

# Pipeline Maverick Project 2 Report

## Cross-Platform Toxicity Dynamics in Geopolitical Content

Tripti Ramesh Anchan

Department of Computer Science, SUNY Binghamton

Binghamton, New York, USA

tanchan@binghamton.edu

### Abstract

Political conversations online look very different depending on where they happen. Anonymous platforms like 4chan move quickly and allow people to say almost anything without consequences, while Reddit's pseudonyms, voting system, and moderation slow down conversations and keep them more regulated. Building on the ingestion pipeline developed in previously, this project extends the analysis by applying Google's Perspective API to measure toxicity in collected posts.

My results show clear differences: 4chan is consistently more toxic (mean=0.346) than Reddit (mean=0.236), with significantly higher rates of identity attacks, severe toxicity, and explicit threats. When major geopolitical events occur, 4chan often reacts before Reddit, suggesting that anonymity allows discussions to escalate more rapidly. High toxicity posts mostly use identity focused terms (like "jew", "muslim") showing that the most hostile conversations on 4chan often revolve around targeting specific groups rather than just arguing about politics in general. Reddit's moderation rules and having usernames seem to discourage this kind of language, which helps keep overall toxicity lower and slows down how quickly conversations escalate. Together, these findings show how platform design fundamentally shapes the tone, speed, and hostility of online political conversations. Anonymity and minimal moderation create conditions for fast moving, highly toxic exchanges, while platform features that enforce accountability reduce the intensity and spread of hostile narratives.

## 1 Introduction

Political conversations online are shaped not only by who participates, but by the design of the platforms where those conversations take place. 4chan, for example, is an anonymous imageboard where users post anonymously, threads disappear quickly, and very little moderation stands in the way of extreme or fast-moving discussions. Reddit on the other hand, is a large, community organized discussion site built around persistent usernames, subreddit specific rules, and voting systems that decide which posts rise or fade. These structural differences create entirely different communication environments, meaning that the same geopolitical event can produce very different conversations depending on whether it unfolds on 4chan or on Reddit.

The previous project built the cross-platform pipeline that made this work possible, giving us a way to continuously collect posts, comments, timestamps, and other metadata from multiple 4chan boards and Reddit communities centered on global politics. That initial analysis revealed early differences in how the two platforms talk about geopolitical events, especially in sentiment patterns and the pace of discussion. But it also raised deeper questions that the

original pipeline couldn't yet answer questions about how hostility emerges, how narratives shift when conversations heat up, and whether the two platforms react in sync during major events.

The proposal laid out the next steps of adding automated toxicity scoring, aligning conversations in time, and examining how keywords and narratives change as toxicity increases. Building on that direction, this project takes a closer look at how toxic discourse forms, escalates and differs across platforms. Three research questions guide this analysis:

- (1) **RQ1:** Do toxicity spikes happen at the same time on both platforms, or does one tend to react sooner when major geopolitical events occur?
- (2) **RQ2:** What kinds of language, keywords, and narrative patterns show up during high-toxicity periods, and how do they differ from moments when conversations are relatively calm?
- (3) **RQ3:** How do platform design choices like anonymity versus pseudonymity and ephemeral versus persistent content shape the overall amount and distribution of toxicity?

By combining toxicity scores, sentiment information, and keyword patterns in the collected posts, this project goes further than the descriptive comparisons in the previous project. The goal is to understand how each platform's design shapes the way conversations build, which narratives take hold, and how quickly discussions turn hostile as major geopolitical events unfold.

## 2 Background and Related Work

### 2.1 Platform Architecture and Social Behavior

Different platforms create different social dynamics. 4chan is built around full anonymity and fast-moving threads that disappear quickly, which encourages unfiltered and low accountability expression [1]. Reddit, meanwhile, uses pseudonymous accounts, community moderation, and voting systems that shape what users see and reward more norm aligned behavior [2]. These structural contrasts affect how political conversations unfold, forming the basis for expecting higher toxicity on anonymous, lightly moderated platforms.

The "online disinhibition effect" [4] explains part of this: fewer social cues and no persistent identity tend to increase hostile expression, while pseudonymity introduces enough accountability to moderate behavior. This conceptual framework motivates comparing the two platforms.

### 2.2 Measuring Toxicity

Google's Perspective API provides standardized scores for multiple forms of hostile language, including toxicity, severe toxicity, identity

attacks, insults, profanity, and threat [5]. It has been used to study conflict on Wikipedia [6], political aggression on Twitter [7], and moderation effects on Reddit communities [3]. Using the same scoring model across platforms allows for direct comparison of toxicity distributions in 4chan and Reddit discussions.

### 2.3 Narrative Change and Hostility

Prior work shows that toxic or high-conflict environments often rely on identity-targeted language and emotionally charged narratives [8]. Spikes in negative affect can also accompany emerging conspiratorial frames or rapid narrative shifts [9]. Keyword-based methods and collocation patterns are commonly used to detect these changes. This project applies these ideas by comparing high-toxicity and low-toxicity posts to understand how narratives shift as discussions escalate.

## 3 Dataset

This project builds on the data collected previously, which used a custom pipeline to track political discussions across both 4chan and Reddit. For 4chan, the pipeline collected posts from /pol/, /int/, and /news/, while Reddit data came from r/politics, r/worldnews, and r/geopolitics. These communities were chosen because they consistently discuss global events and provide a natural cross-platform comparison.

- 4chan boards: /pol/, /int/, /news/
- Reddit subreddits: r/politics, r/worldnews, r/geopolitics

To keep the topics aligned across platforms, the pipeline filtered for major geopolitical keywords such as “ukraine,” “gaza,” “israel,” “china,” and “election,” following the approach outlined in the proposal. In total, the system collected several hundred thousand posts, but only a subset could be processed through Google’s Perspective API due to rate limits. This yielded 34,721 toxicity-scored posts from 4chan and 13,504 from Reddit, giving a combined analysis set of 48,225 items. Each entry includes the post text, timestamps, thread information, and platform-specific metadata.

This toxicity-scored subset forms the basis for all analyses in the report.

### 3.1 Dataset Summary

Table 1 summarizes the dataset.

**Table 1: Dataset Summary Statistics**

	4chan	Reddit
Total items collected	816,854	33,695
Toxicity-scored items	34,721	13,504
Collection period	210 days	48 days
Mean toxicity	0.346	0.236
Boards/Subreddits	3	3

### 3.2 Data Collection Architecture

The dataset comes from the cross-platform pipeline built previously, which continuously collected political discussions from both 4chan and Reddit.

- The **4chan crawler** uses the platform’s public JSON endpoints to capture new threads every minute before they disappear, storing each post with its timestamp, thread ID, and ephemeral user label.
- The **Reddit crawler** uses undocumented .json endpoints to avoid OAuth overhead; recursively collects comment trees.
- All collected data is stored in TimescaleDB, which handles time-based partitioning and helps keep the ingestion process efficient. Duplicate entries are removed using composite keys based on platform-specific metadata.

## 3.3 Toxicity Annotation

For this project, the pipeline was extended with automated toxicity scoring. Each post selected for analysis was sent to Google’s Perspective API, which returns scores for several types of hostile language, including toxicity, identity attacks, and threats. Because the API processes only one request per second, I could score only a portion of the full dataset—but the resulting 48,225 posts still provide a large and reliable sample for comparing how hostility appears and spreads across the two platforms.

## 4 Methodology

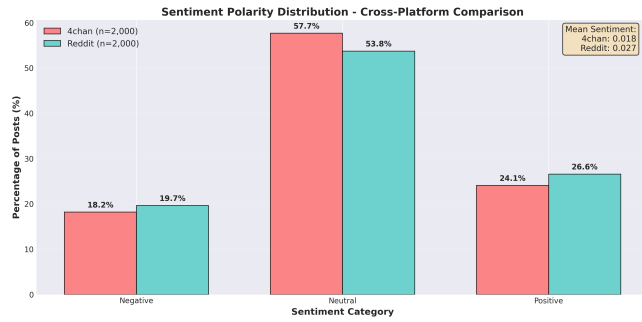
To compare toxicity patterns across platforms, I applied five analysis steps to the 48,225 toxicity-scored posts:

- **Sentiment Analysis.** Using TextBlob polarity scores, I sampled 2,000 posts from each platform and classified them as positive, neutral, or negative to compare emotional tone with toxicity levels.
- **Toxicity Distribution Analysis.** Perspective API toxicity scores were summarized using overlapping histograms and cumulative distribution functions (CDFs) to compare how toxicity is distributed across 4chan and Reddit.
- **Keyword Frequency Comparison.** Posts were split into high-toxicity ( $> 0.35$ ) and low-toxicity ( $\leq 0.35$ ) groups. For each group, I computed normalized frequencies (per 100 posts) for geopolitical terms, identity labels, and conflict related keywords.
- **Multi-Attribute Toxicity Profiling.** For all six Perspective API attributes general toxicity, severe toxicity, identity attack, insult, profanity, and threat, I calculated mean platform scores and 4chan-to-Reddit ratios to identify which hostile language types differ the most.
- **Temporal Alignment.** For specific keyword discussions, I identified peak keyword days on each platform and aligned post activity across a  $\pm 3$ -day window to compare reaction timing. For November 1–14, I computed daily thread counts and hourly post volumes for 4chan’s /pol/ board.

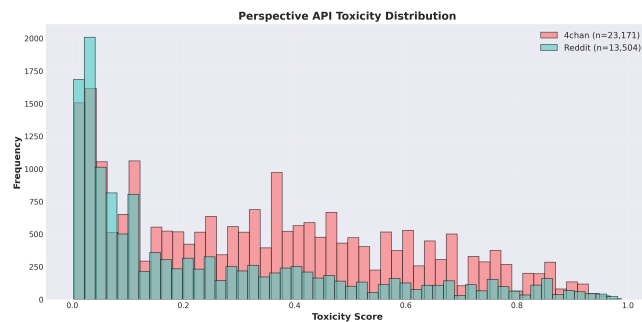
## 5 Results

### 5.1 Sentiment Polarity

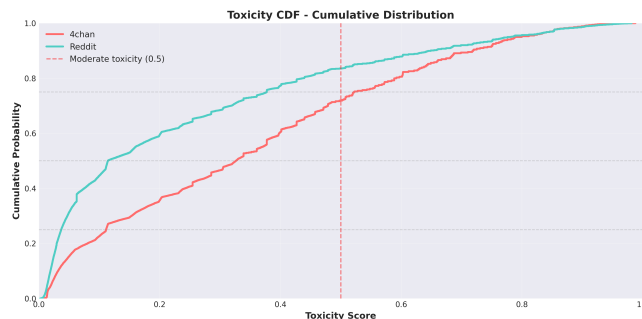
Figure 1 shows that both platforms lean heavily toward neutral sentiment, with similar proportions of positive and negative posts. This confirms that emotional tone and toxicity behave differently, i.e. a post can sound neutral and still be hostile.



**Figure 1: Sentiment polarity distribution for 4chan and Reddit.**



**Figure 2: Toxicity distribution for 4chan and Reddit.**



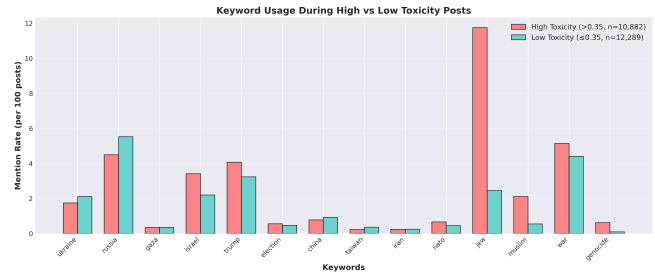
**Figure 3: CDF comparison of toxicity between platforms.**

## 5.2 Overall Toxicity Levels

Figures 2 and 3 together show how differently toxicity is distributed across platforms. Reddit's scores cluster near zero, while 4chan spreads across the entire 0–1 range with many mid- and high-toxicity posts. The CDF shows that only about 70% of 4chan posts fall below a toxicity score of 0.5, compared to roughly 95% on Reddit, highlighting how routine higher toxicity is on 4chan.

## 5.3 Language Shifts During High Toxicity

Figure 4 shows that identity focused terms (e.g., “jew”, “muslim”) increase sharply in high toxicity posts, especially on 4chan. In contrast, geopolitical terms like “russia” or “china” appear consistently



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

across both toxicity groups, suggesting they reflect topic rather than hostility.

## 5.4 Types of Hostility

Figure 5 breaks toxicity into six components. 4chan scores higher on all attributes, but the largest gap appears in identity attacks, followed by severe toxicity and threats. These results align with the keyword analysis, showing that identity-targeted rhetoric drives much of the platform difference.

## 5.5 Temporal Dynamics

Figure 6 demonstrates that 4chan reacts to geopolitical events sooner, peaking on October 20 for Ukraine and Gaza discussions, while Reddit peaks a day later. This consistent 24-hour lag suggests that anonymous platforms enable faster reactivity to breaking news.

## 5.6 4chan /pol/ data from November 1 to 14

Figures 7 and 8 show /pol/'s activity during the U.S. election period (November 1–14, 2025). Daily thread creation averaged approximately 550 threads per day, representing threads successfully captured by the collection methodology. While the crawler ran continuously with 5-minute polling intervals, my design only captured longer lived threads, which may have resulted in undersampling of short lived threads.

## 6 Discussion

The results give a clearer picture of how geopolitical conversations behave across two very different platforms. Rather than treating toxicity as a single outcome, the analysis shows how timing, language, and platform design each play a role in shaping hostile discourse. Below, I revisit the three research questions and interpret what the findings suggest.

### 6.1 RQ1: Do toxic spikes occur at the same time across platforms?

The timing analysis shows that toxicity does not peak simultaneously. Across both the Ukraine and Gaza discussions, 4chan consistently reached its activity peak about 24 hours before Reddit. This pattern makes sense given 4chan's anonymity and fast-moving threads, compared to Reddit's moderation, and longer comment chains. This lead-lag relationship suggests that activity on highly

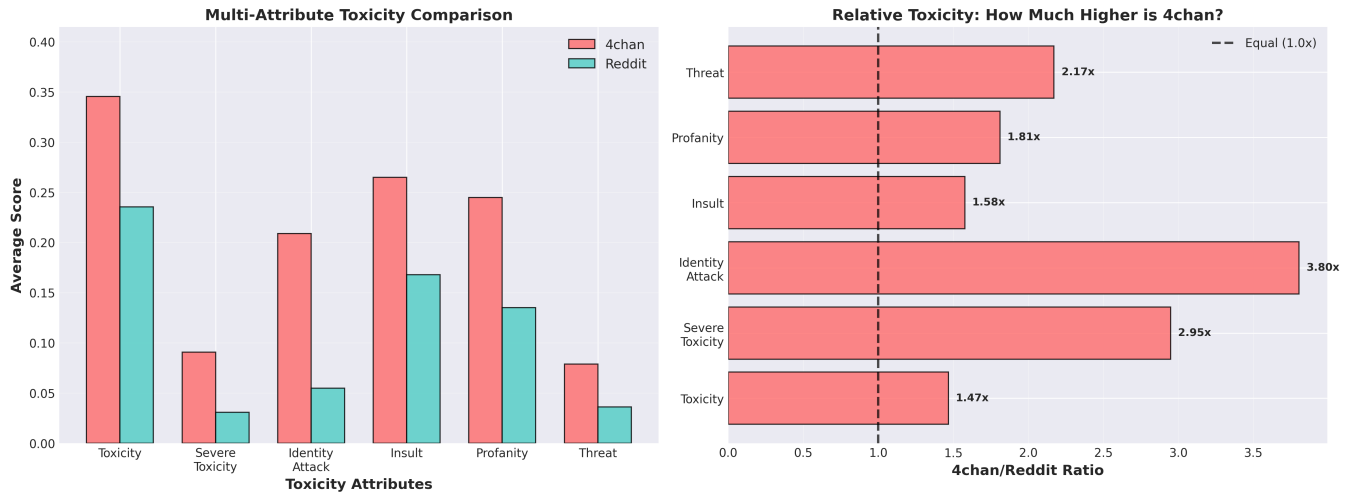


Figure 5: Multi-attribute toxicity comparison.

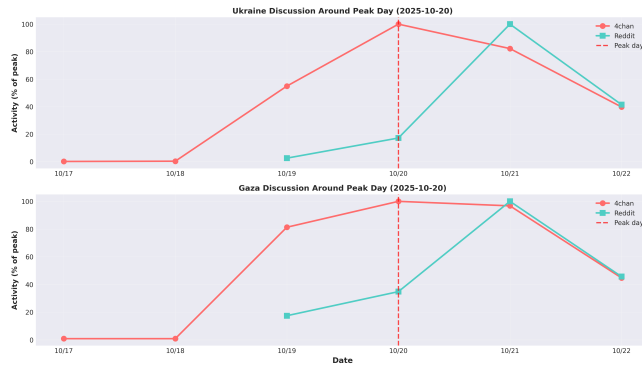


Figure 6: Ukraine and Gaza activity around peak day.

anonymous platforms may serve as an early indicator for when some narratives or hostile discussions are about to start elsewhere.

## 6.2 RQ2: What language patterns mark high toxicity?

The keyword analysis shows that the biggest shifts during toxic moments come from identity-focused language. Words like “jew” and “muslim” appear far more often in high toxicity posts than in low toxicity ones, aligning with the jump in Perspective API identity attack scores. Meanwhile, geopolitical terms such as “Ukraine,” “Russia,” and “China” remain relatively stable across toxicity levels. This suggests that toxicity on these platforms is not simply a matter of discussing topics, it’s more directly connected to how users talk about people and groups. Hostility spikes when conversations turn personal or identity-based, not just when political events escalate.

## 6.3 RQ3: How does platform design shape toxicity levels?

The differences in average toxicity across platforms are substantial: 4chan’s mean toxicity is roughly 47% higher than Reddit’s, and the gaps are even wider for severe toxicity, identity attacks, and threats. These differences match what platform design theories predict. 4chan has no limits on expression, no identity, minimal moderation, reducing both social and technical consequences for hostile behavior. Reddit, in contrast, has usernames, subreddit rules, and active moderation, which helps to reduce targeted hostility.

## 6.4 Implications

- **Platform design matters.** Anonymity and low moderation on 4chan makes it easier for hostile, identity targeted language to spread, while Reddit’s usernames and rules help reduce escalation.
- **Early warning signals.** The 24-hour lead on 4chan can suggest any spikes in toxic activity as an early indicator of what may appear next on more mainstream platforms.
- **Sentiment  $\neq$  toxicity.** Emotional tone does not indicate hostility.
- **Value for research.** Multi-platform pipelines that saves timing and context give a more accurate picture of how narratives spread and evolve across different online communities.

## 7 Limitations and Future Work

- Only a portion of the dataset could be scored because the Perspective API is slow, which limits how much toxicity data could be analyzed at once.
- The data collection window was uneven, 4chan data was collected continuously, while Reddit’s window was uneven due to json method used for collection. This gave multiple rate limits that made it slightly difficult to extract Reddit data continuously.

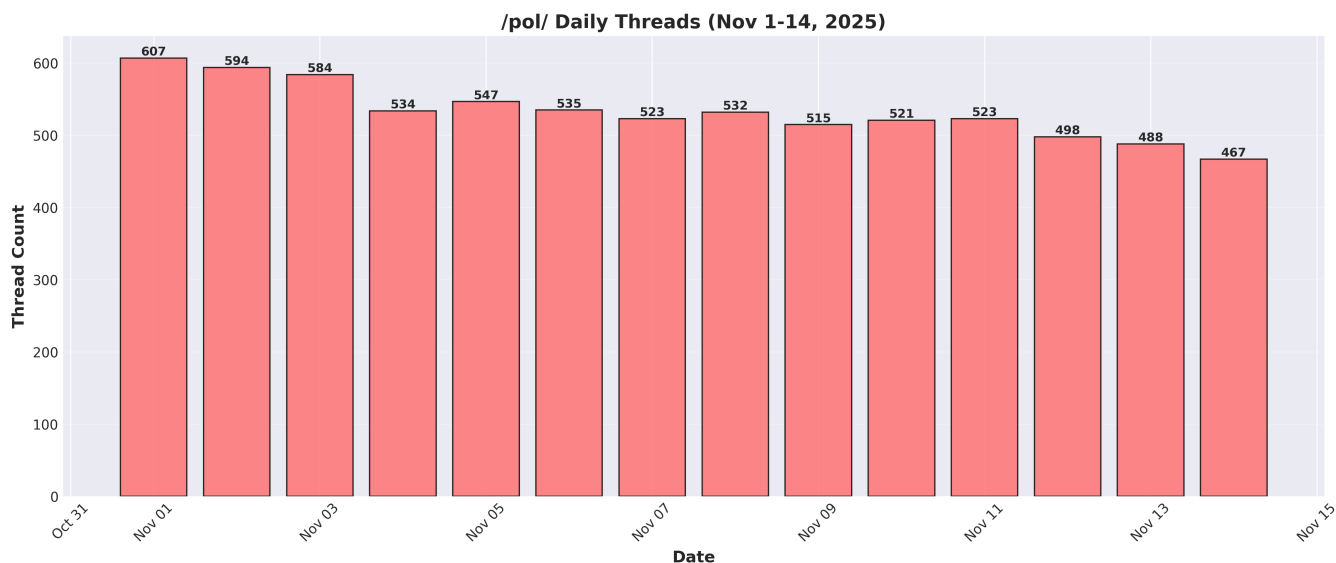


Figure 7: Daily /pol/ thread counts (Nov 1–14, 2025).

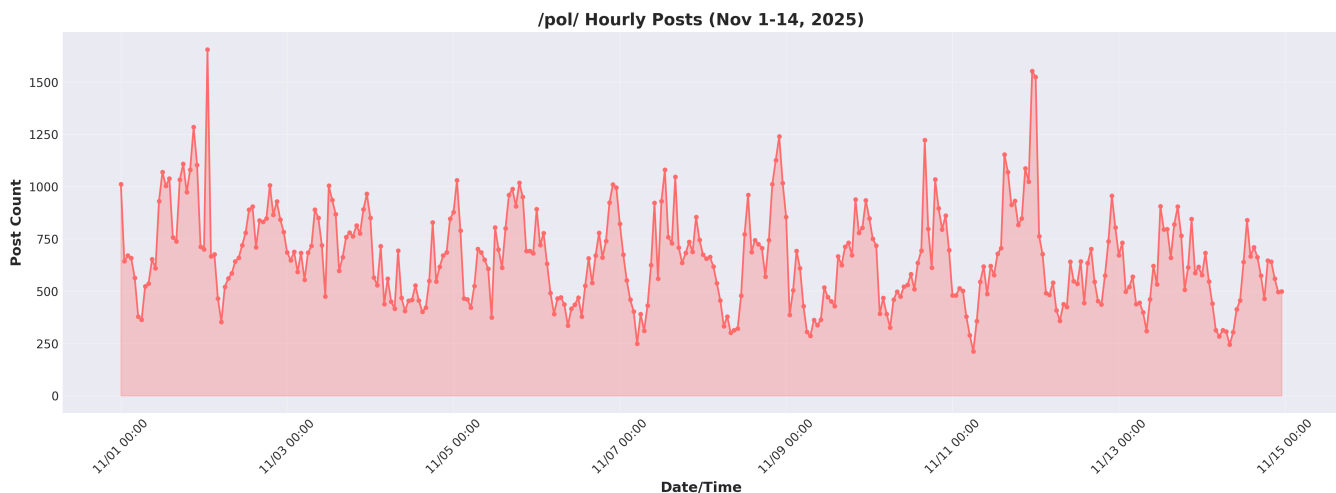


Figure 8: Hourly /pol/ post volume during Nov 1–14.

- Frequency counts capture broad patterns but miss narrative structure, sarcasm, and semantic relationships.
- Thread capture completeness: My 4chan collection methodology crawled threads when they disappeared from the active board catalog. However, threads with very short lifespans (created and archived between 5-minute polling intervals) may have been missed. This design captured the discussions effectively but may have undersampled the ephemeral content. The 550 threads per day represents successfully captured dead threads rather than the total board activity, though the sample represents major content.

Building on this project, a follow-up study could explore several new directions. Three concrete research questions for the next phase include:

- **RQ1:** What language actually marks a discussion as toxic? Using TF-IDF analysis, can we identify platform-specific toxic vocabularies beyond the basic keywords already examined?
- **RQ2:** Are the same people responsible for most toxic posts, or do regular users sometimes post toxic content too?
- **RQ3:** Do conversations get more toxic as they get longer, or does it stay consistent throughout a thread?

## 8 Conclusion

This project shows that platform design fundamentally shapes how political conversations unfold online. 4chan's anonymity enables faster, harsher discussions centered on identity-based attacks, while Reddit's moderation and usernames keep toxicity lower and slow down how quickly conversations escalate. These differences matter for understanding how hostile narratives emerge and spread during major geopolitical events.

The dataset has some limitations like API rate limits restricted toxicity scoring, and Reddit's collection window was shorter than 4chan's but the findings point toward concrete next steps. Future work can build on this foundation by identifying the specific language that distinguishes toxic discussions, examining whether toxicity comes from repeat offenders or occasional users, and tracking how conversations escalate over the course of a thread. Together, these extensions would provide a deeper understanding of the toxic political discourse online.

## References

- [1] Michael S. Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. 2011. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM '11)*, 50–57.
- [2] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), Article 31, 1–22. <https://doi.org/10.1145/3134666>
- [3] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), Article 32, 1–25. <https://doi.org/10.1145/3274301>
- [4] John Suler. 2004. The Online Disinhibition Effect. *CyberPsychology & Behavior* 7, 3 (2004), 321–326. <https://doi.org/10.1089/1094931041291295>
- [5] Jigsaw and Google. 2017. Perspective API. Retrieved from <https://www.perspectiveapi.com/>
- [6] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, 1391–1399. <https://doi.org/10.1145/3038912.3052591>
- [7] Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020. Characterizing Twitter Users Who Engage in Adversarial Interactions against Political Candidates. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, 1–13. <https://doi.org/10.1145/3313831.3376548>
- [8] Alice Marwick and Rebecca Lewis. 2018. Media Manipulation and Disinformation Online. Data & Society Research Institute. Retrieved from <https://datasociety.net/library/media-manipulation-and-disinfo-online/>
- [9] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira Jr. 2020. Auditing Radicalization Pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT\* '20)*. Association for Computing Machinery, 131–141. <https://doi.org/10.1145/3351095.3372879>