

Data Hunters Project 3 Proposal

Shubhendu Jadhav

State University of New York at Binghamton
Binghamton, NY, USA
sjadhav@binghamton.edu

Avanti Kopulwar

State University of New York at Binghamton
Binghamton, NY, USA
akopulwar@binghamton.edu

Abstract

This proposal describes our plan for Project 3 of CS 515: *Social Media Data Science Pipelines*. Building on the Reddit and 4chan data pipeline developed in Project 1 and the analytical framework from Project 2, we will now focus on answering one of our proposed research questions through a hands-on, interactive dashboard. The system will let users explore sentiment, toxicity, and posting activity across both platforms, as well as track stock related discussions, using an interface designed for interactivity, clarity, and smooth data visualization.

Keywords

Social Media Analysis, Reddit, 4chan, Stock Market, Toxicity, Sentiment, Dashboard

ACM Reference Format:

Shubhendu Jadhav and Avanti Kopulwar. 2025. Data Hunters Project 3 Proposal. In *CS 515 Project 3 Proposal*. ACM, New York, NY, USA, 2 pages.

1 Introduction

In Project 1, we built a live data-ingestion pipeline for Reddit and 4chan that continuously collects posts related to finance and market discussions. Project 2 focused on analyzing this data to measure sentiment using VADER, toxicity using Google's Perspective API, and overall posting activity trends. For Project 3, we aim to extend that analysis by answering a specific research question about how toxicity and sentiment change over time across platforms. We will also design an interactive web dashboard that allows users to explore these patterns directly, with flexible parameters and smooth visualization.

2 Research Question

The main research question we will answer in our final report is:

RQ3: How strongly do sentiment and toxicity correlate across platforms, and what do those correlations suggest about discourse tone?

This question builds naturally on our previous work and allows a focused comparison of cross-platform discussion patterns.

3 Analyses Included in the Interactive Tool

The dashboard will include at least three analyses from Project 2, each with adjustable parameters so users can explore specific time periods or metrics.

3.1 Posting Activity Over Time

- **Parameters:** platform (Reddit / 4chan), date range.

- **Output:** time series chart showing changes in post volume over time, highlighting major activity spikes.

3.2 Sentiment Distribution

- **Parameters:** platform, date window.
- **Output:** visualization of the share of positive, neutral, and negative posts based on VADER sentiment scores.

3.3 Toxicity Trends (Perspective API)

- **Parameters:** platform, toxicity threshold, date range.
- **Output:** trend plot showing how average toxicity changes over time and how it relates to posting surges.

3.4 Optional: Stock-Ticker Mentions

- **Parameters:** top N tickers.
- **Output:** frequency and sentiment trend for selected stock tickers.

These fulfill the requirement for at least three parameterized analyses and directly support our chosen research question.

4 Tools, Libraries, and Frameworks

We will use the following tools and libraries to build our dashboard and analysis system:

- **Flask:** backend framework for serving the dashboard and handling user queries.
- **Pandas:** for data processing, aggregation, and transformation.
- **Psycopg2:** for connecting to our PostgreSQL database.
- **VADER Sentiment Analyzer:** for text sentiment analysis.
- **Google Perspective API:** for toxicity scoring.
- **Matplotlib:** for generating static figures used in the report.
- **Plotly.js:** for interactive data visualization with zoom and hover features.
- **HTML/CSS/JavaScript:** for the front end layout and interactivity.

All of these are supported on the course VM and ensure full reproducibility.

5 Planned Style Enhancements (For Extra Style Points)

We plan to include several features that make the dashboard more interactive and visually polished.

5.1 Interactive Plotly Hover Tooltips

Each plot will include hoverable data points showing:

- average sentiment,
- average toxicity,

- post volume,
- and date or timestamp.

This adds useful context and helps users understand changes over time.

5.2 Side-by-Side Reddit vs. 4chan Comparison Toggle

A toggle will let users compare results for Reddit and 4chan side by side or as overlapping plots, highlighting differences in tone and toxicity between platforms.

5.3 Live Stock-Ticker Bar

We will include a small scrolling ticker showing trending stocks with the number of posts and sentiment direction. This provides a real time snapshot of what tickers are being discussed most actively.

5.4 Dark Mode Toggle

We will also add a dark mode option for improved readability and modern design using CSS class toggling.

6 Expected Outcome

By the end of Project 3, we will deliver:

- A working, interactive Flask based web dashboard,
- Three parameterized analyses with visualizations,
- An ACM format final report answering RQ3 with supporting evidence,
- A recorded 10-minute demo showcasing the tool and results.

7 Conclusion

This proposal outlines a clear, achievable plan that meets all course requirements for Project 3. Our goal is to combine solid data analysis with an intuitive, interactive dashboard that makes it easy to explore toxicity and sentiment dynamics across Reddit and 4chan. The final product will provide both meaningful insights and a polished presentation that reflects our full social media data pipeline.