Data Hunters Project 2 Proposal

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

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

Abstract

This proposal outlines our plan for Project 2 of *CS 515: Social Media Data Science Pipelines*. Building on our earlier datacollection system, we will now focus on measurement and analysis of financial discussions from Reddit and 4chan. Our goal is to describe how user sentiment, posting activity, and toxicity evolve over time across these two communities that often drive social-media-based market movements. We will integrate the Google Perspective API for toxicity measurement and conduct a series of experiments to visualize, compare, and better understand the emotional and behavioral patterns behind stock-related online discussions.

ACM Reference Format:
Shubhendu Jadhav and Avanti Kopulwar. 2025. Data Hunters Project 2 Proposal. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The increasing influence of social media on the stock market popularly seen during events like the GameStop surgehas made online communities a valuable lens for studying investor behavior. Reddit boards such as `r/wallstreetbets` and 4chan's `/biz/` board host thousands of daily posts reflecting market sentiment, collective anxiety, and meme driven speculation. While our Project 1 system continuously collected posts from these platforms, this phase shifts toward understanding what that data actually reveals. Through sentiment, toxicity, and activity analyses, we aim to capture how online discussions mirror or even anticipate broader market trends.

2 Dataset Description

Our dataset consists of continuously collected posts from multiple stock-related subreddits (`r/wallstreetbets`, `r/stocks`, `r/investing`, `r/Daytrading`, etc.) and 4chan's financial board `/biz/`. Each record includes post ID, subreddit or board, text content, timestamp, and engagement metadata. All data are stored in a PostgreSQL database hosted on a Binghamton VM, with automatic hourly backups to ensure data integrity. As of late October 2025, the dataset includes roughly 13,000 Reddit posts and 640,000 4chan threads.

Conference'17, Washington, DC, USA

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

3 Proposed Methodology

Our analysis will combine data querying, sentiment scoring, and toxicity measurement.

- **Data Extraction & Cleaning:** Posts will be pulled from the database and preprocessed to remove duplicates, URLs, emojis, and non-English text. Basic counts and distributions will verify that the crawlers remain consistent.
- **Sentiment Analysis:** Using the VADER model, we will compute sentiment scores for each post and categorize them as positive, neutral, or negative. These values will later be aggregated by day, subreddit, and stock ticker to identify mood shifts around specific events.
- **Toxicity Measurement via Perspective API:** To gauge the tone of conversations, we will integrate the Google Perspective API. Each comment will be cleaned and sent to the API, which returns a score between 0 and 1 representing the probability of toxic language. Scores will be stored alongside sentiment results, enabling joint analysis of mood vs. toxicity. Over time, we will observe how toxicity levels rise or fall with market volatility or news-driven surges.
- **Comparative Analysis:** We will compare Reddit and 4chan on metrics such as posting frequency, sentiment distribution, and average toxicity to identify platform-specific behavioral differences.
- **Visualization & Reporting:** All figures will be created in Python using Matplotlib and Pandas to ensure full reproducibility.
- **Libraies Used:** matplotlib, datetime, pandas, numpy.

4 Planned Figures and Tables

To clearly present findings, we plan to include several key visualizations:

- **Figure 1 – Post Growth Over Time:** A time series plot showing cumulative post counts from Reddit and 4chan. This will highlight crawler consistency and activity surges linked to major financial events.
- **Figure 2 – Sentiment Distribution:** A bar chart comparing proportions of positive, neutral, and negative posts across both platforms to highlight emotional differences.
- **Figure 3 – Top 10 Stock Tickers:** A horizontal bar plot of the most-mentioned tickers (e.g., AAPL, TSLA, GME) with their average sentiment values, showing which tickers trigger the strongest reactions.
- **Figure 4 – Toxicity Trends Over Time:** A line chart based on Perspective API scores illustrating daily average toxicity changes, indicating when market discussions turn hostile.

- **Figure 5 – Cross-Platform Comparison:** A dual-axis chart comparing sentiment trends and posting volume across Reddit and 4chan over time.
- **Table 1 – Dataset Summary:** A table summarizing total posts, average sentiment, average toxicity, and active subreddits/boards to provide quick dataset context.

These visuals together will show how mood, activity, and toxicity shift and overlap across platforms.

5 Planned Research Questions

(1) How does posting activity differ between Reddit and 4chan communities discussing the stock market?
(2) Which stock tickers are most frequently mentioned, and how does sentiment shift around major financial events?
(3) How do toxicity levels vary across platforms, and do spikes in toxicity align with heightened sentiment or activity?
(4) Is there a correlation between discussion volume and overall market volatility?
(5) Can sentiment and toxicity indicators help predict emerging trading trends or investor mood shifts?

6 Expected Outcome

By the end of Project 2, we expect to produce a comprehensive analysis of our dataset, supported by quantitative evidence and visual exploration. We aim to uncover how online financial discourse differs between Reddit and 4chan, how emotions fluctuate during market events, and how toxic language shapes these conversations.

Acknowledgments

We thank Professor Jeremy Blackburn and the CS515 course staff for their continuous guidance.

References

- Reddit API Documentation. Reddit, Inc. Available at: https://www.reddit.com/dev/api/
- Perspective API Documentation. Google. Available at: https://developers.perspectiveapi.com/s/about-the-api
- Academic paper: "Incivility or Invalidity? Evaluating Perspective API Scores as a Proxy for Toxicity." *American Behavioral Scientist, SAGE Publications, 2024.* Available at: https://journals.sagepub.com/doi/full/10.1177/1532673X241309627