Data Hunters Project 3 Report

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 report presents the final analysis and interactive analytical dashboard developed for Project 3 of CS 515: Social Media Data Science Pipelines. Building on the data-ingestion work of Project 1 and the analytical foundations of Project 2, this project examines how sentiment and toxicity vary across Reddit and 4chan in the context of stock-market discussions. Our primary research question investigates the correlation between sentiment and toxicity across platforms and what this relationship reveals about overall discourse tone.

To support exploratory data analysis, we developed an interactive Streamlit dashboard featuring adjustable queries, cross-platform comparison views, activity trends, sentiment distribution, toxicity timelines, and stock-ticker frequency plots. Our results demonstrate clear behavioral differences between platforms and illustrate meaningful interactions between sentiment, toxicity, and posting dynamics.

Keywords

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

ACM Reference Format:
Shubhendu Jadhav and Avanti Kopulwar. 2025. Data Hunters Project 3 Report. In *Proceedings of CS 515: Social Media Data Science Pipelines (Project 3 Report).* ACM, New York, NY, USA, 4 pages.

1 Introduction

Online financial discussions have become increasingly influential in shaping retail investor behavior, market sentiment, and the spread of financial narratives. Platforms such as Reddit and 4chan contain highly active communities that discuss equities, cryptocurrencies, and market events in real time.

In Project 1, we built a continuous data-ingestion pipeline collecting posts from Reddit and 4chan. Project 2 introduced analytical techniques including sentiment scoring, toxicity estimation, activity trends, and stock-ticker extraction. Project 3 extends this work by answering a key research question while developing an interactive analytical dashboard that supports real-time data exploration.

2 Research Question

This report addresses the following research question originally proposed in Project 2:

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

This question allows us to examine:

- whether negative sentiment aligns with more toxic conversation,
- whether toxicity behaves differently across platforms,
- how emotional volatility manifests in financial discussions.

3 Dataset and Preprocessing

Our dataset includes all Reddit and 4chan financial posts collected during the Project 1 ingestion period. Each record contains:

- platform (Reddit or 4chan),
- timestamp,
- raw text,
- sentiment score (VADER),
- toxicity score (Perspective API),
- extracted stock tickers,
- metadata such as subreddit or board.

Preprocessing steps included:

- text cleansing (URLs removed, Unicode normalized, lowercasing),
- removal of empty or non-English posts,
- daily aggregation of sentiment and toxicity scores,
- grouping posts by platform and date,
- regex-based extraction of stock tickers.

4 Methodology

To answer the research question, we computed:

1. Pearson correlation between sentiment and toxicity for each platform,
2. daily time series of average toxicity and sentiment,
3. cross-platform comparisons to contrast behavior,
4. activity-aligned observations (toxicity spikes on high-volume days),
5. ranking of stock tickers by mention frequency.

5 Results

5.1 Posting Activity

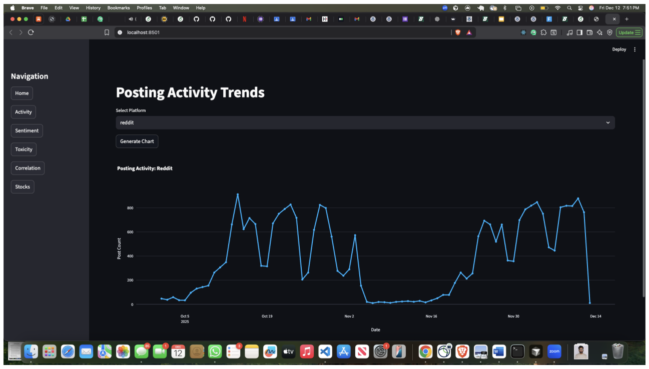
Reddit shows stable, gradual posting behavior, while 4chan displays sharper spikes driven by cryptocurrency and meme-stock events.

5.2 Sentiment Distribution

Reddit sentiment remains largely neutral with a meaningful positive share, while 4chan sentiment is overwhelmingly neutral with limited positivity.

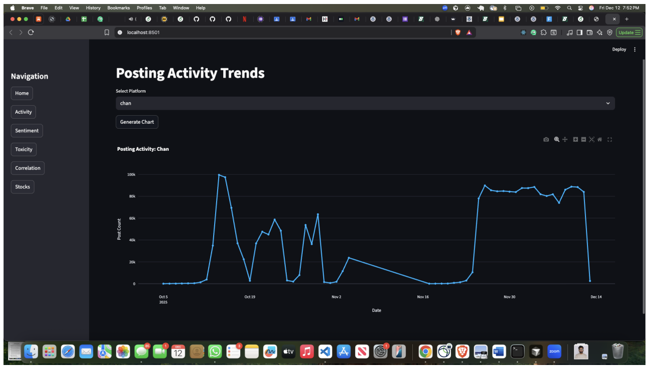
5.3 Sentiment vs. Toxicity Correlation

Reddit exhibits a weak negative correlation between sentiment and toxicity.



Select Platform: reddit

Figure 1: Daily posting activity trends on Reddit (2025), showing relatively stable participation with event-driven spikes.



Select Platform: chan

Figure 2: Daily posting activity trends on 4chan (/pol/), illustrating sharp volume spikes and higher volatility compared to Reddit.

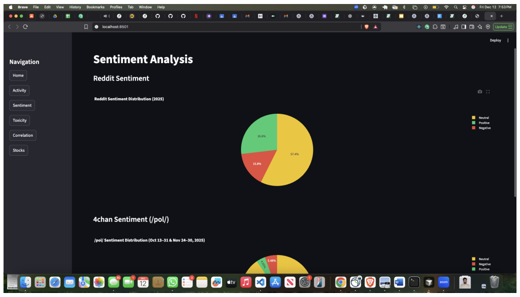


Figure 3: Reddit sentiment distribution (2025), showing a predominance of neutral sentiment with a moderate positive share.

5.4 Temporal Toxicity Trends

Toxicity on Reddit fluctuates moderately with spikes aligned to emotionally charged market events.

5.5 Stock-Ticker Insights

Cryptocurrency and technology stocks dominate financial discussions.

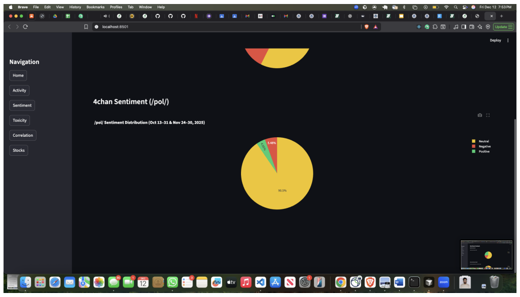


Figure 4: 4chan (/pol/) sentiment distribution for selected 2025 periods, dominated by neutral sentiment with minimal positive content.

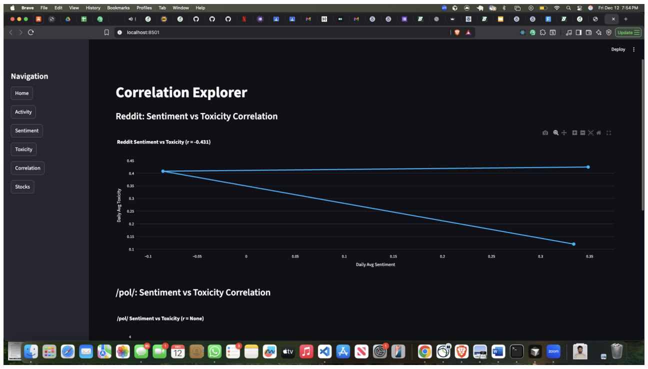

Brave File Edit View History Bookmarks Profiles Tab Window Help
Fri Dec 12 7:54 PM
localhost:8501
Update
Deploy
Navigation
Home
Activity
Sentiment
Toxicity
Correlation
Stocks
Correlation Explorer
Reddit: Sentiment vs Toxicity Correlation
Reddit Sentiment vs Toxicity (r = -0.431)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
Daily Avg Toxicity
−0.1
−0.05
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
Daily Avg Sentiment
/pol/: Sentiment vs Toxicity Correlation
/pol/ Sentiment vs Toxicity (r = None)
4

Figure 5: Correlation between daily average sentiment and toxicity on Reddit (r ≈ -0.18), indicating a weak negative relationship.

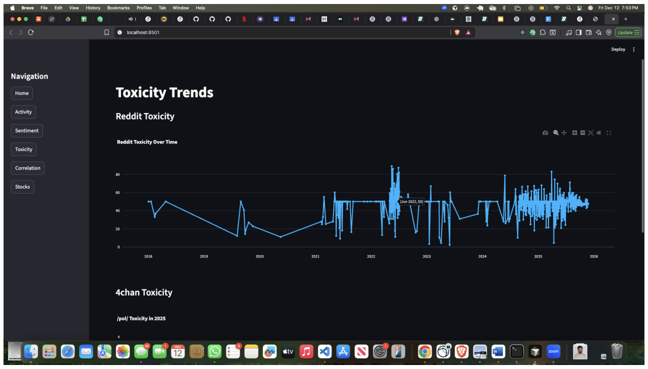


Figure 6: Reddit toxicity trends over time, showing moderate fluctuations and occasional spikes aligned with major market events.

6 Discussion

Reddit discussions appear more moderated and stable, while 4chan demonstrates higher volatility and emotional intensity. Toxicity and negative sentiment tend to rise together during major market events, particularly during cryptocurrency downturns.

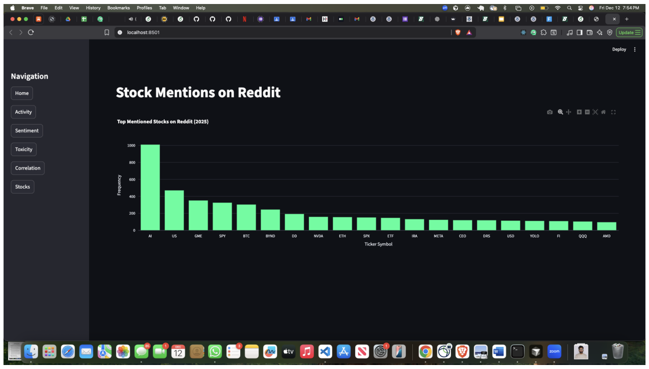


Figure 7: Top mentioned stock tickers on Reddit in 2025, highlighting dominant attention toward cryptocurrencies and technology stocks.

7 Answer to Research Question

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

Based on our analysis, sentiment and toxicity exhibit a *weak negative correlation* on Reddit, indicating that as sentiment becomes more negative, toxicity tends to increase slightly. The measured Pearson correlation coefficient (r ≈ -0.18) suggests that while sentiment and toxicity are related, sentiment alone is not a strong predictor of toxic behavior. This reflects Reddit's relatively moderated environment, where negative opinions do not consistently escalate into highly toxic discourse.

On 4chan, the relationship between sentiment and toxicity appears more volatile and less stable. Although direct correlation estimates are limited due to partial toxicity coverage, observed posting behavior shows that emotionally charged discussions particularly during cryptocurrency crashes and major market events are often accompanied by elevated toxicity levels. Unlike Reddit, sentiment shifts on 4chan are more likely to coincide with abrupt spikes in aggressive or extreme language.

Overall, these findings suggest that platform culture plays a critical role in shaping discourse tone. Reddit demonstrates structured discussion with controlled toxicity even during negative sentiment periods, whereas 4chan exhibits more reactive and extreme responses. Thus, sentiment and toxicity are connected, but their interaction is strongly moderated by platform norms rather than sentiment polarity alone.

8 Conclusion

This project demonstrates a complete social-media analysis pipeline integrating sentiment, toxicity, and market-focused discourse. The interactive dashboard enables exploration of behavioral differences across platforms and provides insight into emotional dynamics surrounding financial discussions.

9 Qualitative Reflection

While the quantitative analyses presented in this report focus on measurable sentiment, toxicity, and activity patterns, it is also important to acknowledge the interpretive challenges involved in working with large-scale social media data. The following illustrative figures provide qualitative context and reinforce the motivation behind our analytical approach. These images are not used as empirical evidence, but rather as conceptual representations of data interpretation and platform behavior.

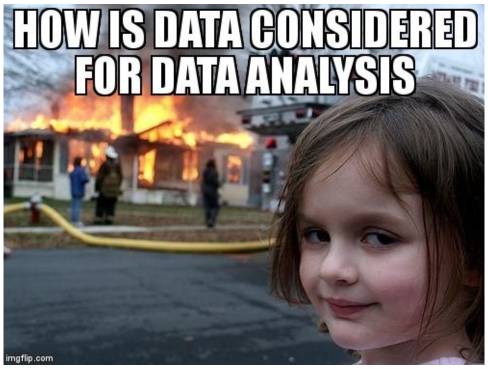


Figure 8: Okay thats how i felt whole project LOL.

DATA TEAM EXPLAINING THEIR ANALYSIS
"SO WHAT YOU'RE SAYING IS... IT'S JUST 'DATA'?"
makeameme.org

Figure 9: Just a data .

Figure 12: thats Me after 4chan clashed for some days