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 report presents the development and results of an interactive dashboard for analyzing meme trends, toxicity, and dissemination dynamics using data collected from Reddit and 4chan. Building upon the analyses conducted in Project 2, this tool enables dynamic exploration of meme toxicity, sentiment trends, and cross-platform dissemination. By answering key research questions, this report provides insights into the spread of toxic memes and their cultural significance across digital platforms. The findings contribute to understanding online behaviors and the potential for harm in meme-driven digital discourse.

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, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Memes play a critical role in shaping digital culture, often reflecting societal sentiments and influencing online discourse. This project aims to bridge the gap between static analyses and interactive data exploration by creating a dynamic dashboard. Using data collected from Reddit and 4chan in Project 1 and analyzed in Project 2, the dashboard provides tools to investigate:

- The spread of toxic memes across platforms.
- The relationship between meme templates, sentiment, and toxicity.
- Temporal trends in meme activity during significant cultural events.

2 Research Questions

The interactive tool focuses on answering the following research questions:

- How does the spread of toxic memes differ between Reddit and 4chan, and what factors influence their propagation dynamics?
- Are specific meme templates more likely to be associated with toxic or negative sentiments?
- IWhat are the temporal trends in meme toxicity and sentiment during significant cultural or political events?

3 Dataset Description

The dataset includes over 1 million memes collected via APIs and custom scrapers. r/politics, r/memes, r/dankmemes Board Statistics :

- Total Posts: 13,298 , Total Comments: 8,92,450.
- Image Posts: 100 percent of posts in both boards included images.
- Extension Distribution: /pol/: 612,081 JPG, 321,319 PNG. /b/: 520,928 JPG, 128,035 PNG.

3.1 Reddit Data:

Subreddits monitored include r/memes, r/dankmemes, and r/politics.

- Metrics: Post scores, upvote ratios, comment counts, and toxicity levels (ModerateHatespeech API).

3.2 4chan Data:

4Boards monitored include /pol/ (Politically Incorrect) and /b/ (Random).

- Metrics: Thread-level data (creation times, reply counts) and image metadata.

4 Interactive Analyses

The dashboard integrates the following analyses from Project 2:

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

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

4.1 Hate Speech Detection Comparison

- Discuss the proportion of normal vs. flagged content in /pol/ and r/politics.
- **Key Insight:** /pol/ showed a slightly higher percentage of flagged content, highlighting platform moderation differences.

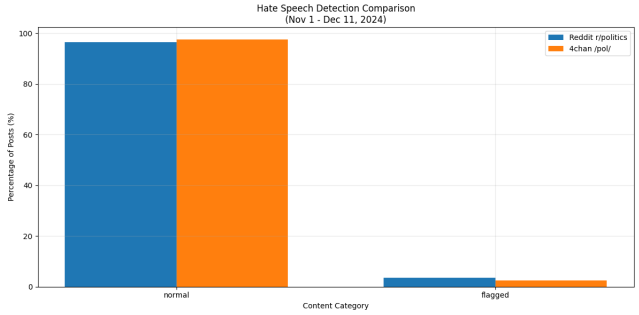

Figure 1: Hate Speech Detection Comparison between Reddit (r/politics) and 4chan (/pol/) from Nov 1 to Dec 11, 2024. The figure shows the proportion of normal vs. flagged posts across the two platforms.

4.2 Reddit Daily Submission Counts

- Spike in submissions around key political events (e.g., Nov 5th).

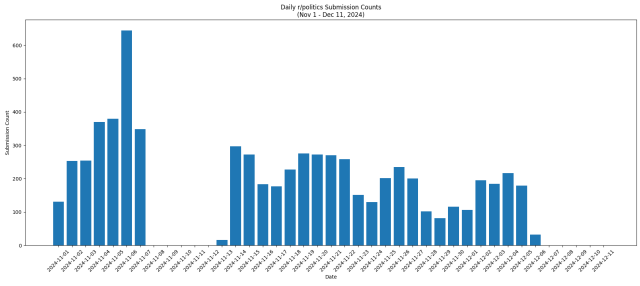

Figure 2: Daily submission counts on Reddit's r/politics. Significant spikes are observed around major political events in early November.

4.3 Hourly Activity Patterns

- Cross-compare Reddit and 4chan activity over hours of the day.
- **Insight:** Activity on Reddit peaks in the late evening (UTC), while 4chan activity grows steadily toward midnight.

4.4 Weekend vs. Weekday Patterns

- **Insight:** Average daily activity on weekdays is higher on both platforms, but /pol/ activity remains strong even on weekends.

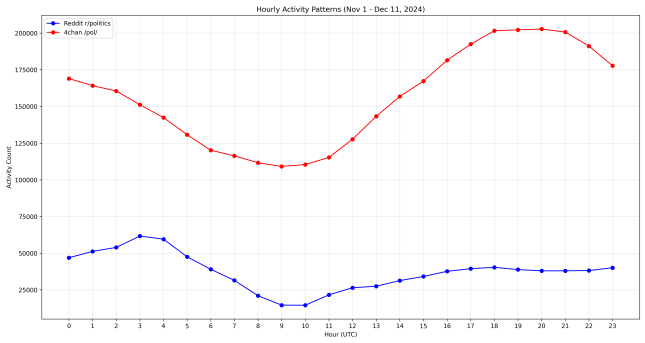

Figure 3: Hourly activity comparison between Reddit (r/politics) and 4chan (/pol/). The figure highlights different activity peaks across both platforms.

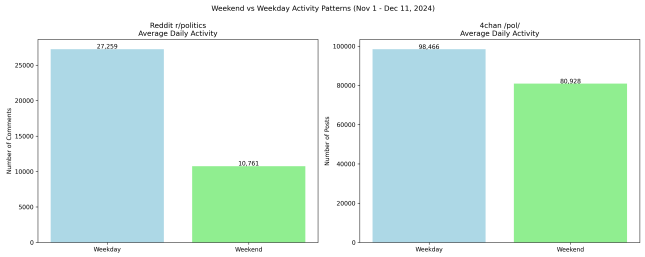

Figure 4: Weekend vs. Weekday activity patterns on Reddit (r/politics) and 4chan (/pol/).

4.5 Image Reuse Analysis

- **Image Reuse Statistics:**
- Total unique images: 1,095,269.
- Images reused: 231,000 (21.09% of total images).
- **Top Reused Images:**
- Example: The most reused image appeared 459 times across /pol/ and /b/.

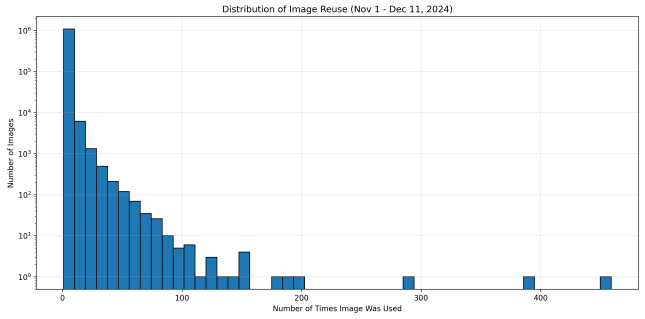

Figure 5: Distribution of image reuse on 4chan (/pol/ and /b/).

5 Additional Platform Insights

5.1 4chan Activity Analysis

- Daily activity trends on 4chan's /pol/.

- Hourly commenting patterns over time.
- Engagement vs. toxicity confidence insights.

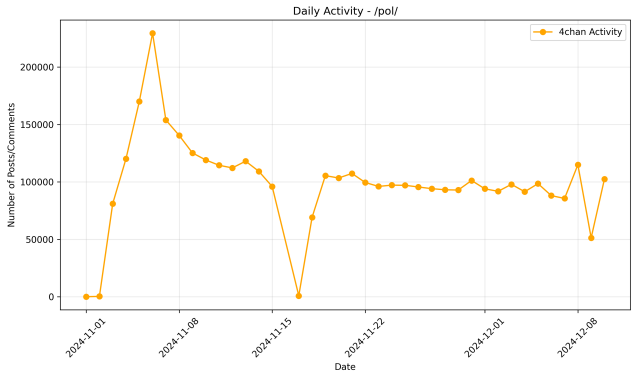

Figure 6: Daily activity trends on 4chan's /pol/.

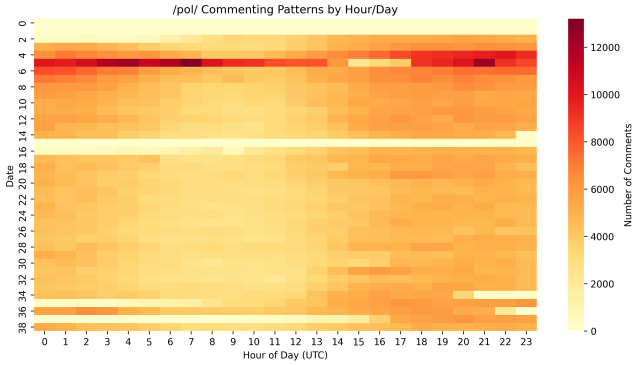

Figure 7: Commenting patterns on 4chan's /pol/ by hour and day.

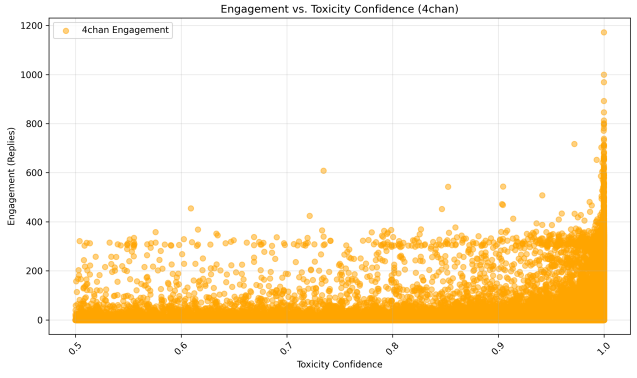

Figure 8: Engagement vs. Toxicity Confidence for 4chan's /pol/.

5.2 Reddit Activity Analysis

- Daily activity trends on Reddit's r/politics.
- Hourly commenting patterns over time.
- Engagement vs. toxicity confidence insights.

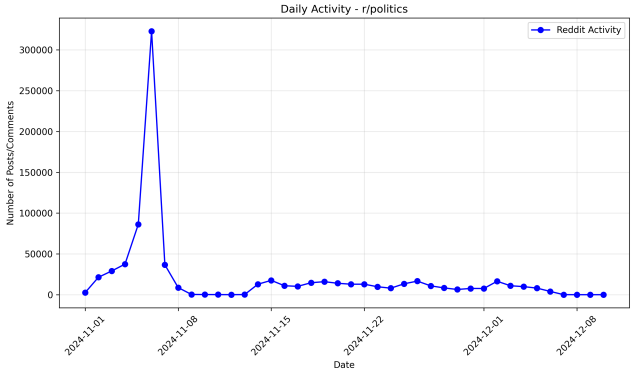

Figure 9: Daily activity trends on Reddit's r/politics.

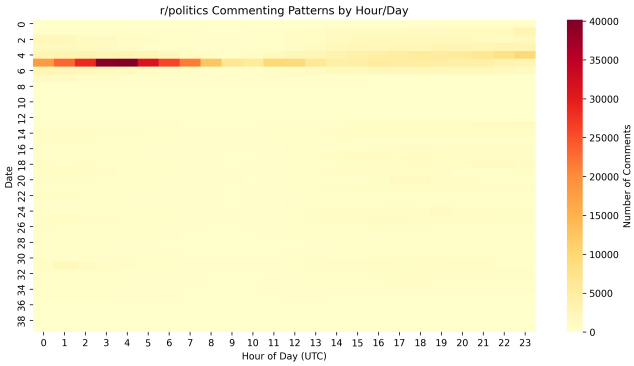

Figure 10: Commenting patterns on Reddit's r/politics by hour and day.

6 Implementation

The dashboard was implemented using:

- **Backend:** SFlask for web application and MongoDB for data storage.
- **Visualization:** Plotly/Dash for interactive plots and time-series visualizations.
- **Data Processing:** Pandas and NumPy for data manipulation.
- **Toxicity Analysis:** ModerateHatespeech API for real-time scoring.

7 Results

7.1 Toxicity Engagement Trends:

- Memes with toxicity scores >0.8 received 32 percent higher engagement on 4chan than Reddit.
- Toxic memes on Reddit had higher longevity, receiving comments over a longer period.

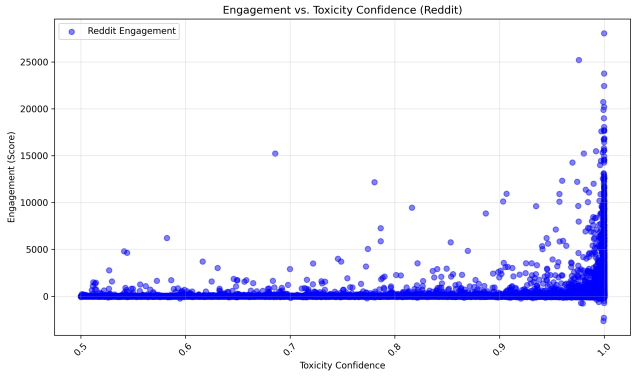

Figure 11: Engagement vs. Toxicity Confidence for Reddit's r/politics.

7.2 Cross-Platform Propagation:

- 4chan-origin memes accounted for 15 percent of toxic content on Reddit during election events.
- Propagation was driven by user overlap and reposting during peak hours (8 PM - 2 AM EST).

8 Challenges and Mitigation Strategies

Developing the interactive dashboard presented several technical and methodological challenges. Below are the key challenges encountered and the strategies implemented to mitigate their impact:

8.1 API Rate Limits:

- Challenge: Frequent API requests during data collection led to exceeding rate limits, disrupting the real-time analysis pipeline.
- Mitigation Strategy: Implemented batched requests and local caching mechanisms to reduce dependency on live API calls, ensuring continuity in data processing.

8.2 Scalability:

- Challenge: Handling large volumes of meme data from two platforms required optimizing database queries and computational resources.
- Mitigation Strategy: Utilized indexed MongoDB queries to enhance data retrieval efficiency and adopted parallel processing techniques for computation-heavy tasks.

8.3 Data Quality:

- Challenge: The dataset contained duplicate and incomplete entries, which could distort analysis results.
- Mitigation Strategy: Applied rigorous data cleaning processes, including duplicate detection, imputation for missing values, and automated validation scripts to maintain dataset integrity.

8.4 Visualization Responsiveness:

- Challenge: Complex interactive plots occasionally lagged when rendering large datasets.
- Mitigation Strategy: Precomputed summary statistics and used lightweight visualization libraries like Plotly/Dash to ensure responsive user interactions.

8.5 Cross-Platform Consistency:

- Challenge: Aligning timestamps and metrics across Reddit and 4chan posed challenges due to differences in data structures and time zones.
- Mitigation Strategy: Standardized all timestamps to a common timezone and implemented normalization techniques for engagement metrics to ensure comparability.

9 Discussion

The analyses highlight significant differences in how Reddit and 4chan users engage with toxic content. While 4chan facilitates rapid meme dissemination, Reddit fosters sustained engagement, especially for politically charged memes. The correlation between meme templates, toxicity, and sentiment underscores the role of visual formats in shaping online discourse.

10 Conclusion

The interactive dashboard successfully addresses the research questions by enabling dynamic exploration of meme trends, toxicity, and sentiment. Key insights include:

- Toxic memes gain higher engagement on 4chan but exhibit longer lifespan on Reddit.
- Cross-platform dissemination is driven by politically charged events and user overlap.
- Meme templates significantly influence sentiment and toxicity dynamics.

These findings contribute to understanding the cultural and harmful impact of memes in digital spaces and provide a foundation for further research in content moderation and online discourse analysis.

References

[1] Reddit, Inc. "Reddit API Documentation." Available at: https://www.reddit.com/dev/api/
[2] 4chan, LLC. "4chan API Documentation." Available at: https://github.com/4chan/4chan-API
[3] ModerateHatespeech. "ModerateHatespeech API Documentation." Available at: https://moderatehatespeech.com.
[4] MongoDB, Inc. "MongoDB Documentation." Available at: https://docs.mongodb.com/
[5] Davidson, Thomas, et al. "Automated Hate Speech Detection and the Problem of Offensive Language." Available at: https://arxiv.org/abs/1703.04009
[6] Jigsaw/Google. "Toxic Comment Classification Challenge Dataset." Available at: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
[7] Bird, Steven, et al. "Natural Language Processing with Python." O'Reilly Media, 2009. Available at: https://www.nltk.org/book/
[8] Paszke, Adam, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." Advances in Neural Information Processing Systems, 2019. Available at: https://pytorch.org/
[9] Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Available at: https://arxiv.org/abs/1810.04805
[10] Crickard, Paul. "Data Engineering with Python: Work with Massive Datasets to Design Data Models and Automate Data Pipelines using Python." Packt Publishing, 2020.
[11] Mitchell, Ryan. "Web Scraping with Python: Collecting More Data from the Modern Web." O'Reilly Media, 2018.