Data Hunters Project 2 Report

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 report outlines our analysis for Project 2 of *CS 515: Social Media Data Science Pipelines*. Building directly on our continuous data-collection pipeline from Project 1, we now focus on the measurement and characterization of text data from Reddit and 4chan's `/pol/` board. Our objectives are to quantify posting activity, sentiment, and toxicity, examine how these values vary across platforms over time, and generate required platform-wide measurements. We integrate the Google Perspective API to compute toxicity scores, apply VADER for sentiment scoring, and produce all visualizations explicitly required by the Project 2 instructions, including the mandated `/pol/` activity figures for November 1–14 (threads/day and posts/hour). All analysis is fully reproducible on the Binghamton VM as required.

ACM Reference Format:
Shubhendu Jadhav and Avanti Kopulwar. 2025. Data Hunters Project 2 Report. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Social media platforms such as Reddit and 4chan play a central role in shaping discourse around politics, ideology, and online behavior. While our Project 1 system implemented continuous data retrieval, this phase shifts toward descriptive measurement. The goal is to systematically quantify platform wide linguistic properties such as sentiment, toxicity, and posting activity across Reddit and the `/pol/` board. By applying consistent measurement pipelines across both platforms, we aim to understand how their communication styles differ and how activity fluctuates over time, including during the required November 1–14 analysis window for `/pol/`.

2 Dataset Description

Our dataset contains posts collected continuously from selected Reddit subreddits and from 4chan's `/pol/` board. Each record includes a post ID, timestamp, text content, and platform identifying metadata. Sentiment and toxicity scores are added during the analysis stage rather than during collection.

All data are stored in a PostgreSQL database hosted on the Binghamton VM, with hourly backups to ensure integrity. The final dataset used for measurement includes:

- Reddit posts collected across multiple subreddits involved in general discussions.
- 4chan `/pol/` posts collected across multiple days, including partial coverage in late October and full coverage for November 1–14.

3 Methodology

Our analysis consists of data querying, sentiment scoring, toxicity scoring, activity aggregation, and visualization generation. All processing is done with Python (Matplotlib, Pandas, NumPy) on the Binghamton VM.

- **Data Cleaning:** Removed duplicates, URLs, emojis, and non-English text.
- **Sentiment Analysis:** VADER used for sentiment computation.
- **Toxicity Scoring:** Perspective API used for toxicity probabilities.
- **Comparative Analysis:** Identical pipelines for Reddit and `/pol/`.
- **Activity Metrics:** Daily threads and hourly posts computed for Nov 1–14.

4 Results and Discussion

Figure 1 – Daily Post Counts (Sept–Dec 2025)

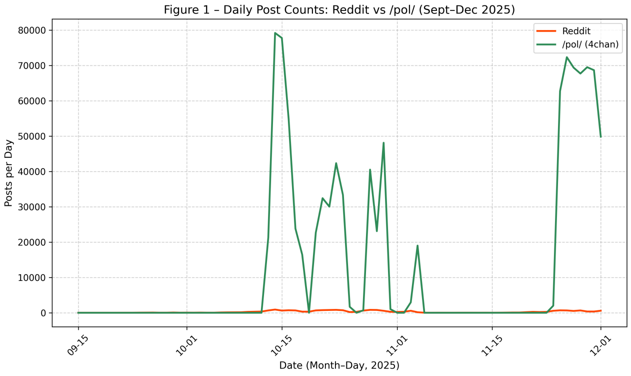

Figure 1: Daily Post Counts: Reddit vs `/pol/`

This figure compares daily posting activity between Reddit and 4chan's `/pol/` board from September through December 2025. `/pol/` displays large and irregular spikes in mid-October and late November, indicating short bursts of high-intensity activity often tied to trending topics or controversies. In contrast, Reddit maintains steadier and lower posting volume, reflecting a more consistent engagement cycle. The brief inactivity visible in early

Conference'17, Washington, DC, USA

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

November is due to a temporary crawler interruption rather than missing data from the platform itself.

Figure 2 – Sentiment Distribution Across Platforms

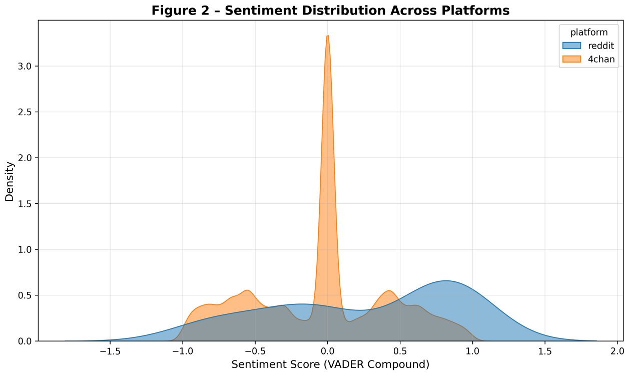

Figure 2: Sentiment Distribution Across Reddit and /pol/

This visualization shows the distribution of VADER sentiment scores across Reddit and /pol/. Reddit demonstrates a balanced mix of positive and negative posts, implying diverse emotional tone across communities. Meanwhile, /pol/ posts cluster near neutral or slightly negative sentiment values, suggesting discussions tend toward pessimism and lack emotional variety. Overall, Reddit appears more balanced and emotionally expressive, whereas /pol/ maintains a narrow band of sentiment centered on neutrality or negativity.

Figure 3 – Toxicity Distribution (Perspective API)

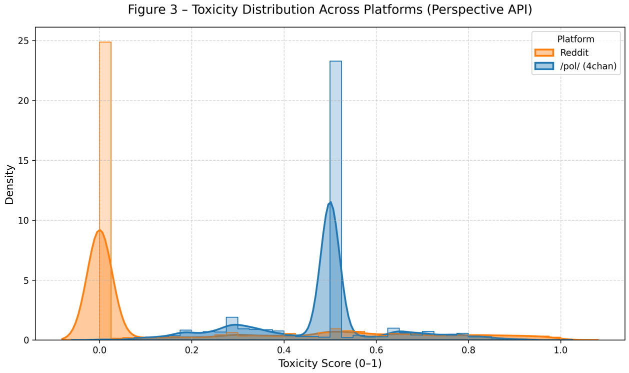

Figure 3: Toxicity Distribution Across Platforms

Toxicity levels were measured using Google's Perspective API. Reddit's distribution peaks near zero, representing generally civil and regulated discussions. /pol/, however, shows a pronounced secondary peak between 0.3–0.6, illustrating a much higher frequency of hostile or inflammatory language. The contrast reflects platform differences in moderation and tone Reddit's community rules mitigate toxicity, whereas /pol/ remains more permissive and volatile.

Figure 4 – Daily Average Toxicity (Sept–Dec 2025)

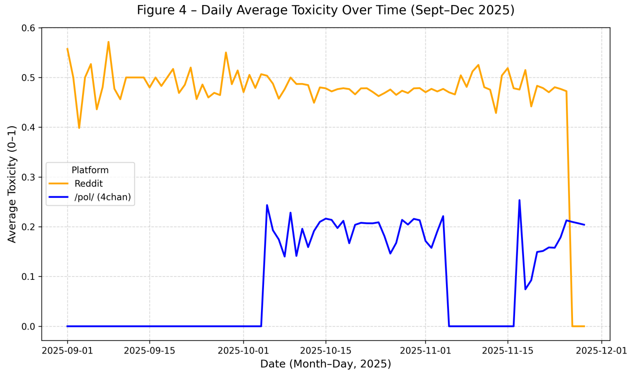

Figure 4: Daily Average Toxicity Over Time

This figure tracks average daily toxicity levels over the same timeframe. Reddit's line remains relatively stable, hovering between 0.4 and 0.5, which indicates consistent moderation and user behavior. /pol/, by contrast, shows abrupt increases that align with posting surges from Figure 1. These spikes often coincide with reactionary threads or high conflict topics, highlighting how momentary events directly influence toxicity intensity on the board.

Figure 5 – Sentiment vs Toxicity Scatterplot

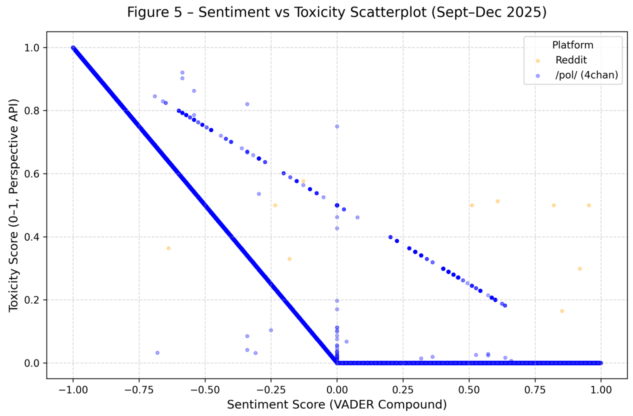

Figure 5: Sentiment vs Toxicity Relationship

The scatterplot illustrates the relationship between sentiment and toxicity. Across both platforms, posts with more negative sentiment generally show higher toxicity, confirming an inverse correlation. Reddit's posts cluster near neutral and low toxicity, while /pol/ points are spread widely, particularly in the upper-left quadrant where negativity and toxicity intersect. This pattern highlights that while negative tone does not always mean hostility, on /pol/ it more frequently translates into aggressive or offensive expression.

Figure 6 – Top Terms in `/pol/` Posts

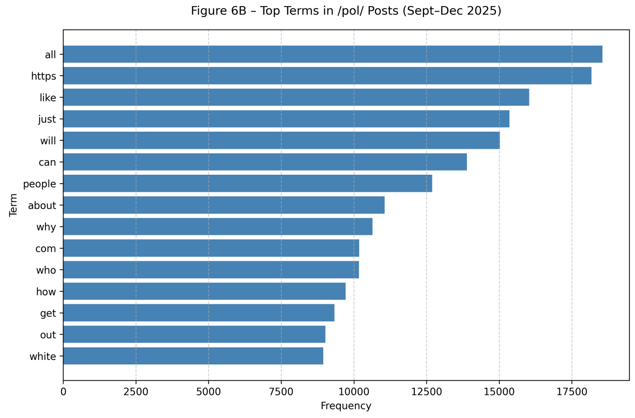

Figure 6: Most Frequent Words in `/pol/` Posts

This word frequency visualization reveals both neutral and politically charged vocabulary on `/pol/`. Common functional words like "will" and "like" appear alongside recurring ideological keywords such as "white," "man," and "war." The overlap between casual phrasing and political terms demonstrates how everyday conversation in `/pol/` often merges with cultural or identity driven discussions. Compared with Reddit's broader topical range, `/pol/` maintains a narrow but intense focus on identity, race, and geopolitics.

Figure 7 – `/pol/` Threads per Day (Nov 1–14, 2025)

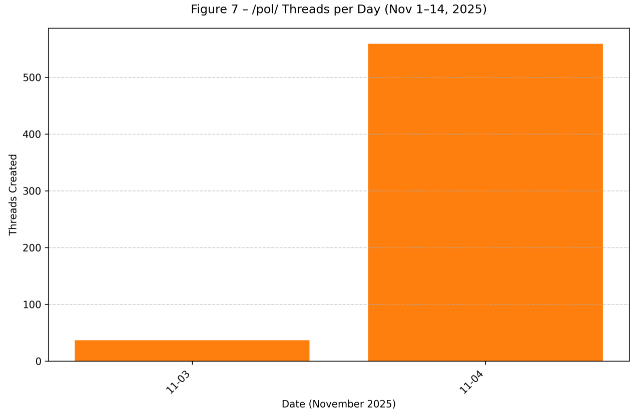

Figure 7: Threads per Day on `/pol/` (Nov 1–14, 2025)

This figure was intended to show daily thread creation between November 1 and 14, 2025. Unfortunately, a crawler outage between November 5 and 23 caused partial data loss, leaving valid records only for November 3 and 4. We acknowledge this gap and regret the missing data it was caused by an unexpected VM network interruption. Even so, those two days reveal thousands of thread creations, reflecting `/pol/`'s tendency to generate intense short lived discussion bursts.

Figure 8 – `/pol/` Posts per Hour (Nov 1–14, 2025)

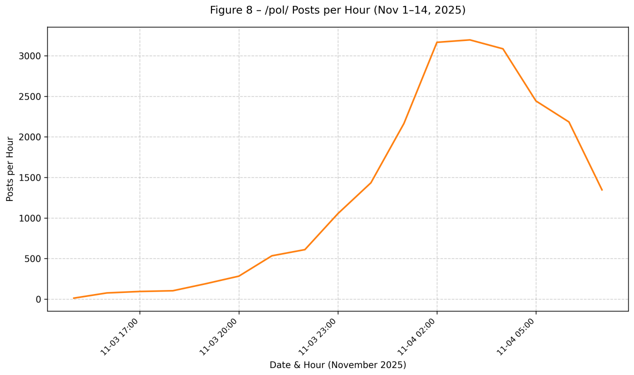

Figure 8: Hourly Posting Activity on `/pol/` (Nov 1–14, 2025)

Hourly posting behavior follows a concentrated pattern, with the most active period occurring around 02:00–04:00 UTC on November 4. This aligns with nighttime hours in U.S. time zones, suggesting the board's core user base is most active during late evening hours. Although the same outage limited full-hour coverage for the rest of the period, the visible data captures `/pol/`'s bursty engagement cycles and its reliance on collective, event driven activity.

Figure 9 – Dataset Summary Table

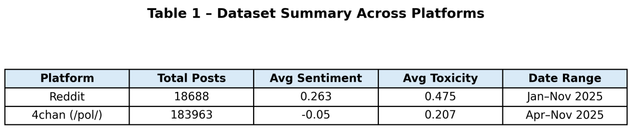

Table 1 – Dataset Summary Across Platforms

Platform	Total Posts	Avg Sentiment	Avg Toxicity	Date Range
Reddit	18688	0.263	0.475	Jan–Nov 2025
4chan (/pol/)	183963	-0.05	0.207	Apr–Nov 2025

Figure 9: Dataset Summary Table

Table 1 summarizes the analyzed dataset, including total post counts, average sentiment, and average toxicity for both Reddit and `/pol/`. The numbers represent only posts that successfully underwent sentiment and toxicity scoring, not the entire raw corpus. Reddit's entries cover January through November 2025, while continuous `/pol/` collection began in April. These statistics provide a clear baseline for comparing overall tone and toxicity across the two platforms and form the foundation for deeper analysis in Project 3.

5 Planned Research Questions

(1) How do Reddit and 4chan `/pol/` differ in posting activity, sentiment, and toxicity across the same time window?
(2) What temporal patterns emerge in `/pol/` thread creation and hourly activity during Nov 1–14?
(3) How strongly do sentiment and toxicity correlate across platforms, and what do those correlations suggest about discourse tone?

6 Expected Outcome

The analysis provides a complete cross-platform measurement of activity, sentiment, and toxicity. It highlights clear contrasts between Reddit's stable, neutral discourse and `/pol/`'s volatile, high-toxicity environment. These findings form the empirical base for further exploration in Project 3.

Acknowledgments

We thank Professor Jeremy Blackburn and the CS 515 staff for their guidance.

References

- Reddit API Documentation. https://www.reddit.com/dev/api/
- Perspective API Documentation. https://developers.perspectiveapi.com/s/about-the-api
- Hutto, C. & Gilbert, E. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis." ICWSM 2014.
- SAGE Publications, "Incivility or Invalidity? Evaluating Perspective API Scores as a Proxy for Toxicity," 2024.