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 Introduction

Online discussions in fitness communities reveal evolving trends in health, exercise, and nutrition, influencing both public perception and individual behavior. This project builds on our initial data collection from fitness-focused subreddits like /r/fitness, /r/bodyweightfitness, /r/HealthyFood, /r/exercise, and /r/nutrition by adding a toxicity analysis layer using the ModerateHatespeech API. Through real-time measurement and analysis of toxicity and sentiment, we aim to identify key fitness topics and the overall tone of these discussions, offering valuable insights into the positive and potentially harmful aspects of online fitness communities.

2 Proposed Methodology

Data Preprocessing

We will clean the collected data using standard NLP preprocessing techniques to ensure accuracy and relevance in our analysis. This process includes removing special characters, punctuation, and irrelevant symbols that may interfere with accurate text analysis. Additionally, we will tokenize the content, breaking down posts and comments into individual words or structured units, which prepares the text for further processing and analysis steps.This preprocessing step will yield a refined dataset, allowing for accurate real-time toxicity measurement and more reliable trend analysis across platforms.

Toxicity Analysis

We will measure the toxicity level of each post and comment in real time using the ModerateHatespeech API, with a focus on fitness-related keywords. This API provides toxicity scores that allow us to categorize content based on toxicity levels, making it possible to identify high, medium, and low-toxicity posts and comments instantly. Given that toxicity can vary across specific topics like extreme diets and body image, we will analyze these scores by category within each subreddit. This approach will help pinpoint particularly sensitive or contentious topics in fitness communities, creating a clear distribution of toxicity across different areas of discussion.

Sentiment Analysis

Sentiment analysis will involve evaluating each post and comment to determine its overall tone, assigning a sentiment score between -1 (very negative) and +1 (very positive). This scoring approach will capture the sentiment expressed across fitness discussions, identifying whether topics like workout routines, diet plans, or body image discussions carry predominantly positive, negative, or neutral tones. Once each post and comment has been scored, these values will be aggregated to calculate the average sentiment for each specific topic and subreddit. This aggregation will reveal general sentiment trends within the community, showing which topics and subreddits tend towards positive or negative sentiment. By examining these trends, we can gain insights into the underlying tone of fitness discussions, supporting an understanding of community dynamics and user engagement across fitness-focused platforms.

3 Data Visualizations and Insights

We propose the following visualizations to analyze toxicity and sentiment across fitness-related discussions:

Histogram of Toxicity Score Distribution:
This visualization will show the distribution of toxicity scores across various fitness topics on each subreddit, with the X-axis representing toxicity scores and the Y-axis showing the number of posts/comments. This will help us understand the prevalence of different toxicity levels within specific discussions.

Scatter Plot of Sentiment Analysis by Subreddit:
To examine the relationship between toxicity and user engagement, this scatter plot will display toxicity scores against engagement counts (e.g., number of comments), using the X-axis for toxicity scores and the Y-axis for engagement count. This will help reveal any correlation between toxicity levels and the extent of user interaction.

Line Graph of Topic Toxicity Over Time:
This visualization will track toxicity scores over time for each topic on each subreddit. The X-axis will represent time, while the Y-axis will show average toxicity scores, highlighting any fluctuations in toxicity across topics over time.

4 Data Scaling Strategy

This proposal provides a structured plan for measuring and analyzing fitness-related data with an emphasis on toxicity. By leveraging real-time toxicity assessment and structured measurement, this project aims to deliver meaningful insights into the behaviors, trends, and sentiment dynamics within fitness communities. These findings will serve as a foundation for further analysis in Project 3, contributing to a nuanced understanding of online fitness discourse.

5 Libraries

Requests - To interact with Reddit and ModerateHatespeech APIs.
Pandas - For efficient data manipulation and structuring.
NLTK - For NLP preprocessing, tokenization, and text normalization.
Matplotlib and Seaborn - For data visualization, enabling customizable and clear visuals.

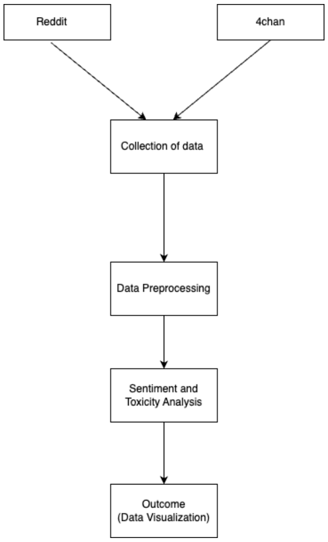

Figure 1: System Architecture

6 Data Source Enhancement

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.