Tracking Political Sentiments: Analyzing Election-Driven Discussions on Reddit and 4chan

Asmita Singh
Binghamton University
Binghamton, NY, USA
asingh14@binghamton.edu

Chinmay Purandare
Binghamton University
Binghamton, NY, USA
cpurandare@binghamton.edu

Sanika Shinde
Binghamton University
Binghamton, NY, USA
sshinde2@binghamton.edu

Abstract

This project examines online discourse during the 2024 U.S. presidential election, focusing on toxic sentiment, engagement patterns, and thematic trends on Reddit and 4chan. By analyzing flagged comments, it reveals how political events shape public sentiment, amplify polarization, and drive engagement. A command-line tool (CLI) enables interactive querying and visualization of sentiment trends, toxic language, and recurring narratives. This study offers insights into the impact of digital platforms on political discourse, providing valuable perspectives for researchers, policymakers, and platform moderators to navigate the challenges of online communication during high-stakes event.

Keywords

Reddit API, Crawler, 4Chan API, PostgreSQL, Data Collection, Data Pipeline, Data Measurement and Analysis, CLI, Named Entity Recognition, Latent Dirichlet Allocation, Sentiment Analysis, Plotting, Visualization, Election Trends, Impact of Social Media

1 Introduction

The 2024 U.S. presidential election marked a pivotal moment in how online platforms shaped political discourse. Reddit and 4chan emerged as key arenas for public opinion, with their contrasting moderation styles offering unique insights into political discussions. Reddit's structured debates contrasted sharply with 4chan's unmoderated, polarizing exchanges.

This project explores the evolution of toxic interactions during politically charged moments, focusing on flagged toxic comments. It identifies key entities and topics associated with toxicity, analyzes how sentiment (emotional tone and subjectivity) influences discussions, and examines recurring themes in high-engagement, high-toxicity conversations.

Using techniques such as Named Entity Recognition (NER), sentiment analysis, and LDA topic modeling, the study addresses questions about online toxicity, sentiment dynamics, and dominant narratives. This analysis highlights the broader implications of digital communication in shaping public sentiment and offers data-driven insights for managing challenges in online discourse during critical events.

2 Methodology

2.1 Data Collection

Data was collected using custom-built crawlers for Reddit (targeting subreddits like r/politics and r/conservative) and 4chan's /pol/ board, capturing posts and comments during key election periods. Data was stored in a PostgreSQL database and classified using the Moderate Hate Speech API into normal and flagged (toxic) categories. This enabled the analysis of toxicity patterns, sentiment variations, and thematic trends across these platforms.

2.2 Tools and Frameworks

- Programming Languages: Python
- Libraries:
  - spaCy for NER
  - TextBlob for sentiment analysis
  - sklearn for LDA topic modeling
  - matplotlib and seaborn for visualization
- Data Storage: PostgreSQL for storing and querying large datasets.

3 Analyses

- **For NER and LDA, we have one parameter that can be varied and it is the –date-range parameter which is in YYYYMMDD-YYYYMMDD format**
- **For sentiment analysis, we have two parameters that can be varied and they are:**
  Reddit: 1) –upvote-threshold and 2) –date-range
  4chan: 1) –country-name and 2) –date-range

3.1 Named Entity Recognition (NER)[2] Analysis for Toxic Comments

Named Entity Recognition (NER) using spaCy was a critical analytical technique employed to identify and categorize significant entities from flagged toxic comments on Reddit and 4chan's /pol/ board. This method provided granular insights into the types of entities being targeted or mentioned in toxic discourse, addressing the research question:

What key entities and topics emerge from flagged toxic comments on different subreddits and 4chan's /pol/ board, and what do they reveal about online toxicity?

- Execution: The commands used for performing NER analysis offered flexibility in terms of data selection:
- Date Range: Specific
- Reddit Command:

```
python3 reddit_analyzer/
    reddit_toxicity_analyzer.py --analysis-
    type="NER" --date-range="
    20241127-20241128" --output
    reddit_analyzer/
    ner_date_range_20241127_20241128_results
    .html
```

NER command for Reddit specific date range

This command focuses on toxic comments within a defined date range (e.g., November 27–28, 2024).

- Platform-Specific Analysis:
  For 4chan:

```
python3 chan_analyzer/chan_toxicity_analyzer.
    py --analysis-type="NER" --date-range="
    ALL" --output chan_analyzer/
    ner_date_range_all_results.html
```

NER command for 4chan all date range

This command allowed platform-specific segmentation to compare toxicity trends and entity mentions between structured (Reddit) and unmoderated (4chan) ecosystems.

- Results:
  Identified Entity Types:
  - PERSON: Public figures like "Kamala Harris," "Donald Trump," and others were frequently mentioned, often in derogatory contexts.
  - ORGANIZATION: Political parties and media outlets (e.g., "Fox News," "Democrats") were recurring entities, reflecting ideological divides.
  - EVENT: Mentions of debates, rallies, and protests highlighted the temporal relevance of discussions.
- Visualization:
  Results were saved in HTML format, organized by entity type (e.g., PERSON, ORGANIZATION). These outputs allowed users to interactively explore data and identify key trends. For example: The PERSON category revealed frequent abusive terms targeting political figures. ORGANIZATION highlighted toxic narratives around news outlets and advocacy groups.
  The commands for 'date=ALL' for Reddit and specified date range for 4chan can be found in readme of Project 3 implementation.[1]

3.2 Latent Dirichlet Allocation (LDA) [3] Topic Modeling

Latent Dirichlet Allocation (LDA) is a machine learning technique designed to uncover hidden topics within large volumes of text data. By clustering co-occurring terms, LDA identifies dominant themes and narratives, making it especially effective for analyzing patterns in high-toxicity and high-engagement discussions. This analysis was applied to toxic comments from Reddit and 4chan, providing insights into recurring narratives and thematic differences across platforms which addresses the research question:

What dominant terms and underlying themes emerge from analyzing flagged toxic comments in high-toxicity, high-engagement discussions?

The command can analyze all available data or a specific date range, generating a detailed report with topics and their associated words.

- Date Range: All
- Reddit Command:

```
python3 reddit_analyzer/
    reddit_toxicity_analyzer.py --analysis-
    type="LDA" --date-range="ALL" --output
    reddit_analyzer/
    lda_date_range_all_topics.html
```

LDA command for Reddit all date range

Analyzes all available toxic Reddit comments, identifying recurring topics and their most frequent words.

- 4chan Command:

```
python3 chan_analyzer/chan_toxicity_analyzer.
    py --analysis-type="LDA" --date-range="
    20241209-20241210" --output
    chan_analyzer/
    lda_date_range_20241209_20241210_topics.
    html
```

LDA command for 4chan specific date range

Extracts themes from 4chan comments during December 9–10, 2024.

- Key Insights
  - Reddit: Recurring themes included political figures (e.g., "Trump"), national identity (e.g., "Americans"), and explicit expressions (e.g., "don," "fuck," "moron").The topics derived align with the platforms' known reputations, lending credibility to the findings. Themes suggest that toxic discourse on Reddit often centers around contentious political figures and debates tied to national identity, intertwined with personal insults and explicit language.
  - 4chan: Discussions prominently featured terms reflecting hate speech and derogatory expressions, such as "nigger," "jews," "faggot," "women," and "retarded." These terms illustrate the platform's unfiltered nature and focus on provocative content.
- Visualization: Word clouds of the most frequent words within each topic effectively summarized key narratives, helping to identify dominant patterns in the toxicity of user-generated content.

The analysis highlights recurring themes across platforms, emphasizing political polarization, hate speech, and identity-driven debates. These findings provide actionable insights for moderation and policy-making, paving the way for more nuanced future studies on the evolution of toxicity in online platforms.

The commands for specified date range for Reddit and 'date=ALL' for 4chan can be found in readme of Project 3 implementation.[1]

3.3 Sentiment Analysis

Sentiment analysis examines the emotional tone (polarity) and opinion levels (subjectivity) of toxic comments on Reddit and 4chan to understand online discourse dynamics during significant political events. Polarity ranges from negative to positive (-1 to 1), reflecting emotional sentiment, while subjectivity measures the degree of opinion versus factuality (0 to 1). Using TextBlob, comments are analyzed for these metrics, revealing how positive and negative sentiments drive engagement on Reddit and highlight the adversarial nature of 4chan's discussions.

In sentiment analysis we are addressing the research question:
What is the relationship between sentiment levels (polarity and subjectivity) and the content of online discussions, based on the analysis of comments?

- Date Range: All
  Reddit Command:

```
python3 reddit_analyzer/
    reddit_toxicity_analyzer.py --analysis-
    type="SNT" --date-range="ALL" --upvote-
    threshold=1000 --output reddit_analyzer/
    snt_date_range_all.txt
```

Sentiment Analysis for Reddit for all dates with upvote threshold

Performs sentiment analysis on all toxic comments with at least 1000 upvotes.
4chan command:

```
python3 chan_analyzer/chan_toxicity_analyzer.
    py --analysis-type="SNT" --date-range="
    ALL" --country-name="United␣States" --
    output chan_analyzer/snt_date_range_all.
    txt
```

Sentiment Analysis for 4chan for all dates with any country

Analyzes all 4chan toxic comments across US.
Date Range: Specified
Reddit command for date November 27-28, 2024 and for country US:

```
python3 reddit_analyzer/
    reddit_toxicity_analyzer.py --analysis-
    type="SNT" --date-range="
    20241127-20241128" --upvote-threshold
    =500 --output reddit_analyzer/
    snt_date_range_20241127_20241128.txt
```

Sentiment Analysis for Reddit for specific dates with upvote threshold

Conducts sentiment analysis on Reddit comments with at least 500 upvotes posted between November 27-28, 2024.
4chan Command for date December 9-10, 2024 and for country UK:

```
python3 chan_analyzer/chan_toxicity_analyzer.
    py --analysis-type="SNT" --date-range="
    20241209-20241210" --country-name="
    United␣Kingdom" --output chan_analyzer/
    snt_date_range_20241209_20241210.txt
```

Sentiment Analysis for 4chan for specific dates with any country

Analyzes toxic 4chan comments from this specific date range for United Kingdom.

- Key Insights
  - Reddit:
    - Polarity Trends: Positive comments often get high upvotes; negative comments drive debates.
    - Subjectivity: Opinion-driven content dominates politically focused subreddits.
  - 4chan
    - Polarity Trends: Negative sentiment dominates discussions on 4chan's /pol/ board, reflecting the toxic and adversarial nature of the discourse.
    - Subjectivity: Comments are a mixed with spikes in subjectivity during politically charged events.
- Visualization
  Polarity vs. Subjectivity: The scatter plots provide a visual representation of the relationship between emotional tone (polarity) and personal opinion (subjectivity). High-subjectivity comments with extreme polarity (positive or negative) are more frequent during politically relevant discussions, particularly during debates or elections. Colored in purple and yellow, indicating sentiment balance on Reddit. Colored in green and red, highlighting stronger polarity trends on 4chan.

Reddit tends to have more diverse sentiment levels, with both positive and negative discussions contributing to engagement metrics like upvotes and comments. Subjectivity is high, showing a strong reliance on opinion-driven discourse. 4chan exhibits an overwhelmingly negative sentiment polarity with lower overall subjectivity compared to Reddit, likely due to the platform's inherent culture and engagement patterns.

These insights suggest that sentiment and subjectivity levels play a critical role in shaping the nature and tone of discussions on both platforms, reflecting their unique community behaviors.

4 Presentation

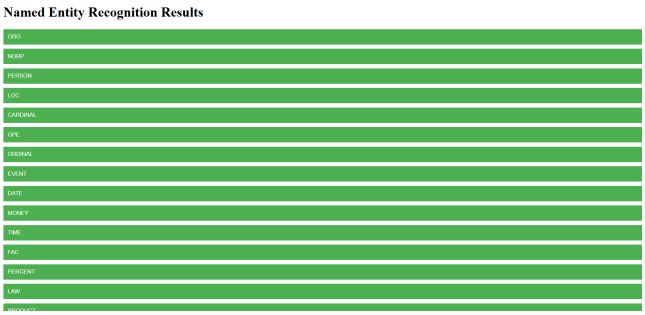

Figure 1: Named Entity Recognition (NER) Categories for Toxic Reddit Comments

In Figure 1, the image displays the results of Named Entity Recognition (NER) applied to toxic Reddit comments. The categories are structured in a dropdown format, where each green button represents an entity type identified by the NER process. Examples of categories include ORG (Organizations), PERSON (People), DATE, and more. The results highlight the diversity of named entities extracted from the dataset, categorized to provide insight into the context and subjects of toxic discussions.

In Figure 2, the image expands the "PERSON" category from the NER results applied to toxic Reddit comments. It provides a detailed list of individuals mentioned within the dataset, including names such as "Adam," "Alex Jones," "Clinton," and others. This

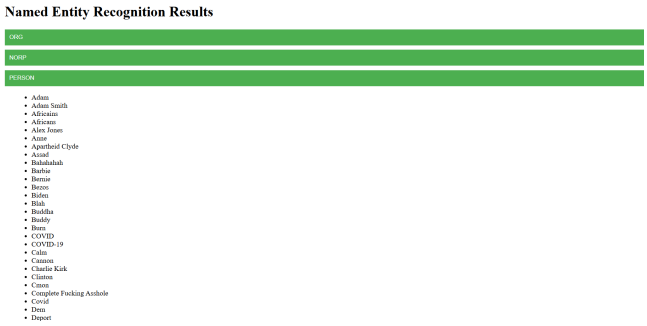

Figure 2: Expanded View of the 'PERSON' Category in NER Results

category captures personal mentions, showcasing how individuals are referenced in toxic discussions. The list reflects the diversity of personal entities identified in the dataset and can be used to analyze patterns or trends in discourse.

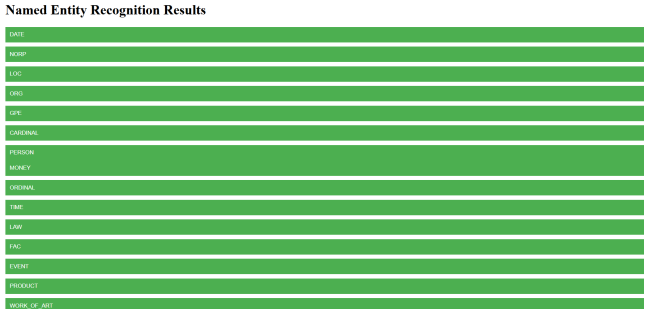

Figure 3: Named Entity Recognition (NER) Categories for Toxic 4chan Comments

In Figure 3, the image displays the results of Named Entity Recognition (NER) applied to toxic 4chan comments. The categories are organized in a dropdown format, where each green button represents an entity type identified by the NER process. Categories include DATE, NORP (Nationalities or Religious/Political Groups), PERSON, LOC (Locations), and others. These results showcase how entities are categorized to provide a comprehensive understanding of the topics and themes present in toxic discussions on 4chan.

In Figure 4, the image expands the "LOC" (Location) category from the NER results applied to toxic 4chan comments. It provides a detailed list of locations mentioned within the dataset, including names such as "Africa," "Asia," "Mars," and "Europe." This category captures geographic mentions, offering insight into how locations are referenced in toxic discussions on 4chan. The diversity of locations listed reflects the broad scope of geographic topics within the dataset.

In Figure 5, the scatter plot visualizes the polarity (sentiment ranging from -1 for negative to +1 for positive) and subjectivity (ranging from 0 for objective to 1 for subjective) of all comments collected from Reddit across all dates.

- Purple Points: Indicate comments with positive sentiment.

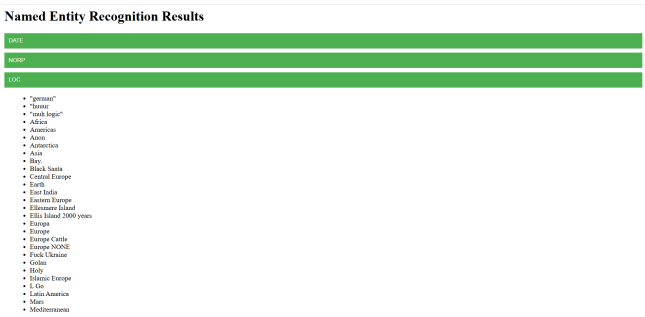

Figure 4: Expanded View of the 'LOC' Category in NER Results

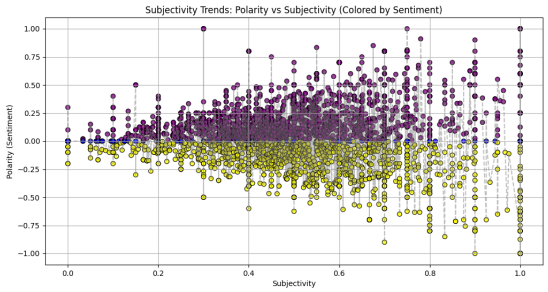

Figure 5: Polarity vs. Subjectivity Analysis of Reddit Comments Across All Dates

- Yellow Points: Indicate comments with negative sentiment.
- Blue Points: Represent comments with neutral sentiment.

The chart provides insights into the overall distribution of sentiment on Reddit, showing a mixture of positive, negative, and neutral comments across varying levels of subjectivity. The clustering around the neutral polarity line suggests a significant proportion of comments with mixed or balanced sentiments. This contrasts with the Chan data, offering a platform-specific perspective on emotional tone.

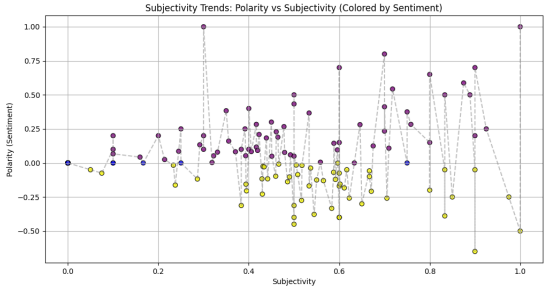

Figure 6: Polarity vs. Subjectivity Analysis of Reddit Comments for Two Selected Dates

In Figure 6, the scatter plot displays the relationship between polarity (sentiment ranging from -1 for negative to +1 for positive)

and subjectivity (ranging from 0 for objective to 1 for subjective) for Reddit comments collected over two specific dates.

- Purple Points: Represent comments with positive sentiment.
- Yellow Points: Represent comments with negative sentiment.
- Blue Points: Represent comments with neutral sentiment.

The focus on two dates offers a narrowed perspective, highlighting the sentiment dynamics within a smaller temporal scope. Compared to the broader dataset, this allows for a closer examination of sentiment patterns or specific events influencing the discussion.

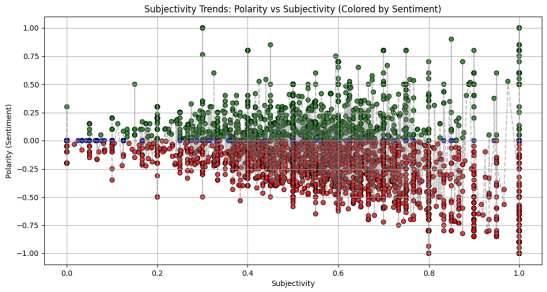

Figure 7: Polarity vs. Subjectivity Analysis of Chan Toxic Posts Over Time

In Figure 7, the scatter plot represents the relationship between polarity (sentiment ranging from -1 for negative to +1 for positive) and subjectivity (ranging from 0 for objective to 1 for subjective) for all comments collected from Chan across all dates.

- Green Points: Indicate comments with positive sentiment.
- Red Points: Indicate comments with negative sentiment.
- Blue Points: Represent comments with neutral sentiment.

The distribution shows a wide range of subjectivity across comments, with clusters of both positive and negative sentiments. The plot provides insights into the overall tone and emotional bias of comments, highlighting patterns such as a denser cluster of subjective negative comments compared to objective or neutral ones.

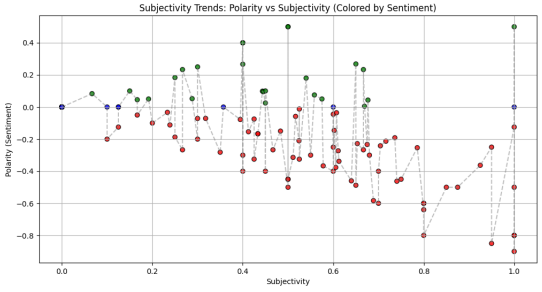

Figure 8: Polarity vs. Subjectivity Analysis of Chan Comments for Two Selected Dates

In Figure 8, the scatter plot shows the relationship between polarity (sentiment ranging from -1 for negative to +1 for positive) and subjectivity (ranging from 0 for objective to 1 for subjective) for comments collected from Chan, filtered for two specific dates.

- Green Points: Represent comments with positive sentiment.
- Red Points: Represent comments with negative sentiment.
- Blue Points: Represent comments with neutral sentiment.

The limited date range provides a more focused view of sentiment trends, making it easier to identify variations in tone and emotional bias during those specific periods. The dashed lines connect points to show potential temporal transitions or clusters within this subset.

Figure 9: Word Cloud of Latent Dirichlet Allocation (LDA) Topics from Reddit Comments Across All Dates

In Figure 9, the word cloud visualizes the most frequent words extracted from topics generated using Latent Dirichlet Allocation (LDA) on Reddit comments spanning all dates. The size of each word reflects its prominence and frequency across the five identified topics.

Figure 10: Word Cloud of Latent Dirichlet Allocation (LDA) Topics from 4chan Comments Across All Dates

In Figure 10, the word cloud visualizes the top words extracted from topics generated using Latent Dirichlet Allocation (LDA) on 4chan comments spanning all dates. The size of each word corresponds to its prominence or frequency across the five identified topics.

5 Conclusion

The analysis of toxic online discourse during the 2024 U.S. presidential election highlighted distinct patterns across platforms. Frequently targeted entities, such as political figures and ideologies, revealed how toxicity often centers around key topics of public interest. Recurring narratives on Reddit reflected structured debates and ideological divides, while 4chan's content was characterized by polarizing and adversarial themes. Sentiment analysis showed that positive sentiment on Reddit drove engagement and debate, whereas 4chan's discussions were dominated by negative and highly subjective exchanges. These insights reveal how toxicity, sentiment, and thematic trends influence engagement and reflect the unique cultures of these platforms.

6 Citations

References

[1] [n. d.]. Readme for Project 3 Implementation. ([n. d.]). https://github.com/2024-Fall-CS-415-515/project-3-implementation-web_crawlers/blob/main/README.md

[2] Emiliano De Cristofaro Gianluca Stringhini Jeremy Blackburn Antonis Papasavva, Savvas Zannettou. 2020. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)* (2020). https://arxiv.org/abs/2001.07487

[3] JEREMY BLACKBURN UTKUCAN BALCI, MICHAEL SIRIVIANOS. 2013. A Data-driven Understanding of Left-Wing Extremists on Social Media. (July 2013). https://arxiv.org/abs/2307.06981