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

Fitness-focused online communities are key platforms for exchanging health advice, sharing experiences, and building support networks. However, these spaces often contend with issues like toxicity and divisive sentiments that can affect user participation. This project explores the interplay between sentiment, toxicity, and engagement on platforms such as Reddit and 4chan. Through predictive modeling, clustering methods, and temporal analysis, the study will identify patterns influencing user behavior and interaction. Predictive models will assess how sentiment and toxicity impact engagement, while clustering will reveal distinct user and topic groups. Temporal trends will highlight the evolution of sentiment and toxicity over time, offering valuable insights into community dynamics. The outcomes will include actionable insights and interactive visualizations designed to support platform administrators in promoting healthier and more constructive discussions within fitness communities.

2 Introduction

Online fitness communities serve as dynamic platforms for discussing health, exercise, and nutrition. These platforms, including Reddit and 4chan, host diverse conversations ranging from workout tips and diet plans to mental health and body image discussions. While these discussions foster community support and shared learning, they also encounter challenges such as toxicity and polarizing sentiments, which can influence user participation and the overall quality of discourse.

This project focuses on analyzing fitness-related discussions to uncover key behavioral patterns and engagement dynamics. By examining the interplay between toxicity, sentiment, and user engagement, the study aims to identify factors that drive or hinder productive interactions. Advanced techniques such as predictive modeling and clustering will be employed to segment user behavior and analyze trends over time. The goal is to provide actionable insights for fostering constructive and meaningful discussions within these communities.

3 Libraries

The following tools, libraries, and frameworks will be used to implement the project:

Backend:

Flask: A lightweight Python framework to build the web-based interactive tool.

Frontend:

HTML/CSS: For structuring and styling the dashboard.

JavaScript: To add interactivity to visualizations, using libraries like D3.js if necessary.

Database: PostgreSQL: For querying and managing the data.

Data Analysis: Pandas: For data cleaning, manipulation, and preprocessing.

NumPy: For efficient numerical computations.

Visualization:

Matplotlib: For generating static visualizations.

Seaborn: For advanced and aesthetically pleasing plots.

Plotly: For dynamic, interactive visualizations.

Deployment: Hosted locally on a university-provided virtual machine, accessible via the university VPN.

4 Web Framework

For this project, the chosen web framework is Flask, a lightweight and flexible Python framework that is well-suited for building interactive web applications. Flask will serve as the backbone of the interactive dashboard, enabling the dynamic querying and visualization of data collected from fitness and politics-related discussions on Reddit and 4chan. Its modular design and simplicity make it ideal for integrating backend data processing with a responsive user interface. Using Flask, the dashboard will facilitate seamless user interactions, allowing users to explore toxicity levels, sentiment trends, and their correlation with engagement metrics. Flask's routing capabilities will handle user requests, process parameters like date ranges or analysis types, and return tailored results, such as scatter plots, histograms, or time-series graphs. This ensures that users can interactively analyze fitness and politics-related data in real time.

5 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).

6 Analysis

1.Toxicity vs. Engagement Analysis This analysis examines how the toxicity levels of posts and comments influence user engagement, measured as the number of comments. Using the ModerateHatespeech API, toxicity levels will be categorized as low, medium, or high. A scatter plot will visualize the relationship between toxicity scores (x-axis) and engagement metrics (y-axis). This analysis will provide insights into whether posts with higher toxicity generate more discussion or discourage user interaction.

2.Sentiment Trends Over Time The sentiment analysis will track the evolution of positive, neutral, and negative sentiments in fitness-related discussions over time. Using time-series visualizations, this analysis will uncover patterns in sentiment changes, revealing how

the tone of discussions fluctuates daily or weekly. This can help understand whether external factors, such as fitness trends or events, influence the sentiment in online conversations.

3.Toxicity Distribution This analysis investigates the overall distribution of toxicity levels within the collected dataset. By using histograms, the frequency of posts categorized into low, medium, and high toxicity will be displayed. This analysis will provide an overview of the general tone of fitness-related discussions, highlighting the prevalence of constructive versus harmful content in the data.