SocialMediaAvengers Project 2 Proposal

Karthik Maganahalli Prakash
Binghamton University
New York, USA
kmaganahalli@binghamton.edu

Sanjana Shivanand
Binghamton University
New York, USA
sshivanand@binghamton.edu

Abstract

Building on Project 1, which established a continuous Reddit–4chan data pipeline, this project proposes analytical methods to study global online discourse. The research examines how communities respond to major policy events, evaluates the quality of user-generated data for machine-learning applications, and analyzes public sentiment toward emerging AI technologies. Using the Perspective API for toxicity scoring and Python-based sentiment models, the study will quantify tone, engagement, and temporal trends across /pol/, /int/, and r/geopolitics, as well as AI-focused subreddits. The resulting analyses aim to transform raw social-media streams into interpretable metrics that reveal cross-platform patterns and support scalable, data-driven modeling in future work.

ACM Reference Format:
Karthik Maganahalli Prakash and Sanjana Shivanand. 2018. SocialMediaAvengers Project 2 Proposal. In . ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Understanding online discourse is essential for studying how information spreads and how communities differ in tone, sentiment, and engagement. Building on Project 1, which developed a continuous Reddit–4chan data pipeline that stores timestamped posts in a PostgreSQL-based TimescaleDB, this project focuses on analyzing global discussions, community behavior, and dataset quality. The goal is to measure how online communities react to major policy and technology events, assess the reliability of user-generated data for machine-learning use, and explore sentiment toward emerging AI systems.

1.1 Motivation

This project moves beyond data collection to understand what people discuss and how they express opinions online. Reddit and 4chan communities differ in tone, topic, and behavior, revealing broader trends in public attitudes toward technology and policy. Studying these differences helps evaluate whether user-generated content can train machine-learning models for sentiment analysis, trend detection, or content moderation. By analyzing text quality and diversity, the project links social-media data to the development of reliable models for studying online discourse.

1.2 Analytical Goals

This project aims to:

- Analyze post activity, sentiment, and toxicity using the Google Perspective API.
- Compare how people discuss AI and policy topics on Reddit and 4chan.
- Check the quality and reliability of the data to see if it can be used for machine-learning research.

2 Research Focus

This study explores three key areas using the continuous Reddit–4chan data stream.

Policy Discussions: Examine how /pol/, /int/, and r/geopolitics react to major policy events (e.g., U.S. immigration reform, EU tech regulation) and whether one region's discussions influence others. Events will be analyzed within ±7 and ±30 day windows, tagged by time, topic, and region.

Dataset Quality: Test if Reddit and 4chan posts are suitable for machine-learning models by checking text clarity, consistency, and toxicity after cleaning.

AI Sentiment: Study how users from different regions respond to new AI technologies (e.g., AutoGPT, large models) through sentiment, toxicity, and location-based analysis.

3 Proposed Methodology

Building on the Reddit–4chan data pipeline from Project 1, this project applies a structured workflow of extraction, transformation, and analysis to study policy reactions, dataset quality, and AI sentiment. The Google Perspective API is used to generate toxicity metrics for each post.

3.1 Extraction

Posts and comments are continuously gathered using existing crawlers via HTTP APIs. Sources include /pol/, /int/, and /g/ on 4chan, and r/geopolitics, r/technology, r/ArtificialIntelligence, and r/AutoGPT on Reddit. Each record stores timestamps and, for 4chan, location tags from post metadata showing the user's country flag. Data are stored in TimescaleDB for efficient temporal and regional analysis.

3.2 Transformation

Text is cleaned by removing HTML, URLs, emojis, and duplicates, with timestamps normalized to UTC. Non-English posts are filtered out, and keywords (e.g., *immigration*, *regulation*, *AI*) are used for tagging. The Perspective API adds TOXICITY, INSULT, and THREAT scores, enabling cross-platform sentiment and behavior comparison.

Conference'17, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

3.3 Analysis

Daily snapshots are exported from TimescaleDB for offline analysis using pandas. Statistical and text-based methods measure post volume, sentiment, and toxicity across platforms and time. Temporarily, trends reveal reactions to major policies, regional activity changes, and opinions on AI. Readability and diversity checks are used to assess dataset quality for future machine-learning use.

4 Plots and Tables to Produce

The following visualizations will capture cross-platform activity, sentiment, and data quality in line with the project's analytical goals. All plots will be generated in Python using matplotlib and pandas; no spreadsheet software will be used.

- **Policy Event Discussion Timeline:** X-axis – Date (±7 and ±30 days around each policy announcement); Y-axis – Number of posts per platform. Shows how discussion volume changes on /pol/, /int/, and r/geopolitics before and after major policy events.
- **Cross-Regional Activity Comparison:** X-axis – Date; Y-axis – Post volume by inferred region (US, EU, India). Detects cross-country effects, such as one region's policy triggering discussion spikes in others.
- **Keyword Frequency Over Time:** X-axis – Date; Y-axis – Mentions of policy or AI-related terms ("immigration," "regulation," "AI," "AutoGPT," "agent"). Reveals dominant topics and their temporal overlap across communities.
- **Average Toxicity by Platform:** X-axis – Board/Subreddit; Y-axis – Mean Perspective API TOXICITY score. Compares tone and civility across Reddit and 4chan.
- **Sentiment Distribution for AI Discussions:** X-axis – Sentiment polarity (negative→positive); Y-axis – Post frequency. Illustrates optimism or anxiety toward emerging AI technologies.
- **Post Length vs. Toxicity:** X-axis – Text length (characters); Y-axis – Mean toxicity score. Examines whether shorter posts are more toxic.
- **Dataset Readability and Diversity Summary (Table):** Columns – Platform, mean sentence length, type–token ratio, mean toxicity. Summarizes dataset representativeness and ML readiness.
- **/pol/ Activity Snapshot:** (a) Daily post counts from November 1 to November 14 2025; (b) Hourly post counts for the same period. Fulfills the required temporal visualization of 4chan activity.

5 Libraries Used

This project uses open-source Python libraries to support data extraction, cleaning, visualization, and analysis.

- **requests**, **httpx** – Access Reddit and 4chan APIs.
- **pandas**, **numpy** – Clean, transform, and aggregate data.
- **matplotlib**, **seaborn** – Visualize policy, sentiment, and toxicity trends.
- **sqlalchemy**, **psycopg2** – Connect to PostgreSQL/TimescaleDB.
- **googleapiclient** – Integrate Google Perspective API for toxicity scoring.

6 Data Availability and Validation

The dataset is continuously collected from selected 4chan boards and Reddit subreddits as part of the ongoing pipeline. All records undergoes deduplication using the unique_name identifier, text normalization, and timestamp alignment to ensure cross-platform consistency. A language filter retains English posts, while non-English entries will be tagged for potential secondary analysis. Event tagging will link posts to policy announcement windows (±7 or ±30 days) and AI-related keyword matches to support temporal comparisons. Planned validation plots include post volume by platform to confirm continuous collection, word-count distributions to assess text completeness, and language composition charts to verify dataset balance. These steps will produce a clean and representative dataset suitable for analyzing cross-regional policy reactions, AI sentiment, and the reliability of user-generated data for machine-learning applications.

7 Expected Outcomes and Next Steps

The analyses are expected to reveal measurable differences in discourse patterns, sentiment, and data quality across Reddit and 4chan communities.

- **Cross-platform differences:** 4chan boards (/pol/, /int/, /g/) are expected to show higher mean toxicity than Reddit due to weaker moderation and anonymous posting.
- **Policy-event response:** Posting activity is expected to spike around major policy announcements (e.g., U.S. immigration reform, EU tech regulations) as users react and discuss, showing cross-country ripple effects in timelines.
- **AI discussions:** Sentiment trends are expected to fluctuate around major AI releases, since Reddit communities tend to emphasize innovation and optimism, while 4chan boards (/pol/, /int/) often frame such topics through skepticism or cultural commentary.
- **Dataset quality:** The cleaned dataset should meet readability and coherence benchmarks through filtering, deduplication, and normalization, confirming its reliability for training and evaluating machine-learning models.
- **Key visual outputs:** Policy timelines, regional activity plots, sentiment distributions, and mean-toxicity charts will show when and how discussions shift, validating cross-platform trends and data consistency.
- **Next steps:** Extend the analytical framework to predictive modeling of toxicity and sentiment in Project 3, using ethical methods and excluding any personal identifiers.

References

[1] A. Smith, J. Zhang, and L. Wang, "Analyzing Cross-Platform AI Discussions on Reddit and 4chan," *arXiv preprint* arXiv:2502.18513, 2025. Available: https://arxiv.org/abs/2502.18513.

[2] K. Lee and M. Roberts, "Reddit-Based Sentiment and Trend Analysis for Online Communities," *arXiv preprint* arXiv:2404.17607, 2024. Available: https://arxiv.org/html/2404.17607v1.

[3] G. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn, "Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web," *Proceedings of ICWSM*, 2017.

[4] Perspective API Documentation, "Using Machine Learning to Understand Toxicity in Online Conversations," Google Jigsaw, 2024. Available: https://www.perspectiveapi.com.