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 proposal outlines the design and implementation of a Dataset Measurements and Analysis System aimed at extracting insights from meme data collected from Reddit and 4chan. Building on the continuous data collection system from Project 1, this project will focus on measuring meme popularity, template reuse, cross-platform dissemination, sentiment, and toxicity. A third-party API (ModerateHatespeech) will be integrated to measure toxicity in real time, enabling analysis of harmful content within meme communities.

CCS Concepts

• **Information systems → Data management systems**; **Web mining**; **Data stream mining**; • **Data collection and analysis**; • **Human-centered computing → Collaborative and social computing**;

Keywords

Data Science, Meme Analysis, Reddit, 4chan, Toxicity, Sentiment Analysis, Data Collection, API Integration

ACM Reference Format:
Amruta Chaudhari, Priyanka Nandwani, and Narendra Khatpe. 2018. Meme Trend Analysis System: Dataset Measurements and Analysis. In . ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The analysis of meme trends and dissemination offers valuable insights into digital culture. Project 1 implemented a system to collect meme data continuously from Reddit and 4chan, focusing on meme popularity, template reuse, and sentiment. However, beyond simple data collection, it is crucial to conduct thorough analysis to extract meaningful patterns from the dataset.

This project expands on the initial system by incorporating toxicity analysis, aiming to study how harmful content spreads within meme communities. Integrating the ModerateHatespeech API, the system will analyze the toxicity of each meme and conduct a variety of experiments, including sentiment analysis and cross-platform dissemination tracking. The goal is to provide a comprehensive understanding of meme culture, both in terms of harmless trends and potentially toxic content.

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

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

2 Description of Data Sources

2.1 Reddit

- **Description:** Meme-related data from subreddits such as `/r/` memes, `/r/` dankmemes, and `/r/` PoliticalHumor.
- **API:** The Reddit API will be accessed via custom HTTP requests, as implemented in Project 1.
- **Data:** SPosts, comments, upvotes, image URLs, and metadata will be analyzed.

2.2 4chan

- **Description:** Data from meme-centric 4chan boards such as `/pol/` and `/b/`.
- **Scrapping:** Custom scrapers will be used to collect image URLs, thread content, and associated metadata.
- **Data:** Images, comments, and thread activity from targeted boards.

3 Data Analysis and Experimentation Plan

This emphasizes that the core work in this phase revolves around measuring, analyzing, and experimenting with the collected data.

3.1 Toxicity Measurement:

A core addition to the system will be the integration of the ModerateHatespeech API for measuring toxicity. The API will evaluate the toxicity of text and image captions collected from Reddit and 4chan. Given potential instability of the API, the system will be designed to handle outages, ensuring that toxicity measurements can be added as the data is ingested.

3.2 Meme Popularity and Spread

Experiments will focus on understanding meme popularity by analyzing upvotes, reposts, and comment counts from both Reddit and 4chan. We will track how memes evolve and spread across both platforms, identifying whether toxic content travels faster or gains more traction than neutral or positive content. Special attention will be given to election-related memes to determine if and how toxic content gains traction during this period.

3.3 Template and Sentiment Analysis

By leveraging OpenCV, meme templates will be recognized and tracked over time to identify the most reused templates. Sentiment analysis using Hugging Face's Transformers will categorize memes

into positive, negative, and neutral sentiments. Cross-analysis with toxicity measurements will show whether certain templates correlate with higher levels of toxic or negative sentiment.

3.4 Scaling and Real-Time Processing

As the dataset grows, ensuring scalability will be a critical part of the system. The analysis will be done both live and via data dumps, with real-time toxicity measurements integrated into the ingestion pipeline. The system will be scalable to accommodate the high volume of election-related data, and MongoDB will serve as the backend database for efficient data storage and querying.

4 Tools and Libraries

- **ModerateHatespeech API:** For real-time toxicity measurement of collected memes.
- **OpenCV:** For image recognition and meme template analysis.
- **Hugging Face Transformers:** For sentiment analysis and text processing.
- **MongoDB:** For storing meme metadata, text, and toxicity scores.
- **Pandas/NumPy:** For data processing and analysis.
- **Matplotlib/Plotly:** For visualizing meme trends and metrics.
- **Python Requests/BeautifulSoup:** For making HTTP requests and scraping 4chan data.
- **New Relic:** For monitoring application performance, tracking response times, and analyzing logs to ensure real-time system stability and scalability.

5 Experiments

The following experiments will be conducted:

5.1 Toxicity and Engagement:

- How does the toxicity level of a meme affect its popularity (engagement metrics like upvotes, comments) on Reddit and 4chan?
- Are specific meme templates more associated with toxic or negative sentiments, and do these templates gain higher engagement?

5.2 Cross-Platform Dissemination:

- What patterns emerge in the spread of toxic vs. non-toxic memes across Reddit and 4chan during the election period?
- Do memes originating from one platform (Reddit or 4chan) show higher toxicity when disseminated to the other platform?

5.3 Temporal and Event-Based Trends:

- How do sentiment and toxicity levels fluctuate around key election events, and is there a noticeable difference in response times between Reddit and 4chan?
- What are the temporal trends in meme engagement during the election week, and how do they correlate with major political events?

6 Conclusion

This project aims to develop a robust measurement and analysis system for the data collected in Project 1. By adding toxicity measurements and conducting experiments on meme popularity, sentiment, and cross-platform dynamics, the system will yield valuable insights into both the cultural and harmful aspects of memes. It will contribute insights into how political memes spread and how toxic content impacts digital engagement on Reddit and 4chan. These findings will set the foundation for future research in understanding the role of memes in digital society.

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] Hugging Face. “Transformers Documentation.” Available at: https://huggingface.co/docs/transformers.
[5] Bradski, G. “The OpenCV Library.” Dr. Dobb’s Journal of Software Tools, 2000. Available at: https://opencv.org/.
[6] MongoDB, Inc. “MongoDB Documentation.” Available at: https://docs.mongodb.com/.
[7] Richardson, L. “Beautiful Soup Documentation.” Available at: https://www.crummy.com/software/BeautifulSoup/.