Pipeline Maverick Project 3 Report

Interactive Toxicity Analysis Dashboard

Tripti Ramesh Anchan
Department of Computer Science, SUNY Binghamton
Binghamton, New York, USA
tanchan@binghamton.edu

1 Introduction

Project 2 revealed several important patterns in online toxicity across 4chan and Reddit. It showed that 4chan is consistently more toxic, that identity-based attacks drive much of the difference between the two platforms, and that 4chan often responds more quickly to major geopolitical events. While these insights were valuable, they represented only static snapshots of behavior. Users could see the patterns, but they couldn't interact with the data, adjust thresholds, or explore how toxicity shifts across different communities and over time.

Project 3 builds on this work by turning those static results into a fully interactive dashboard. Instead of passively viewing results, users can now filter by platform, community, time window, and toxicity level, allowing them to actively explore how toxic behavior develops and changes across platforms.

In addition to enhancing interactivity, this project also answers one of the central research questions proposed in Project 2:

RQ1: What language actually marks a discussion as toxic?

To address this question, the dashboard integrates a TF-IDF analysis that compares the vocabulary of high-toxicity and low-toxicity posts. This allows the system to move beyond simple keyword counts and uncover the specific language patterns that distinguish genuinely toxic discourse. Together, the interactive visualizations and TF-IDF analysis provide a hands-on, exploratory view of how toxicity manifests across platforms.

2 Research Questions and Analyses

This report presents the following analyses:

- **Analysis 1: Toxicity Distribution.** This analysis visualizes how toxicity scores are distributed across Reddit and 4chan using histograms and CDFs. Users can filter by platform, and date range to observe how toxicity behavior shifts.
- **Analysis 2: Keyword Frequency in High vs. Low Toxicity Posts.** This module compares the vocabulary used in high toxic posts versus low toxic posts. Users can adjust toxicity thresholds, apply date filters, and examine which keywords spike during toxic discussions.
- **Analysis 3: Multi-Attribute Toxicity Breakdown.** This analysis decomposes toxicity into six attributes, including insults, profanity, threats, toxicity, severe toxicity, and identity attacks. Users can compare platforms and examine how different forms of toxicity vary across communities.
- **Analysis 4: TF-IDF Toxic Vocabulary Analysis (RQ1).** This new analysis identifies platform-specific toxic vocabularies by computing TF-IDF scores for high-toxicity versus low-toxicity posts. This reveals which terms are most strongly associated with toxic discourse on each platform.

Together, these analyses transform the static findings of Project 2 into a dynamic, user-driven exploration system that both visualizes toxicity patterns and reveals the specific language that defines toxic discussions.

3 System Overview

The interactive dashboard is built using a lightweight Python stack connected to the TimescaleDB database used in Projects 1 and 2. Users interact with browser-based controls that trigger API calls to the Flask backend, which retrieves posts, runs analyses, and generates interactive figures returned to the frontend.

Figure 1 illustrates the full architecture.

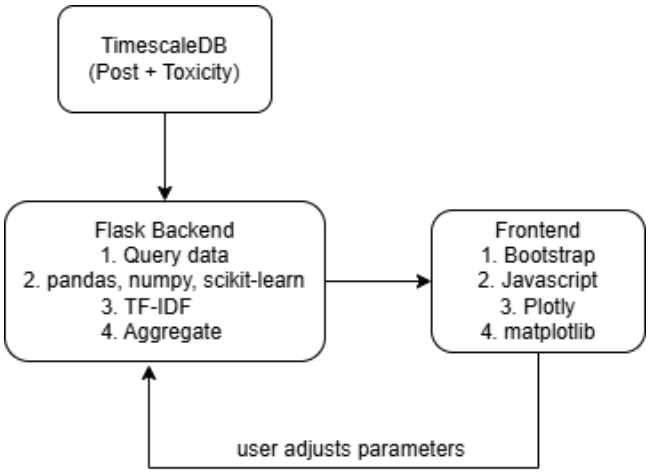

Figure 1: System architecture connecting the dashboard frontend, Flask backend, and TimescaleDB.

4 Results

The dashboard integrates three core analyses from Project 2 along with a new TF-IDF module to answer RQ1. Together, these results show consistent cross-platform differences in both the level and nature of toxic language.

4.1 Toxicity Distribution

Figures 2 and 3 show how toxicity scores are distributed across 4chan and Reddit. Reddit's scores cluster near zero, while 4chan spreads across a much wider range of mid- and high-toxicity values, indicating both higher and more variable toxicity. The CDF shows that nearly 95% of Reddit posts fall below a toxicity score of 0.5, compared to only about 70% on 4chan, making elevated toxicity

routine on 4chan rather than rare. Users can interactively filter by platform and time period and see both histogram and CDF views.

Figure 2: Perspective API toxicity distribution for 4chan and Reddit.

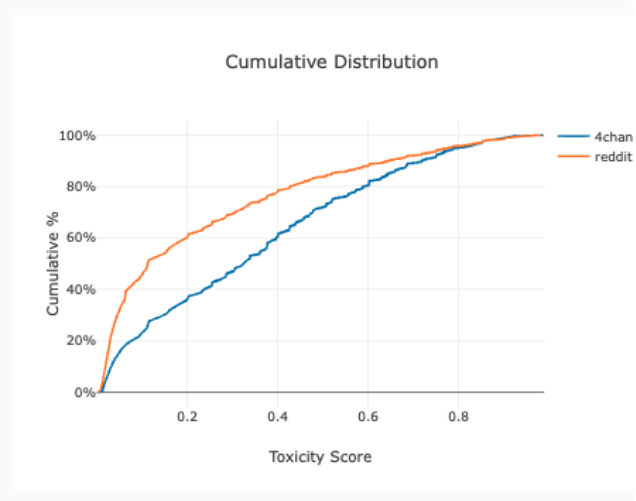

Figure 3: Cumulative distribution of toxicity scores.

4.2 Keyword Frequency in High vs Low Toxicity

Figure 4 compares keyword frequencies between high-toxicity (score > 0.35) and low-toxicity (score ≤ 0.35) posts. Identity-focused terms like "jew" and "muslim" show sharp increases in high-toxicity contexts, especially on 4chan. In contrast, geopolitical terms such as "russia" and "china" appear consistently across both toxicity groups, suggesting they reflect topic rather than hostility. This demonstrates that toxicity is driven by identity-based language rather than simply discussing controversial political subjects.

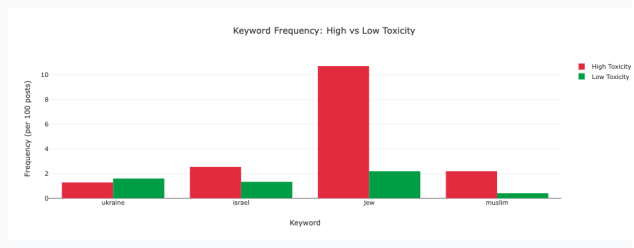

Figure 4: Keyword frequencies for high vs. low toxicity posts.

4.3 Multi-Attribute Toxicity Breakdown

Figure 5 breaks toxicity into six components. 4chan scores higher on every attribute, with the largest gaps appearing in identity attacks, severe toxicity, and threats. These results indicate that platform differences are driven not just by general negativity, but by targeted and extreme forms of hostility.

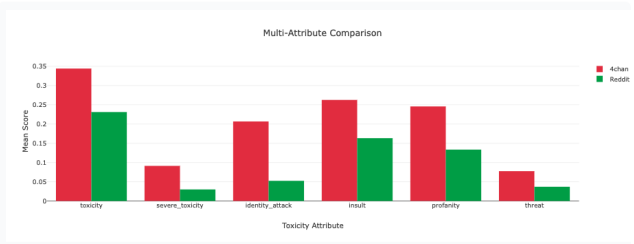

Figure 5: Multi-attribute toxicity comparison between platforms.

4.4 TF-IDF Toxic Vocabulary (RQ1)

Figures 6 and 7 show the top TF-IDF-scoring words in toxic posts for both platforms. The most distinctive terms reveal platform-specific toxic vocabularies: 4chan features explicit identity slurs, while Reddit shows political language ("trump", "fuck", "shit") and dismissive terms ("stupid", "guy"). These words appear frequently in toxic posts but rarely in non-toxic ones, indicating a strong association with hostile discourse.

Table 1 compares the top toxic words across platforms. 4chan's toxic vocabulary is dominated by identity-based slurs, while Reddit emphasizes political figures and profanity. Both platforms share common profanity ("fuck", "fucking"), but the context and co-occurring terms differ substantially.

The TF-IDF analysis reveals that toxic conversations are characterized by:

- **Platform-specific vocabularies:** 4chan uses explicit slurs, while Reddit focuses on political figures and accusations.
- **Direct identity-based slurs:** More prevalent on 4chan ("jews", explicit identity terms) versus Reddit's political labeling ("nazi", "trump").
- **Shared profanity:** Both platforms use aggressive language ("fuck", "fucking", "shit"), but 4chan scores higher on TF-IDF scale (0.09 max vs. Reddit's 0.04 max).

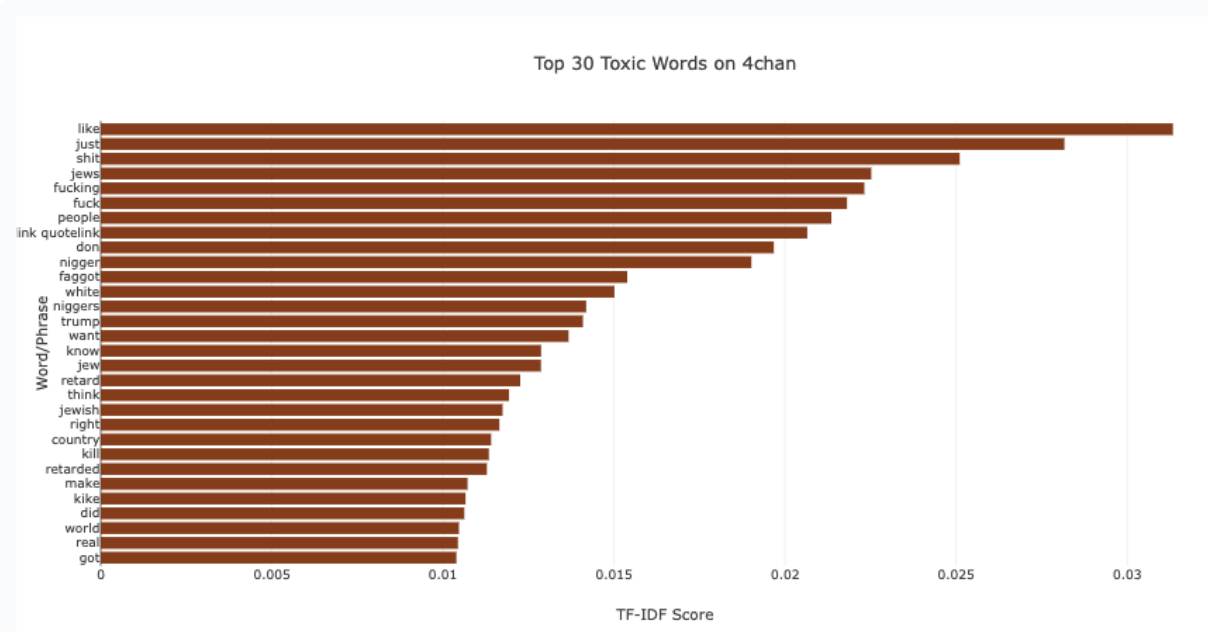

Figure 6: Top 20 TF-IDF toxic vocabulary for 4chan.

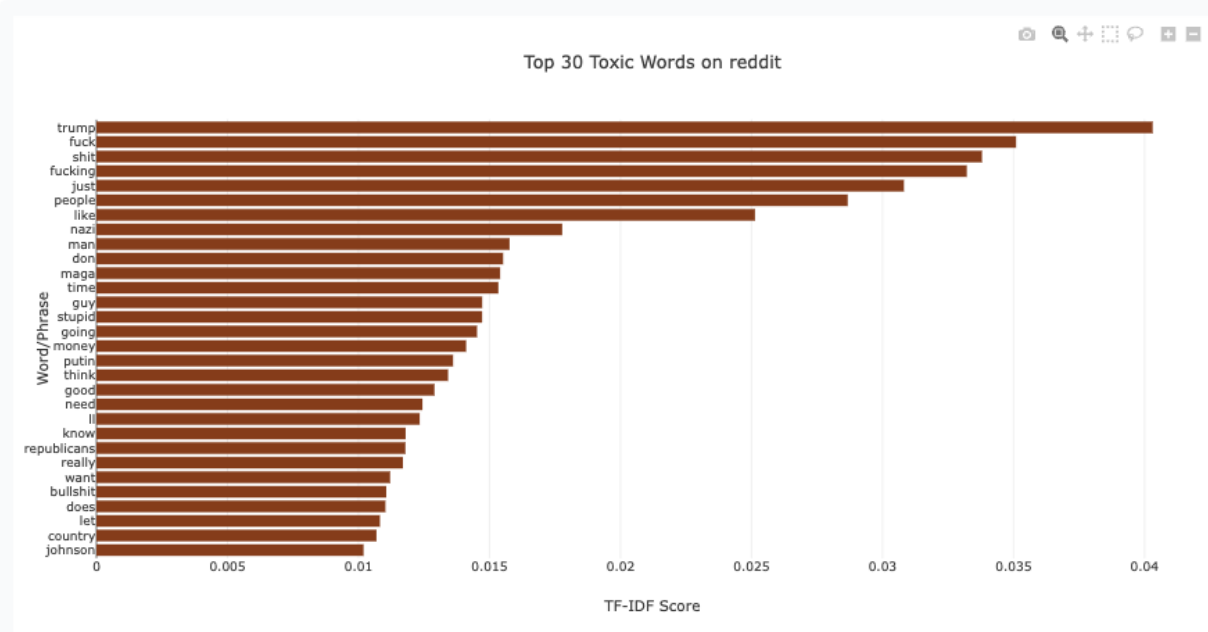

Figure 7: Top 20 TF-IDF toxic vocabulary for Reddit.

Overall, the results confirm that toxicity on 4chan is not only higher in magnitude than on Reddit, but also linguistically distinct. Toxic discussions are primarily marked by identity-based hostility, targeted slurs, and intensified profanity, directly answering RQ1.

5 Challenges

5.1 Event-Response Analysis

The event-response analysis from Project 2 was not included in the interactive dashboard because it relies on keywords having clear peak activity days. While this works well for static analyses of known events (e.g., "ukraine", "gaza"), it becomes unreliable in

Table 1: Top 10 TF-IDF Toxic Words by Platform

Rank	4chan	Reddit
1	jews	trump
2	fuck	fuck
3	fucking	shit
4	shit	fucking
5	people	just
6	like	people
7	just	like
8	don	nazi
9	quot	man
10	br	maga

an open-ended interactive setting. Many keywords in the dataset show steady activity rather than sharp spikes, which can lead to empty or misleading results when used for temporal alignment.

5.2 Date Range Limitations Across Analyses

A key limitation across all date-based analyses is that toxicity scores are not available for every post. Due to API rate limits, scored data is unevenly distributed over time. As a result, some user-selected date ranges may contain very few or no scored posts, leading to incomplete or empty visualizations. For more reliable results, broader date ranges often work better than narrow time filters.

6 Conclusion

This project builds an interactive dashboard that extends Project 2's static findings into a dynamic, user-driven exploration tool. Through a combination of distributional analyses, multi-attribute comparisons, and TF-IDF vocabulary extraction, the system reveals how toxicity manifests across platforms and what language actually defines toxic discussions.

The final result is a more accessible, interpretable view of online toxicity that allows users to explore patterns, adjust assumptions, and form their own insights.

I attempted to work on 5 analyses in total as proposed, of which I could successfully implement 4.