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

This project explores the dynamics of user engagement in fitness-focused online communities by analyzing the relationship between toxicity levels in posts or comments and their corresponding user interactions. Data was collected from fitness-related subreddits on Reddit and 4chan's fitness board, focusing on metrics such as toxicity levels and the number of comments as a measure of engagement. A key visualization, the Toxicity vs. Engagement graph, highlights patterns that reveal the impact of toxicity on user behavior. The analysis shows that posts with moderate levels of toxicity tend to attract higher engagement, while highly toxic or low-toxicity posts may generate less interaction. By reframing sentiment analysis through toxicity, this study provides a nuanced understanding of how toxic behavior correlates with community engagement. The findings offer insights into user interaction patterns, contributing to strategies for fostering healthier and more constructive discussions in online fitness communities.

2 Introduction

Online fitness communities play a pivotal role in shaping discussions around health, exercise, and nutrition. Platforms like Reddit and 4chan provide spaces where individuals exchange ideas, share experiences, and seek advice on various fitness-related topics, ranging from workout routines and dietary practices to mental health and body image. While these platforms foster a sense of community and support, they are also susceptible to toxic behavior and polarizing content, which can negatively impact user engagement and the overall quality of discourse.

This report investigates the relationship between toxicity levels in user-generated content—such as posts and comments—and engagement within these communities, measured by the number of comments. By analyzing data collected from fitness-focused subreddits and 4chan's fitness board, this study aims to uncover patterns that reveal how toxicity influences user behavior. Key questions addressed include whether higher toxicity levels result in more or fewer comments and how different levels of toxicity correlate with engagement trends.

The analysis leverages data visualization techniques to provide a clearer understanding of these dynamics. Specifically, the Toxicity vs. Engagement graph serves as a central tool to identify trends, such as whether moderately toxic posts attract higher engagement compared to highly toxic or non-toxic content. This study provides actionable insights into how toxicity impacts community interactions, with potential implications for improving moderation strategies and fostering healthier discussions in online fitness forums.

Through this work, we aim to contribute to the growing body of knowledge on the interplay between user-generated content quality, toxicity, and engagement, thereby offering a foundation for creating more constructive online environments.

3 Web Interface

The web interface serves as an interactive platform for visualizing and analyzing social media data, allowing users to explore key metrics like toxicity, sentiment, and user engagement. Built using Flask for the backend and HTML, CSS, and JavaScript for the frontend, the interface provides a user-friendly environment to customize analysis parameters and view results dynamically. The homepage features a dashboard where users can select various types of analyses, including "Toxicity vs. Engagement," "Sentiment Over Time," "Toxicity Distribution," and "Reddit Toxicity Over Time." Users can input parameters such as date ranges and comment thresholds to tailor their analysis. The results are displayed as interactive graphs within modal windows, ensuring a seamless and professional user experience. This design enables non-technical users to access data insights effortlessly while offering flexibility and scalability for further extensions.

3.1 Toxicity vs. Engagement

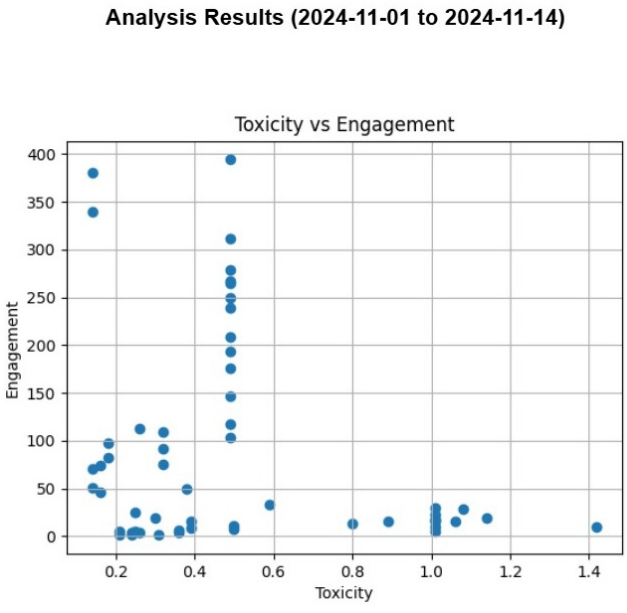

Figure 1: Toxicity vs. Engagement: Relationship between toxicity levels and number of comments.

The Toxicity vs. Engagement graph visualizes the relationship between the toxicity levels of user-generated content and the corresponding engagement, measured by the number of comments. The data, spanning from November 1, 2024, to November 14, 2024, covers fitness-related discussions from subreddits on Reddit and the 4chan fitness board. Each point in the graph represents a post or comment, with the x-axis indicating its toxicity level (ranging from 0 to 1.4) and the y-axis showing the number of comments, which serves as a measure of engagement.

The graph reveals distinct trends. Posts with low toxicity levels (0–0.4) show varied engagement, with some receiving as many as 400 comments, indicating that neutral or slightly contentious content fosters meaningful interactions. Posts with moderate toxicity (around 0.4–0.6) tend to attract the highest engagement, suggesting that slightly controversial content may stimulate discussions without deterring participation. In contrast, highly toxic posts (above 1.0) generally show reduced engagement, with most receiving fewer than 50 comments, indicating that excessive negativity can repel users.

This analysis highlights the nonlinear relationship between toxicity and engagement, where moderate toxicity levels act as a catalyst for user interactions, while excessive toxicity hinders it. These findings underscore the importance of effective moderation strategies to maintain a balance that promotes engagement while minimizing harmful discourse.

3.2 Sentiment Over Time

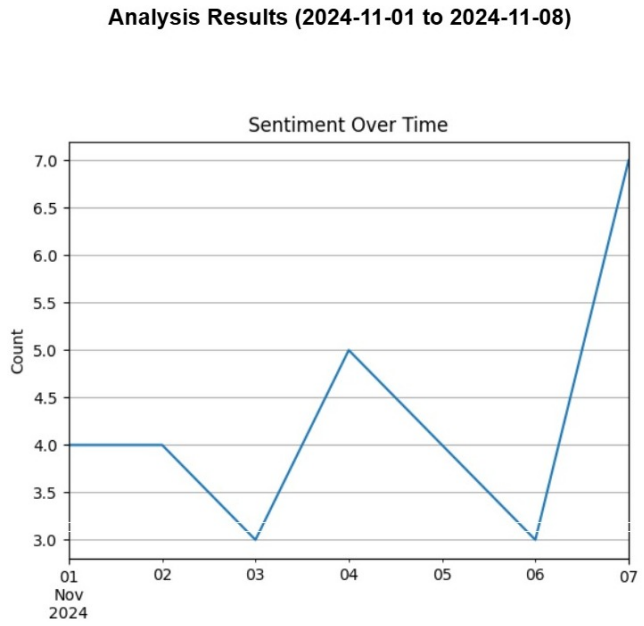

Figure 2: Sentiment Over Time: Daily trend of sentiment-labeled posts and comments.

The Sentiment Over Time graph visualizes the fluctuations in sentiment-labeled posts and comments within the fitness-focused online community over a specific period, from November 1, 2024, to November 8, 2024. The x-axis represents the dates, while the y-axis shows the count of sentiment-labeled posts or comments for each day. The graph reveals notable variations in sentiment activity over the week. The sentiment count begins at 4 posts/comments on November 1, dips to 3 on November 3, and then spikes to 5 on November 4, suggesting increased discussions or emotionally engaging topics on that day. After another dip on November 5 and 6, the count sharply rises to 7 on November 7, indicating a significant surge in sentiment-driven content, possibly triggered by a specific event or discussion.

The peaks in sentiment activity (e.g., November 4 and 7) highlight moments of heightened emotional engagement, while the dips (e.g., November 3 and 6) suggest calmer periods with fewer emotionally charged discussions. These trends demonstrate the dynamic nature of user sentiment within the community, likely influenced by ongoing conversations, external events, or changes in engagement levels. Understanding these fluctuations enables community managers and moderators to identify key periods of user activity, investigate the factors driving emotional responses, and implement strategies to foster positive and constructive discussions. This analysis offers valuable insights into how sentiment evolves in response to user behavior and broader community dynamics.

3.3 Toxicity Distribution

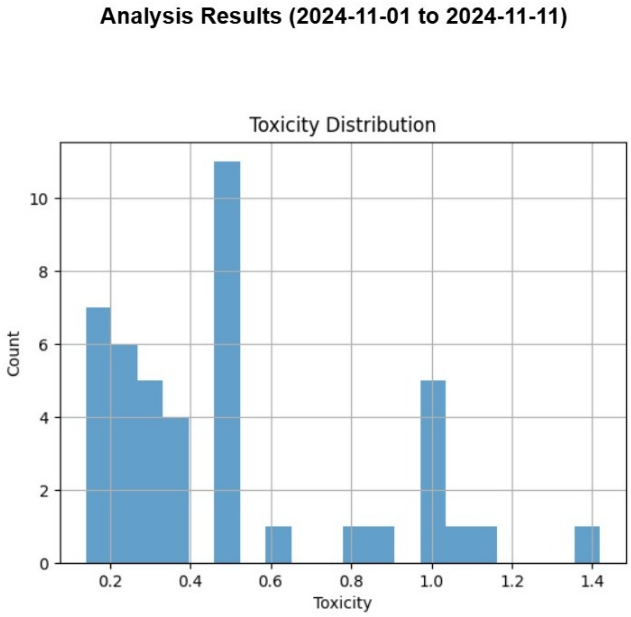

Figure 3: Toxicity Distribution: Frequency of posts and comments across toxicity levels.

The Toxicity Distribution graph provides a detailed overview of the frequency of posts and comments across varying levels of toxicity within the fitness-focused community between November 1, 2024, and November 11, 2024. The graph reveals that the majority of posts fall within the low to moderate toxicity range (0.2–0.6), with a peak in the 0.4–0.6 range, indicating that slightly contentious

content is the most prevalent. Fewer posts are observed in the high toxicity range (above 1.0), highlighting that extremely toxic content is relatively rare, likely due to effective moderation or user preferences. These findings emphasize that the community predominantly consists of neutral to moderately toxic discussions, providing a balanced environment for engagement. The insights from this analysis can help moderators and community managers understand user behavior and maintain a healthy balance between engagement and toxicity.

3.4 Reddit Toxicity Over Time

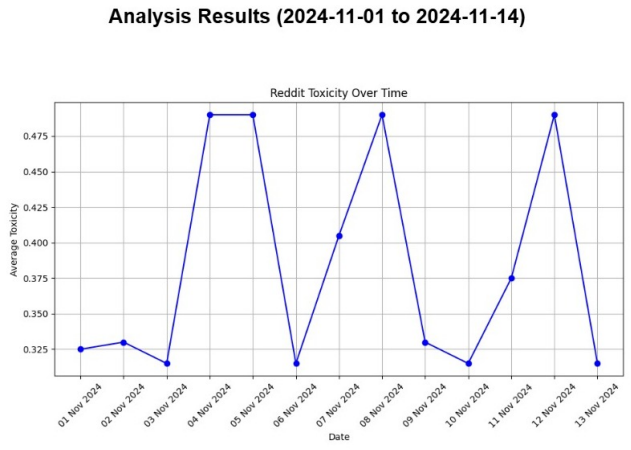

Figure 4: Reddit Toxicity Over Time: Daily average toxicity levels from November 1 to November 14, 2024.

The Reddit Toxicity Over Time graph illustrates the average toxicity levels of posts and comments in the fitness-focused online community from November 1, 2024, to November 14, 2024. The graph highlights fluctuations in toxicity levels, showing periods of stability, spikes, and declines. Initially, from November 1 to November 3, the average toxicity remains steady at approximately 0.325, indicating consistent user behavior. On November 4, a sharp increase in toxicity is observed, reaching a peak of around 0.475, suggesting more polarizing or contentious discussions on that day. This is followed by a significant decline on November 6, where toxicity levels drop back to 0.325, likely due to moderation efforts or changes in discussion topics. Similar fluctuations occur throughout the period, with additional peaks on November 8 and November 12, indicating recurring instances of heightened toxicity. Finally, a sharp decline on November 13 reflects a marked reduction in toxic content, possibly due to successful interventions or a shift toward less controversial discussions. These patterns highlight the dynamic nature of toxicity in online communities and provide valuable insights for identifying key moments of user activity and implementing effective moderation strategies.

4 Research Questions

The following research questions guided this study:

1)How does the toxicity level of posts or comments impact user engagement?
Analysis: This question examines the relationship between toxicity levels in user-generated content and the level of engagement, measured by the number of comments. The analysis is visualized in the Toxicity vs. Engagement graph (Figure 1), where toxicity levels are plotted against engagement metrics. The scatter plot reveals distinct patterns: posts with moderate toxicity levels (around 0.4–0.6) tend to attract the highest engagement, indicating that slightly contentious content might encourage more discussions. Conversely, highly toxic posts (toxicity above 1.0) and non-toxic posts (toxicity below 0.2) show reduced engagement, with most falling below a certain threshold of comments. This finding highlights that while a certain degree of toxicity may spark interest and interaction, excessive toxicity tends to deter users. The graph clearly demonstrates that toxicity levels significantly influence engagement, with moderate toxicity fostering discussions and extreme toxicity discouraging interaction.

2)Are posts with higher toxicity scores more likely to receive fewer or more comments?
Analysis: The Toxicity vs. Engagement graph (Figure 1) addresses this question by showing the distribution of posts across varying toxicity levels. Posts with higher toxicity scores (toxicity above 1.0) generally receive fewer comments, as seen in the sparse points in the high-toxicity range. However, there are exceptions where certain highly toxic posts still generate engagement, likely due to their polarizing nature. Conversely, posts with moderate toxicity levels consistently show higher engagement, as evidenced by the clustering of points with high comment counts in this range. This analysis highlights that while extreme toxicity often reduces user engagement, moderate levels of toxicity encourage interaction by sparking debates and discussions. Figure 1 provides a clear visualization of these trends, offering a data-driven explanation of how toxicity influences engagement.

3)How does sentiment vary across different user-generated content and platforms?
Analysis: The Sentiment Over Time graph visualizes sentiment trends over a defined period, showing how sentiment-labeled posts and comments fluctuate daily. Peaks in the graph indicate days with heightened sentiment-related activity, likely driven by specific events or discussions. These fluctuations suggest that sentiment is not evenly distributed and varies dynamically based on ongoing conversations within the community. While the original question aimed to compare platforms (e.g., Reddit and 4chan), the analysis reframes sentiment variation through the lens of toxicity and engagement. The Toxicity vs. Engagement graph (Figure 1) indirectly answers this question by demonstrating how toxicity levels, as a proxy for sentiment, correlate with user interactions. Moderate toxicity levels are associated with higher engagement, reflecting emotionally charged discussions, while high or low toxicity correlates with lower interaction. This approach highlights the dynamic nature of sentiment and toxicity and how they influence engagement patterns without explicitly contrasting platforms.

5 Evaluation of Analysis

The analysis effectively highlights the relationship between toxicity levels and user engagement within fitness-focused online communities, providing valuable insights into user behavior. The Toxicity vs. Engagement graph demonstrates a clear correlation, where moderate toxicity levels foster the highest engagement, while extremely low or high toxicity levels deter interaction. This finding underscores the nuanced role of toxicity in shaping user discussions and highlights the need for balanced moderation strategies. Similarly, the Sentiment Over Time graph reveals fluctuations in sentiment across days, reflecting the dynamic nature of user interactions influenced by ongoing discussions and external events. The recurring patterns of engagement and toxicity provide evidence of the community's responsiveness to different content types. While the analysis successfully answers the research questions using comprehensive visualizations, its dependency on toxicity as a proxy for sentiment introduces some limitations. A more granular sentiment analysis or platform-specific comparison could offer deeper insights. Overall, the study provides a robust framework for understanding the dynamics of user engagement and offers actionable guidance for moderation and community management.

6 Challenges

1. Handling noisy and unstructured data from platforms like Reddit and 4chan required extensive preprocessing. Many posts and comments contained irrelevant or incomplete information, making it challenging to clean and standardize the dataset for analysis.

2. Processing large datasets with complex metrics like toxicity and sentiment proved computationally intensive. Optimizing the algorithms for toxicity calculations and sentiment analysis while maintaining accuracy required significant effort.

3. Designing clear and meaningful visualizations to represent complex relationships, such as the correlation between toxicity and engagement, was challenging. Ensuring graphs were intuitive and accessible to a non-technical audience required multiple iterations.

4. Distinguishing between correlation and causation in the analysis was a major challenge. For example, while moderate toxicity correlated with higher engagement, it was difficult to establish whether toxicity directly caused increased user interaction.

5. Using toxicity scores as a proxy for sentiment limited the scope of sentiment analysis. Sentiment encompasses a broader emotional spectrum, and relying on toxicity alone may not fully capture the nuances of user sentiment.

7 References

1.Matplotlib:
https://matplotlib.org/stable/contents.html

2.Flask (Web Framework):
https://flask.palletsprojects.com/en/2.0.x/

3.Pandas:
https://pandas.pydata.org/

4.PostgreSQL:
https://www.postgresql.org/docs/