Social Media Data Science Pipelines Project Proposal: 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 Proposal: Dataset Measurements and Analysis. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Social media platforms like Reddit and 4chan play a pivotal role in shaping public opinion on political topics, providing spaces where users discuss, react to, and debate current events. Political discussions on these platforms offer unique insights into public sentiment, controversy, and engagement around significant issues. With the 2024 election approaching and political discourse intensifying, analyzing user interactions on these platforms has become essential for understanding how online communities respond to political developments.

In this project, we will collect and analyze political data from Reddit and 4chan to uncover patterns in user behavior, sentiment, and controversy over time. By examining trends in activity and emotional responses within political discussions, we aim to gain insight into how users on each platform respond to topics surrounding the 2024 election. This analysis will help identify trends in political discourse, offering a foundation for understanding the dynamics of social media interactions around elections.

2 PROPOSED METHODOLOGY

Our methodology includes four distinct analysis components, each aimed at uncovering patterns within the data and facilitating comparisons between Reddit and 4chan:

(1) **Toxicity Over Time:** Using ModerateHatespeech API, we will assess and compare toxicity trends over time between Reddit and 4chan. Data on toxicity scores will be plotted against time to observe changes, focusing particularly on weeks leading up to the election.
(2) **CDF Analysis on Comment Volume and Length:** We will generate a cumulative distribution function (CDF) graph, comparing the number of comments across time as well as variations in text length for each comment as the election approaches.
(3) **Daily Submissions and Engagement Analysis:** For Reddit's r/politics, we will plot the number of daily submissions from November 1 to November 14, 2024. This will be used to examine possible spikes in activity.
(4) **Comparison of Comments per Hour:** For the additional requirement, we will analyze comment frequency by hour for both Reddit and 4chan, which may highlight specific hours of increased activity. This analysis will span from November 1 to November 14, 2024.

3 ANTICIPATED CHALLENGES AND SOLUTIONS

Several challenges are anticipated, particularly around:

- **API Limitations**: Potential outages or throttling issues with ModerateHatespeech will be mitigated by adding retry mechanisms and cache layers for unprocessed data.
- **Storage and Scalability**: To handle increasing data volume, we will create batch-processing scripts to offload historical data to external storage periodically, ensuring we meet storage constraints. To enhance data scalability, we will request additional virtual machine resources to accommodate increased data processing needs. This will allow us to efficiently manage larger datasets and optimize performance as our data requirements grow.

4 PLANNED EXPERIMENTS AND EXPECTED OUTPUTS

Three research questions are central to this project:

- How does the volume and nature of content change over time across both platforms?
- What are the observable trends in toxicity scores in relation to discussion topics?
- How does content engagement differ between Reddit and 4chan?

5 ADDITIONAL DATA COLLECTION AND RESOURCES

Our data collection pipeline includes a continuously operating crawler for both Reddit and 4chan that stores data at regular intervals, ensuring we have adequate volume and completeness for analysis. For hate speech score we will use ModerateHatespeech 3rd party api. No additional datasets are required for this stage, and the current dataset should be sufficient.

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

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

6 DATA EXTRACTION AND TRANSFORMATION PROCESS

In this section, we detail our approach for processing data collected from Reddit and 4chan to facilitate our analyses on user engagement, sentiment, and controversy. In this project all the analysis will be done after dumping relevant data into database. We will use python for analysis where required rows will be fetched according to our needs.

6.1 Reddit Data Structure

The data from Reddit includes various attributes for each comment, such as:

- **id:** Unique identifier for the comment
- **body:** The text content of the comment
- **author:** The username of the commenter
- **created_utc:** The timestamp of when the comment was created
- **ups:** The number of upvotes the comment received
- **downs:** The number of downvotes the comment received
- **link_url:** The URL of the related discussion or article
- **subreddit:** The subreddit from which the comment originates
- **num_comments:** The number of replies to the comment
- **score:** The total score (upvotes - downvotes) of the comment
- **permalink:** The direct link to the comment
- **hateScore:** Hate score generated through api

6.2 4chan Data Structure

The data from 4chan includes various attributes for each post, such as:

- **h:** Height of the image (if applicable)
- **w:** Width of the image (if applicable)
- **no:** Unique identifier for the thread
- **com:** The text content of the post
- **ext:** The file extension of the image (if applicable)
- **md5:** The MD5 hash of the image file
- **now:** Timestamp indicating when the post was created
- **tim:** Time in Unix timestamp format
- **name:** The username of the poster (often "Anonymous")
- **time:** The time when the post was created in Unix timestamp format
- **tn_h:** Height of the thumbnail image (if applicable)
- **tn_w:** Width of the thumbnail image (if applicable)
- **fsize:** File size of the image (if applicable)
- **replies:** The number of replies to the post
- **archived:** Indicates whether the post is archived (1 for yes, 0 for no)
- **filename:** The name of the uploaded image file
- **country:** The country of the poster
- **country_name:** The full name of the country
- **semantic_url:** A human-readable URL representation of the post
- **hateScore:** Hate score generated through api

6.3 Data Transformation Steps

To transform this data for analysis, we will follow these steps:

(1) **Data Cleaning:** Remove irrelevant or incomplete entries to ensure only meaningful comments and posts contribute to our analysis.
(2) **Timestamp Conversion:** Convert ‘created_utc‘ (Reddit) and ‘now‘, ‘time‘ (4chan) timestamps to a readable date format, aiding chronological organization and trend analysis.
(3) **Sentiment Analysis:** Classify the emotional tone of the text in ‘body‘ (Reddit) and ‘com‘ (4chan) into positive, negative, or neutral categories.
(4) **Hate Speech Scoring:** Assign Hate Speech scores based on engagement metrics like replies and sentiment analysis outcomes.
(5) **Keyword Extraction:** Extract keywords from ‘body‘ (Reddit) and ‘com‘ (4chan) fields to track prevalent themes over time.
(6) **Data Aggregation:** Aggregate the transformed data for trend analysis on daily, weekly, and monthly intervals.

7 GRAPH ANALYSIS

In this section, we describe the visualizations we plan to generate to better understand user interactions and sentiment trends on Reddit and 4chan. Each subsection details the specific graph type, its purpose, and the data metrics used.

7.1 Controversy Timeseries Graph

The controversy timeseries graph aims to visualize trends in controversy over a period of time, comparing data from both Reddit and 4chan within the same plot to determine which platform experiences more controversial discussions. In this graph, the x-axis represents dates, while the y-axis shows the number of comments. For each peak in the graph, we will analyze significant events occurring on those dates to understand user reactions to relevant news or discussions. By tracking these peaks across both platforms, we hope to draw conclusions about how users on each platform responded to specific topics or incidents, as well as identify which platform is generally more controversial.

7.2 Sentiment Analysis Graph

The sentiment analysis graph illustrates changes in various emotions over time, with lines for different emotions such as fear, anger, enjoyment, sadness, and disgust. On the x-axis, we will have dates, and on the y-axis, we will display the count of each sentiment. By extracting specific keywords from comments and tallying them daily, we can observe changes in sentiment trends. For each peak in a particular emotion line, we will investigate significant events around that date, aiming to understand how users reacted to those events.

7.3 Toxicity Over Time

This graph will plot the toxicity levels over time for both Reddit and 4chan using data from the ModerateHatespeech API. By comparing toxicity scores from both platforms, we can determine shifts in user

sentiment leading up to the election. The x-axis will represent dates, and the y-axis will display toxicity scores.

7.4 CDF Analysis on Comment Volume and Length

A cumulative distribution function (CDF) graph will compare the volume of comments and their text lengths across both platforms. This graph aims to highlight differences in engagement and verbosity between Reddit and 4chan as the election approaches. In above graph x axis will hold

7.5 Daily Submissions on Reddit's r/politics

This figure will display the number of daily submissions on Reddit's r/politics subreddit from November 1 to November 14, 2024. By observing submission volume on this subreddit, we can detect possible spikes in user engagement leading up to and after the election.

8 LIBRARIES TO BE USED

- Pandas
- NumPy
- matplotlib and seaborn
- Scikit-learn
- Plotly

9 CONCLUSION

Our project provides a comprehensive methodology for collecting, transforming, and analyzing political discourse on Reddit and 4chan. By focusing on toxicity trends, CDF analysis, daily engagement, and hourly activity, we aim to present a robust overview of user behavior and emotional responses leading up to the 2024 election. Through detailed visualizations and sentiment analysis, we expect to uncover meaningful insights into online political discourse and its influence on public opinion.