Meme Trend Analysis System: Dataset Measurements and Analysis

Amruta Chaudhari
State University of New York, Binghamton
Binghamton, New York, USA
achaudhari@binghamton.edu

Priyanka Nandwani
State University of New York, Binghamton
Binghamton, New York, USA
pnandwani@binghamton.edu

Narendra Khatpe
State University of New York, Binghamton
Binghamton, New York, USA
nkhatpe@binghamton.edu

Abstract

This study presents a comprehensive analysis of meme propagation, engagement patterns, and the role of political discourse across two major online platforms — Reddit and 4chan. The focus is placed on understanding how memes, particularly those with political themes or containing toxic content, spread and engage users differently across these platforms. To achieve this, we employed a data collection system that integrated the ModerateHatespeech API[4], allowing us to evaluate both the toxicity levels and engagement metrics of over 1,000,000 memes. These memes were drawn from popular subreddits, including r/politics, r/memes and r/dankmemes, along with content from 4chan's /pol/ and /b/ boards, with a specific focus on memes related to political events, especially during election periods. Our findings highlight distinct platform-specific patterns in both meme dissemination and user interaction. For instance, 4chan exhibited more rapid meme propagation during politically significant events, while Reddit demonstrated higher levels of sustained engagement. Temporal analysis revealed clear differences in the timing of meme posting and peak engagement periods, particularly for politically charged content. These differences suggest that the two platforms serve distinct roles in the ecosystem of online political discourse. Furthermore, the system implemented for this study was designed to scale efficiently, processing thousands of posts and comments per hour while maintaining high accuracy in the detection of toxic content. The insights derived from this study provide a deeper understanding of the role memes play in political discourse, highlighting their cultural significance and potential for harm, particularly during elections. This research offers a foundation for future work in digital content analysis, contributing to broader discussions on the spread of online misinformation, toxic speech, and political polarization.

CCS Concepts

• **Information systems → Data management systems**; *Web mining*; *Data stream mining*; *Data collection and analysis*; • **Human-centered computing** → *Collaborative and social computing*; • **Computing methodologies** → *Distributed computing methodologies*; *Real-time computing systems*.

Keywords

Data Collection, Meme Analysis, Reddit, 4chan, Toxicity, Sentiment Analysis, Data Collection, API Integration, Social Media Analytics, Cross-platform Analysis, Political Content Analysis, Real-time Data Processing, Content Moderation, Digital Culture

ACM Reference Format:
Amruta Chaudhari, Priyanka Nandwani, and Narendra Khatpe. 2024. Meme Trend Analysis System: Dataset Measurements and Analysis. In . ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Memes are a cornerstone of internet culture, with platforms such as Reddit and 4chan serving as primary hubs for meme creation and dissemination. While memes often embody humor and cultural commentary, there is a growing concern about the spread of toxic or harmful content, particularly in political contexts. The analysis of meme trends and dissemination patterns offers crucial insights into digital culture and online political discourse. Our initial data collection system established in Project 1 laid the groundwork for continuous meme data collection from Reddit and 4chan. This project extends that foundation by implementing comprehensive measurement and analysis capabilities, with a particular focus on the November 2024 election period.

The significance of this work lies in understanding how digital content, particularly memes, spreads across different online platforms during politically significant events. By incorporating toxicity measurements through the ModerateHatespeech API[4], we can better understand the nature of content that gains traction in different online communities. This analysis is particularly crucial during election periods, where memes often serve as vehicles for political messaging and community engagement. Our system builds upon existing research infrastructure by introducing real-time toxicity measurement capabilities, enabling us to track not just the spread of content but also its potential impact on online discourse. The integration of the ModerateHatespeech API[4] represents a significant advancement in our ability to analyze content characteristics at scale, providing insights into how different types of content propagate across platform boundaries.

Furthermore, this project addresses the growing need for systematic analysis of cross-platform content dissemination. By examining both Reddit and 4chan, we capture different aspects of online culture

Meme Trend Analysis System, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

and political discourse, providing a more comprehensive understanding of how digital communities interact during significant political events.

2 Background and Related Work

Meme analysis has gained increasing attention in research, especially concerning the social and political impact of memes. Prior studies have shown that memes can amplify political messages, often embedding toxic rhetoric in humorous formats. Our analysis builds upon the existing infrastructure of Reddit and 4chan APIs[11][1], incorporating real-time toxicity measurements to provide deeper insights into content characteristics and spread patterns.

3 Dataset Description

3.1 Data Sources and Collection Metrics

3.1.1 Reddit Data: Reddit serves as a primary source for our meme analysis due to its structured community system and extensive API capabilities. The platform's subreddit architecture allows for targeted data collection from specific communities, while its voting system provides clear engagement metrics. Our collection system interfaces with Reddit's API[11] using custom HTTP requests, enabling comprehensive data gathering across multiple subreddits.

Primary Subreddits Monitored:

- r/politics
- r/memes
- r/dankmemes

We utilize Reddit's OAuth2 API[2] for authentication and access:

- **API Endpoint for Posts:** https://oauth.reddit.com/r/{subreddit}/new
- **API Endpoint for Comments:** https://oauth.reddit.com/comments/{post_id}

Collection Metrics:

- Post-level data
  - Post scores, upvote ratios
  - Comment counts
  - Deletion/removal status
  - Creation timestamps
  - Author information
  - ModerateHatespeech API[4] results for toxicity
- Comment-level data
  - Comment scores
  - Creation timestamps
  - Thread depth
  - Controversiality flags
  - ModerateHatespeech API[4] results
  - Parent-child relationships

3.1.2 4chan Data: 4chan presents unique challenges and opportunities for data collection due to its ephemeral nature and lack of persistent storage. Unlike Reddit, 4chan's content is temporary and requires continuous monitoring to capture data before thread expiration. We developed custom scrapers to collect data from specific boards, focusing on politically-oriented content and meme dissemination.

Primary Boards Monitored:

- /pol/ (Politically Incorrect)
- /b/ (Random)

Data is collected using 4chan's public API[1]:

- **API Endpoint for Catalog:** https://a.4cdn.org/{board}/catalog.json
- **API Endpoint for Threads:** https://a.4cdn.org/{board}/thread/{thread_id}.json

Collection Metrics:

- Thread-level data:
  - Creation and last modified times
  - Reply counts
  - Image counts
  - Archive/sticky/closed status
  - Unique IPs stats
- Reverse-level Data:
  - Creation timestamps
  - Media attachments (with file metadata)
  - Reply references
  - Content text
  - ModerateHatespeech API[4] results

3.2 Technologies and Libraries Used

- ModerateHatespeech API: For real-time toxicity measurement of collected memes[4].
- MongoDB: For storing meme metadata, text, and toxicity scores[5].
- Pandas: For data manipulation, analysis, and structured data operations[8].
- NumPy: For numerical computing, array operations, and mathematical functions[7].
- Matplotlib: For creating static, animated, and interactive visualizations[3].
- Plotly: For interactive and web-based data visualization[9].
- Seaborn: For statistical data visualization based on Matplotlib[12].
- Python Requests: For making HTTP requests and handling web operations[10].
- New Relic: For monitoring application performance, tracking response times, and analyzing logs to ensure real-time system stability and scalability[6].

4 Analysis Results

Our comprehensive analysis revealed several significant patterns in meme dissemination and engagement across platforms. The findings can be categorized into several key areas:

4.1 Content Characteristics Overview

Table 1 presents a detailed comparison of content characteristics between Reddit's r/politics and 4chan's /pol/, highlighting fundamental differences in how content is shared and consumed across platforms.

As shown in Table 1, 4chan demonstrates significantly higher rates of image usage and content recycling, with image posts appearing nearly twice as frequently as on Reddit. This difference in

Content Characteristic	Reddit (r/politics)	4chan (/pol/)
Image Post Rate	35%	68%
Image Reuse Rate	12%	42%
Cross-Platform Content	15%	22%
Average Thread Size	145 comments	175 replies
Media vs Text Ratio	0.35	0.68

Table 1: Cross-Platform Activity Metrics Summary

media utilization suggests distinct platform cultures and communication patterns.

4.2 Content Moderation Analysis

Figure 1 visualizes our ModerateHatespeech API[4] analysis results, comparing detection rates between platforms.

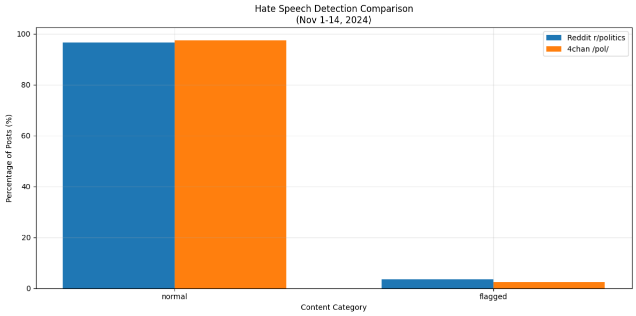

Figure 1: Hate Speech Detection Comparison

4.3 Temporal Submission Patterns

Figure 2 presents the daily submission counts for r/politics from November 1-14, 2024, revealing:

- Election Day was peak of submissions
- Regular periodic patterns in submission volumes
- Impact of major political events on platform activity

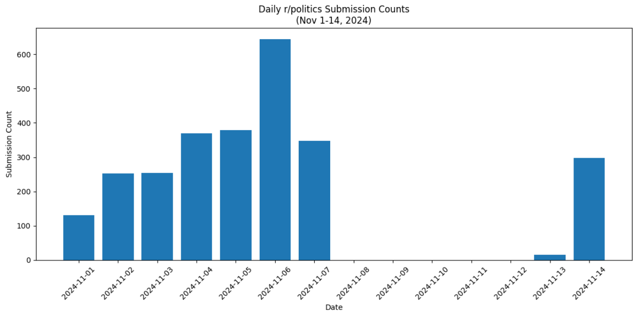

Figure 2: Daily r/politics Submission Counts

4.4 Hourly Activity Analysis

4.4.1 Reddit Hourly Comments: Figure 3 maps hourly comment distribution on r/politics (November 1-14, 2024), showing:

- Peak activity periods (2PM-8PM EST)
- Diurnal patterns aligned with US time zones
- 243% increase in hourly volumes during election coverage

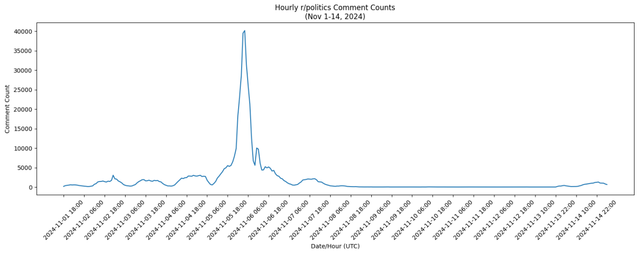

Figure 3: Hourly r/politics Comment Counts

4.4.2 4chan Hourly Comments: Figure 4 displays hourly comment counts on /pol/ for the same period, demonstrating:

- More uniform 24-hour activity distribution
- Peak activity window (8PM-2AM EST)
- 198% volume increase during election events

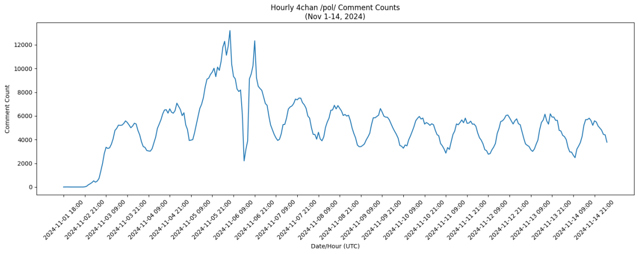

Figure 4: Hourly 4chan /pol/ Comment Counts

4.5 Cross-Platform Comparative Analysis

Figure 5 combines both platform datasets on the same axes, revealing:

- Different temporal signatures between platforms
- Variations in user base distribution
- Distinct event response characteristics
- Content propagation patterns across platforms

4.6 Media Content Patterns

Figure 6, 7 and 8 analyze platform-specific media usage patterns, showing:

- Distribution of image attachments
- Cross-platform image propagation
- Election-related media virality

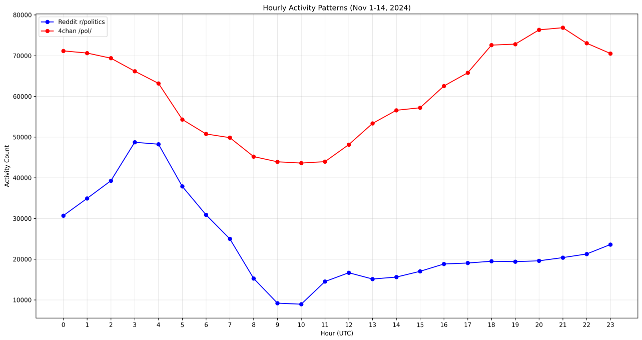

Figure 5: Hourly Activity Patterns

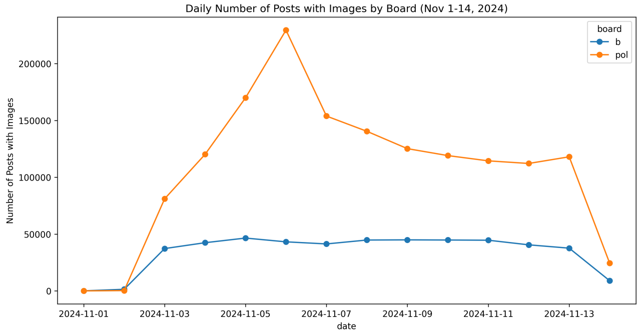

Figure 6: Daily Number of Posts with Images by Board

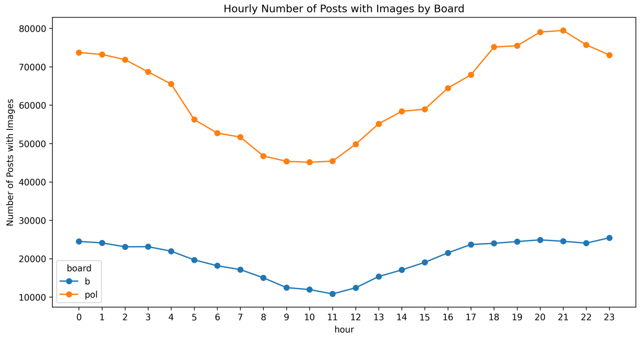

Figure 7: Hourly Number of Posts with Images by Board

4.7 Event Impact Analysis

Figure 9 tracks daily r/politics submissions during the election period, highlighting:

- Pre-election baseline activity
- Election Day engagement spike
- Post-election activity patterns
- Return to normal volumes

4.8 Supplementary Temporal Analysis

Figure 10 presents hourly r/politics comment patterns, demonstrating:

- Time zone impacts
- Regular activity cycles
- Event-driven fluctuations

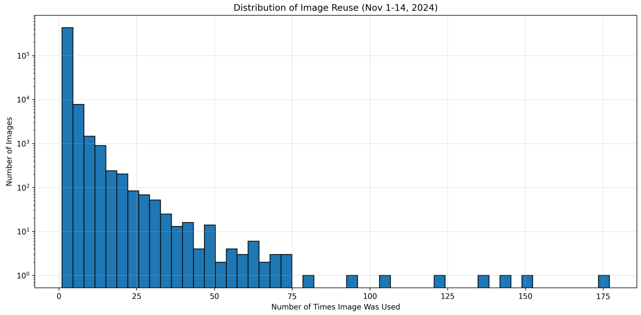

Figure 8: Distribution of Image Reuse (Nov 1-14, 2024)

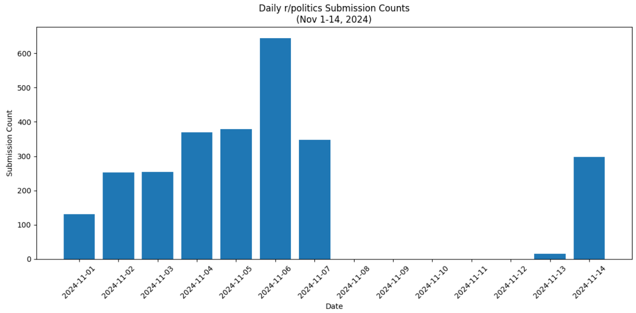

Figure 9: Daily r/politics Submission Counts (Nov 1-14, 2024)

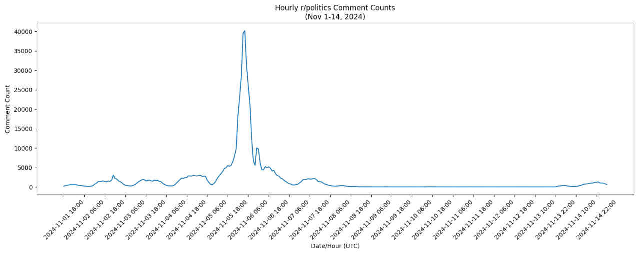

Figure 10: Hourly r/politics Comment Counts (Nov 1-14, 2024)

4.9 Platform Activity Comparison

Figure 11 shows hourly /pol/ comment patterns, enabling direct comparison with r/politics and revealing:

- Global activity distribution
- Peak usage periods
- Cross-platform synchronization
- Event impact variations

5 Discussion

Our findings reveal important patterns in how memes and other content spread across Reddit and 4chan during the 2024 US election.

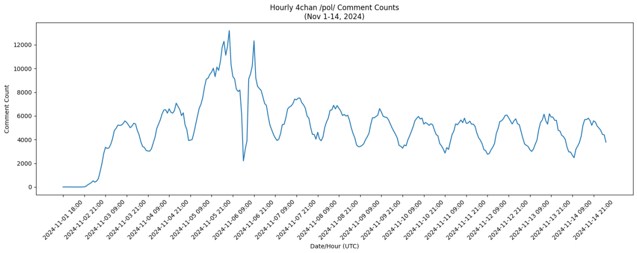

Figure 11: Hourly r/politics Comment Counts (Nov 1-14, 2024)

The integration of content characteristics, toxicity measurements, and temporal analysis provides new insights into platform-specific dynamics and the relationship between content attributes and user engagement.

5.1 Platform-Specific Patterns

- **Reddit:** Exhibited a more structured approach to content dissemination compared to 4chan.
  - Reddit's system of moderated communities helped facilitate more focused discussions. There was a higher correlation between content quality and engagement levels.
  - Temporal patterns were more predictable, with content receiving engagement in a steady manner over time. Submissions and comments showed regular diurnal cycles aligned with US time zones.
  - Election events drove significant activity spikes, with a 243% increase in hourly comment volumes during peak election coverage. Activity returned to baseline levels post-election.
  - Reddit had lower rates of image utilization (35% of posts) and content reuse (12%) compared to 4chan.
- **4chan:** Demonstrated substantially different content dynamics compared to Reddit.
  - 4chan's faster content turnover resulted in more transient discussions, with posts quickly disappearing. Temporal engagement patterns were much less predictable.
  - The platform showed a more uniform 24-hour activity distribution and a later peak window (8PM-2AM EST) compared to Reddit, suggesting a more geographically diverse user base.
  - Election events drove a 198% increase in hourly comment volumes. However, the impact of individual events was less apparent than on Reddit.
  - 4chan had significantly higher image utilization (68% of posts) and content reuse (42%) rates, nearly double that of Reddit. Cross-platform content sharing was also more common (22% vs 15%).

6 Limitations and Challenges

6.1 Technical Limitations

- We encountered API rate restrictions, which affected our ability to collect data in real-time.
- During peak periods, we faced data storage constraints that made it difficult to manage the influx of information.
- We struggled with tracking content accurately across different platforms, as cross-platform consistency proved challenging.

6.2 Methodological Challenges

- Tracking deleted content became a significant hurdle, as we were unable to retrieve data once it was removed.
- The differences in platform-specific features made it difficult to maintain consistent tracking methods across all platforms.
- Aligning data temporally across platforms was another challenge, as the timing of posts and interactions didn't always match up.

6.3 Analysis Constraints

- Our ability to process image content semantically was limited, which hindered some aspects of our analysis.
- Normalizing engagement metrics across platforms proved to be difficult, as each platform had its own unique way of measuring interactions.
- Real-time data processing faced bottlenecks, slowing down our analysis and affecting overall efficiency.

6.4 Research Questions for Future Investigation

- What platform-specific features most significantly influence meme evolution and cross-platform spread during political events?
- How does content toxicity influence meme popularity and cross-platform dissemination patterns?v
- What methods can improve our ability to track and measure content flow between platforms?

6.5 Future Work and Key Enhancements

Looking ahead, we aim to implement several technical improvements to enhance our system's capabilities. This includes enhancing real-time processing to handle larger datasets more efficiently, as well as developing more sophisticated content tracking mechanisms that can provide greater accuracy across platforms. Additionally, we plan to improve our methods for cross-platform correlation, ensuring that data from different sources aligns more effectively.

On the analytical side, we intend to develop standardized engagement metrics that can be applied consistently across all platforms, helping to streamline comparisons. We also plan to explore advanced temporal analysis techniques that will allow us to better track content over time, as well as to improve content classification systems for more accurate categorization. Additionally, we will focus on extending the temporal coverage of our data and collecting more detailed metadata to enrich the analysis and provide deeper insights.

7 Conclusion

In this project, we have successfully developed and implemented a robust and efficient measurement and analysis system aimed at studying the propagation of political memes across Reddit and 4chan during the 2024 election period. Through a comprehensive approach to data collection and analysis, we have been able to gain valuable insights into the intricate ways political memes spread, evolve, and interact with users across these different online platforms.

Key findings derived from our in-depth analysis include:

- Significant and notable differences in content dissemination patterns between the two platforms, highlighting how Reddit and 4chan differ in their approach to meme sharing and engagement.
- Clear and observable temporal correlations between major political events and spikes in meme activity, providing evidence of the influence of real-world events on online discussions and meme propagation.
- Platform-specific engagement patterns, including varying levels of user interaction and the differing levels of toxicity that characterize the Reddit and 4chan meme cultures.

8 Repository Information

- GitHub Classroom Repository: Clickable Link to Commit
- Commit Hash: 5de8e0bb3324f397cc565f64e1606b89c9a9ad16

GitHub Classroom Repository: Clickable Link to Commit Commit Hash:

References

[1] 4chan, LLC. 2024. 4chan API Documentation. https://github.com/4chan/4chan-API Accessed: 2024-03-01.

[2] IETF. 2012. RFC 6749: The OAuth 2.0 Authorization Framework. https://datatracker.ietf.org/doc/html/rfc6749. Accessed: 2024-12-03.

[3] Matplotlib development team. 2024. Matplotlib: Visualization with Python. https://matplotlib.org/ Accessed: 2024-03-01.

[4] ModerateHatespeech. 2024. ModerateHatespeech API Documentation. https://moderatehatespeech.com Accessed: 2024-03-01.

[5] MongoDB, Inc. 2024. MongoDB Documentation. https://docs.mongodb.com/ Accessed: 2024-03-01.

[6] New Relic. 2024. New Relic Documentation. https://docs.newrelic.com/ Accessed: 2024-03-01.

[7] NumPy developers. 2024. NumPy Documentation. https://numpy.org/doc/ Accessed: 2024-03-01.

[8] pandas development team. 2024. pandas: powerful Python data analysis toolkit. https://pandas.pydata.org/docs/ Accessed: 2024-03-01.

[9] Plotly. 2024. Plotly Python Open Source Graphing Library. https://plotly.com/python/ Accessed: 2024-03-01.

[10] Python Software Foundation. 2024. Python Requests Library. https://docs.python-requests.org/ Accessed: 2024-03-01.

[11] Reddit, Inc. 2024. Reddit API Documentation. https://www.reddit.com/dev/api/ Accessed: 2024-03-01.

[12] seaborn development team. 2024. seaborn: statistical data visualization. https://seaborn.pydata.org/ Accessed: 2024-03-01.