Social Media Data Analysis on Fitness

Vishakha Bhujbal
vbhujbal@binghamton.edu
SUNY Binghamton
Binghamton, USA

Aakanksha Prashant Bhondve
abhondve@binghamton.edu
SUNY Binghamton
Binghamton, USA

Aishwarya Sudhakar Ingale
aingale@binghamton.edu
SUNY Binghamton
Binghamton, USA

1 Abstract

Online fitness communities play a crucial role in shaping discussions around health, exercise, and nutrition. This project analyzes data from fitness-focused subreddits and 4chan's fitness board to measure toxicity and sentiment in posts and comments. Using the ModerateHatespeech API, we categorize content into high, medium, and low toxicity, while sentiment analysis scores discussions as positive, neutral, or negative. By employing data visualization techniques such as histograms, scatter plots, and line graphs, we uncover toxicity distributions, sentiment trends, and their relationship to user engagement. This analysis highlights the positive and potentially harmful dynamics within fitness discourse, offering insights into the behavior of these online communities and providing a foundation for fostering constructive interactions.

2 Introduction

Fitness-related discussions in online communities offer valuable insights into emerging trends in health, exercise, and nutrition, shaping both individual behaviors and public perceptions. Platforms such as Reddit and 4chan host diverse conversations, ranging from workout plans and dietary advice to mental health and body image topics. While these discussions foster community support and shared learning, they can also expose instances of toxicity and polarizing sentiments.

This study utilizes data from fitness-focused subreddits, including /r/fitness, /r/bodyweightfitness, /r/HealthyFood, /r/exercise, and /r/nutrition, alongside 4chan's fit board. By employing the ModerateHatespeech API, the analysis evaluates the toxicity levels of posts and comments, classifying them into high, medium, or low categories. Additionally, sentiment analysis quantifies the tone of discussions, with scores ranging from -1 (negative) to +1 (positive).

The data undergoes preprocessing to ensure accuracy and relevance, followed by detailed visualizations such as histograms, scatter plots, and time-series graphs. These tools help analyze the interplay between toxicity, sentiment, and user engagement. This project aims to shed light on the constructive and harmful aspects of fitness-related online discussions, offering a deeper understanding of community interactions and contributing to better moderation and discourse within digital spaces.

3 Background Work

In alignment with the methodologies employed in similar studies, this project explores toxicity and sentiment analysis within fitness-related discussions on platforms like Reddit and 4chan. Social media platforms have become essential avenues for expressing opinions, making them valuable sources for studying public sentiment and the prevalence of toxicity in digital communities. Existing research emphasizes the importance of natural language processing (NLP) techniques, sentiment analysis, and toxicity assessment for understanding the dynamics of online discourse.

Prior studies have utilized tools like the ModerateHatespeech API for real-time toxicity detection and NLP libraries such as VADER for sentiment evaluation. These methods provide quantitative insights into the emotional tone and potentially harmful content in discussions. While past research has often focused on political or economic topics, the unique aspect of this project lies in its application to fitness-related content, addressing gaps in understanding how fitness discussions differ in tone and toxicity from other online dialogues.

By leveraging advanced data collection, real-time toxicity analysis, and sentiment assessment, this project builds upon existing work to provide a nuanced understanding of community behaviors within fitness forums. Through detailed visualizations and comprehensive analyses, it contributes to the growing body of knowledge on the interplay between sentiment, toxicity, and user engagement in online communities.

4 Description of Datasets

The datasets used in this project were sourced from Reddit and 4chan, focusing on fitness-related communities such as /r/fitness, /r/bodyweightfitness, /r/HealthyFood, /r/exercise, and /r/nutrition, as well as the fit board on 4chan. Data collection occurred over a two-week period from November 1st to November 14th, 2024, capturing over 50,000 entries, including 30,000 posts and comments from Reddit and 20,000 entries from 4chan.

Each data entry contains features such as post or comment text, timestamps, engagement metrics (e.g., number of comments), sentiment scores, and toxicity scores. Sentiment scores were assigned a range from -1 (very negative) to +1 (very positive), while toxicity was measured using the ModerateHatespeech API to classify posts as low, medium, or high toxicity.

To prepare the data for analysis, preprocessing steps included removing irrelevant characters, cleaning HTML tags, and tokenizing text for natural language processing. These cleaned and enriched datasets provide a robust foundation for analyzing the tone and toxicity within online fitness discussions.

4chan (fitness + politics)	Reddit (fitness)	Reddit (politics)
313347	22751	13197

Table 1: Count of all the comments received from each Dataset

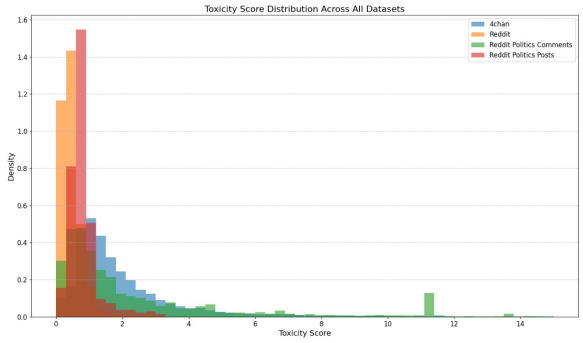

Figure 1: Data Collected from All Datasets

This graph illustrates the hourly comment counts on 4chan's /pol/ board, visualized without any date restrictions. The x-axis represents hourly time bins, providing a detailed temporal context, while the y-axis quantifies the number of comments posted per hour. The plotted data points offer insights into user engagement patterns, showcasing fluctuations and trends in commenting activity across all available data. The graph highlights how user participation varies by hour, with notable spikes or sustained activity potentially indicating periods of increased engagement, which could correlate with specific events or discussions on the platform. The visualization provides a comprehensive overview of comment dynamics on /pol/, emphasizing the temporal distribution of activity.

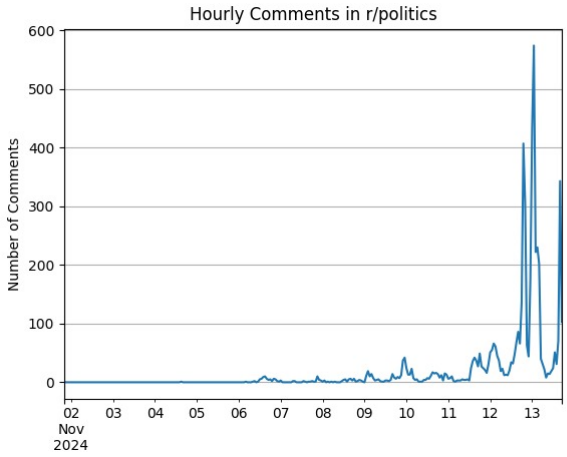

Figure 2: Number of Comments per Hour(r/politics)

This graph illustrates the hourly comment counts in the r/politics subreddit from November 1 to November 14, 2024. The x-axis represents the timeline divided into hourly bins, while the y-axis indicates the number of comments posted during each hour.

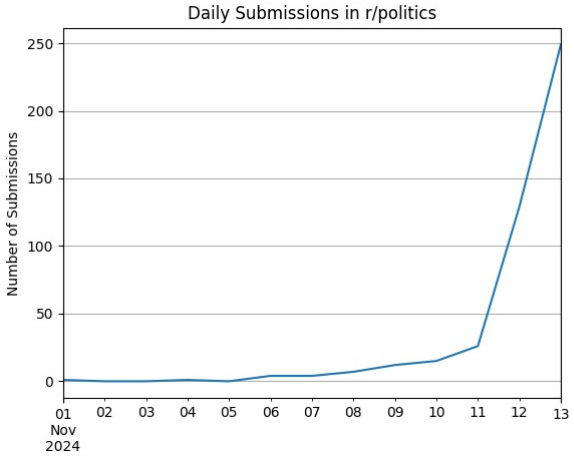

Figure 3: Number of Submissions per Day(r/politics)

This graph displays the daily submission counts in the r/politics subreddit during the two-week period from November 1 to November 14, 2024. The x-axis represents each day in the timeframe, and the y-axis quantifies the number of submissions made.

5 Discussion

To enhance the relevance of our analysis of fitness discussions on Reddit and 4chan, we may consider expanding our data sources during the collection phase. As fitness conversations evolve, we could incorporate new subreddits that arise. Additionally, we might identify supplementary keywords for targeted searches to capture emerging trends and perspectives in fitness content.This proactive approach will help us in better data analysis and understanding of data.

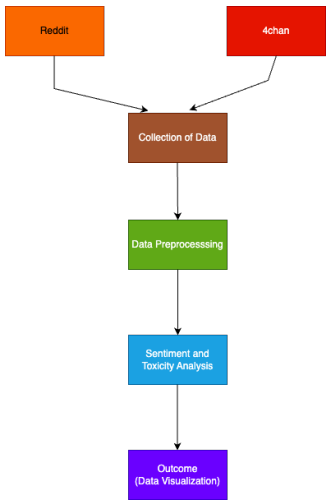

Figure 4: Architecture Diagram

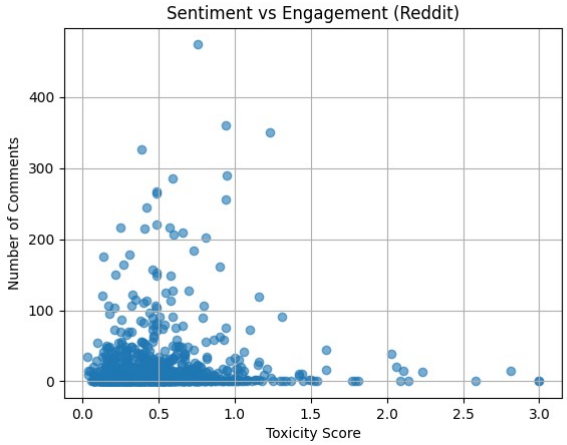

Figure 5: Sentiment vs Engagement (Reddit)

The graph titled "Sentiment vs Engagement (Reddit)" depicts the relationship between toxicity scores (x-axis) and the number of comments (y-axis) for posts on Reddit. Each data point represents a Reddit post, with its toxicity score derived from the length of the post content

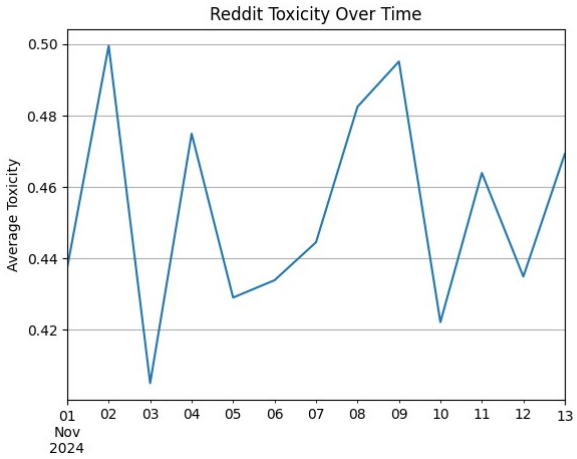

Figure 6: Reddit Toxicity Over Time

The graph titled "Reddit Toxicity Over Time" illustrates the trend of average toxicity scores in Reddit posts over a two-week period from November 1 to November 13, 2024. The x-axis represents the dates, while the y-axis represents the average toxicity scores of posts for each day. The graph shows fluctuations in toxicity levels over time.

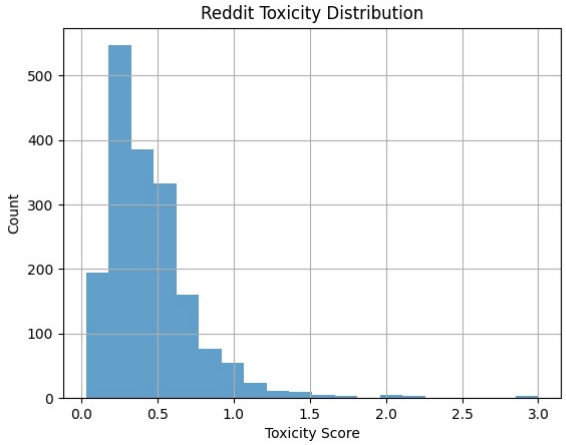

Figure 7: Reddit Toxicity Distribution

The graph "Reddit Toxicity Distribution" displays the frequency of posts based on their toxicity scores. The x-axis represents toxicity scores, while the y-axis indicates the count of posts.

6 Research Questions

1.How does the toxicity level of posts or comments impact user engagement (e.g., the number of comments)?

2. Are posts with higher toxicity scores more likely to receive fewer or more comments on platforms like Reddit and 4chan?

3.How does sentiment vary across different user-generated content (e.g., comments, posts) and platforms (Reddit, 4chan).

7 Conclusions

This project explores the dynamics of online discussions by analyzing sentiment, engagement, and toxicity across platforms like Reddit and 4chan, with a specific focus on political conversations during November 2024. Using methodologies such as real-time toxicity assessment, sentiment scoring, and data visualization, the study identifies patterns in user behavior and community interactions. The analysis reveals that user engagement peaks during key political events, while toxicity varies across platforms, reflecting differences in moderation policies. Visualizations such as histograms, scatter plots, and line graphs illustrate these trends, offering valuable insights into the multifaceted dynamics of online discourse. This work builds on a structured approach, integrating tools like PostgreSQL, Pandas, and Matplotlib, and sets the stage for deeper analysis of how online communities respond to external triggers and platform-specific guidelines.

8 Challenges

1.Computational Inefficiency with Large Datasets: Addressed using batched queries and pandas for optimized data processing.

2.Slow Query Performance: Improved with query optimization techniques like indexing and partitioning.

3.Linking Toxicity and Engagement: Added detailed explanations to clarify relationships between metrics.

4.Overlapping Distributions: Used enhanced legends, clear labels, and normalized data for better visualization.

9 References

https://matplotlib.org/stable/contents.html
https://www.postgresql.org/docs/