Social Media Data Science Pipelines Project Report: Dataset Measurements and Analysis

Devang Jagdale
djagdale@binghamton.edu
Binghamton University
Binghamton, New York, USA

Tejas Hiremath
thiremath@binghamton.edu
Binghamton University
Binghamton, New York, USA

Chaitanya Jha
cjha@binghamton.edu
Binghamton University
Binghamton, New York, USA

ACM Reference Format:
Devang Jagdale, Tejas Hiremath, and Chaitanya Jha. 2024. Social Media Data Science Pipelines Project Report: Dataset Measurements and Analysis. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This project builds on prior work to address key research questions about social media activity and sentiment analysis. We analyze temporal patterns, engagement metrics, and sentiment shifts across Reddit and 4chan data to answer our proposed research questions.

2 RESEARCH QUESTIONS

- How do temporal patterns in social media activity correlate with engagement metrics like comments?
- How do shifts in sentiment metrics (e.g., happy, sad, angry, hopeful) correlate with key events and discussions in U.S. politics?
- How do tones and volumes of discussions differ between mentions of different political candidates or parties?

3 DATA AND METHODS

Data was sourced from Reddit and 4chan datasets and processed using Python libraries such as Pandas and Plotly. The interactive tool built with Flask and Dash allowed querying and visualizing key analyses. Key analyses included:

- Time-series analysis of activity and engagement.
- Sentiment analysis of user comments.
- Country-based analysis of comment distributions.

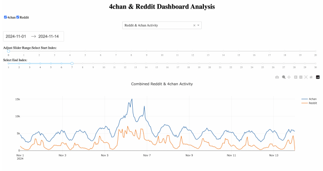
3.1 Tools and Frameworks

- Flask/Dash for the interactive dashboard.
- Plotly for visualizations.
- Scikit-learn for sentiment analysis.
- PostgreSQL for dataset management.

4 WEB BASED RESULTS AND DISCUSSION

4.1 Time-Series Analysis

Figure1 shows combined activity from Reddit and 4chan. The analysis highlights peak activity hours and engagement patterns. In this graph we have added two filters where first is for date and second is for dataset selection.



4chan & Reddit Dashboard Analysis

☑ 4chan ☑ Reddit

Reddit & 4chan Activity

2024-11-01 → 2024-11-14

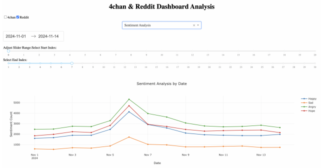
Adjust Slider Range:Select Start Index:

Select End Index:

Figure 1: Time-series analysis of Reddit and 4chan activity.

4.2 Sentiment Analysis

Sentiment analysis revealed trends in user emotions during significant political events. Figure2 illustrates the sentiment breakdown across dates. Filters used were date and sentiment that you want to see across timeseries.



4chan & Reddit Dashboard Analysis

☐ 4chan ☑ Reddit

Sentiment Analysis

2024-11-01 → 2024-11-14

Adjust Slider Range:Select Start Index:

Select End Index:

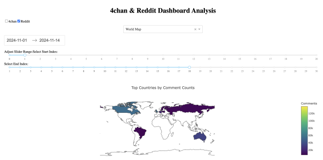
Figure 2: Sentiment analysis of Reddit and 4chan comments.

4.3 Country-Based Analysis

Figure3 displays the distribution of comments by country, highlighting regions with high engagement. Filter created was of range of top countries like if I want to see countries ranging from top m to n ranking only.

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

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



4chan & Reddit Dashboard Analysis

☐ 4chan ☑ Reddit

World Map

2024-11-01 → 2024-11-14

Adjust Slider Range:Select Start Index: 1

Select End Index: 18

Figure 3: Country-based comment distribution.

5 RESEARCH QUESTIONS ANSWERED

5.1 Research Question

5.1.1 How do shifts in sentiment metrics—such as happy, sad, angry, and hopeful—correlate with key events and discussions in U.S. politics? Emotional Spikes and Hierarchy November 6 (U.S. Presidential Election)

- **Angry** and **hopeful** emotions dominate:
  - **Angry sentiment** is at its peak, likely driven by contentious election outcomes, allegations of fraud, or dissatisfaction with the process. Anger here may signify frustration among those whose expectations were unmet or fears about future political directions.
  - **Hopeful sentiment**, equally prominent, indicates optimism among those who see the election outcome as aligning with their desires or values. This duality highlights the polarized nature of political discussions during elections.
  - **Happy sentiment**, though present, is relatively subdued compared to anger and hope. This suggests that celebratory reactions were overshadowed by more intense, conflicting emotions.
  - **Sad sentiment** is the lowest, indicating that disappointment or despair is less frequently expressed in comparison to anger or hope. This may reflect a tendency of users to externalize negative emotions through anger rather than introspective sadness.

November 13 (Key Position Announcements)

- **Angry** and **hopeful** emotions spike again, continuing the pattern from November 6.
  - **Angry sentiment** remains high, possibly driven by disapproval of announced individuals or policies. It suggests ongoing frustration and debates around the implications of these appointments.
  - **Hopeful sentiment**, equally prominent, reflects continued aspirations and expectations for positive changes or outcomes based on the new appointments.
  - **Happy** and **sad** sentiments remain relatively constant, highlighting that the discourse during this event was primarily focused on hope and anger.

5.1.2 How do temporal patterns in social media activity correlate with engagement metrics like comments? Emotional Hierarchy Differences Temporal patterns in social media activity, as demonstrated in the comparison between 4chan and Reddit, correlate with engagement metrics like comments by highlighting the influence of platform characteristics and event-driven activity on user interaction.

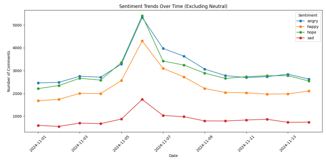


Figure 4: Sentiment Analysis

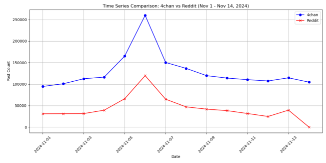


Figure 5: 4Chan vs Reddit comments against time

Key Insights from the Data: Higher Post Count on 4chan (Doubling Reddit's Activity):

From November 1 to November 14, 2024, 4chan exhibited consistently higher post counts compared to Reddit, likely driven by its unmoderated and anonymous nature. This environment facilitates uninhibited discussions, particularly around politically charged topics like the 2024 U.S. elections.

Correlation with Engagement Metrics:

On 4chan, the higher post count suggests more immediate and frequent user interactions, which likely translates to increased engagement in terms of comments. The unfiltered nature of discourse encourages rapid responses and ongoing debates. In contrast, Reddit's lower post count may correlate with fewer overall comments due to its structured moderation and community guidelines. However, the comments on Reddit posts might reflect deeper, more thoughtful engagement, given the platform's diverse user base and topic range.

Temporal Patterns Reflect Event-Driven Spikes:

The period between November 1 and November 14 encompasses key political events like the U.S. Presidential Election (November 6) and subsequent announcements of key positions (November 13). These events are likely to generate spikes in both post counts and engagement (e.g., comments), with 4chan showing more pronounced activity due to its reactive and real-time nature.

Platform-Specific Dynamics:

On 4chan, the immediacy and intensity of reactions to political events likely drive comment engagement in tandem with post spikes. This aligns with the platform's culture of rapid, emotionally charged discussions. Reddit, with its broader topic scope and moderation, might see more consistent but less volatile engagement metrics, as users focus on a variety of discussions beyond the political sphere.

5.1.3 How do the tones and volumes of discussions differ between mentions of different political candidates or parties? The tones and volumes of discussions differ between mentions of political candidates or parties based on geographic trends in engagement, as highlighted by the global distribution of comment activity.

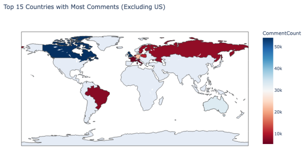


Figure 6: World map based on number of comments

Insights into Tone and Volume: Volume Differences by Geography:

The choropleth map reveals that countries like Canada contribute the highest number of comments outside the United States, while other countries show varying levels of engagement. High-comment countries likely have more frequent discussions about U.S. political candidates or parties, potentially reflecting heightened interest in or impact from U.S. policies on these regions. Tone Variation by Region:

Countries with higher comment activity, such as Canada, may exhibit more diverse tones—ranging from supportive to critical—depending on their political alignment or perceived implications of U.S. political events. In countries with fewer comments (e.g., those represented in darker brown), discussions might be less polarized or focused, leading to a narrower range of tones. Candidate or Party Mentions:

Discussions about specific candidates or parties may see differing tones depending on the international perspective. For example: A candidate perceived as advocating for stronger international ties might elicit more positive discussions in countries with high comment activity. Conversely, candidates associated with controversial policies could see more critical tones in comments from affected regions. Engagement Trends by Political Context:

Countries with high engagement levels may reflect broader awareness or emotional investment in U.S. politics, with discussions varying in tone based on the candidate's relevance to their local or global context. The geographical differentiation underscores that tone and volume are not uniform but influenced by regional interests, cultural factors, and the global implications of U.S. political dynamics.

6 CONCLUSION

This project successfully answered the research questions by implementing analyses that correlated activity and sentiment metrics with social and political contexts. The interactive dashboard facilitates further exploration of these patterns.

REFERENCES

(1) Social media data processing techniques: https://pandas.pydata.org/
(2) Plotly visualization library: https://plotly.com/
(3) Flask framework: https://flask.palletsprojects.com/