis techniques, like sentiment scoring with NLTK, and clear time-series visualizations, this project bridges the gap between raw data and actionable insights. It not only answers key research questions about online discussions and trends but also provides a reusable framework for future social media data analysis. This system offers an accessible and intuitive tool for researchers and decision-makers to explore evolving conversations and make informed decisions based on meaningful data.

8 References

- https://www.reddit.com/dev/api/
- 4chan crawler discussed in the class
- ModernHateSpeech API
- Pandas
- Sci-kit Learn
- NLTK
- Flask
- Blackburn et al. - 2012 - Branded with a scarlet C Cheaters in a gaming social Network [01]
- Balci et al. - 2023 - Beyond Fish and Bicycles Exploring the Varieties of Online Women's Ideological Spaces[02]
- Kwak et al. - 2015 - Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games[03]
- Papasavva et al. - 2020 - Raiders of the Lost Kek 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board[04]
- Hamilton et al. - 2014 Streaming on Twitch: Fostering Participatory Communities of Play within Live Mixed Media[05]
- Pavlopoulos et al. - 2020 - Toxicity Detection Does Context Really Matter[06]