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

Abstract

Fitness and health trends shape public behavior, wellness industries, and consumer habits. Our project leverages social media data from platforms like Reddit and 4chan to analyze evolving fitness trends, workout routines, and dietary habits. Using APIs, we integrate, filter, and analyze data based on fitness-related keywords to identify popular practices and emerging trends. These insights will deepen our understanding of public fitness behaviors, supporting informed decisions in the health and wellness sectors.

1 Data Source

We plan to collect data from Reddit's active fitness-related communities, focusing on subreddits where users discuss workout routines, diet plans, and general fitness advice. Data will be collected using the Reddit API. Our targeted subreddits include /r/fitness, /r/bodyweightfitness, /r/HealthyFood, /r/nutrition, and /r/exercise. These subreddits provide discussions ranging from workout strategies to nutrition and healthy eating, offering us a rich dataset for our project.

We will specifically query these subreddits for posts and comments containing keywords such as "workout," "diet," "calories," "muscle building," and "fat loss." The collected data will be parsed and organized into a structured data frame for further analysis. This data will be used to identify trends and insights in fitness, diet, and exercise discussions.

API: https://www.reddit.com/dev/api/

We plan to collect data from 4chan's /fit/ board, an active space where users discuss fitness routines, nutrition, and share fitness-related memes and advice. Using the 4chan API, we will dynamically extract threads and posts, focusing on key fitness topics such as "routine," "gains," "protein," "strength training," and "cutting." This data will be structured and organized to identify recurring discussions, popular fitness challenges, and meme-based advice circulating within the community.

The 4chan API returns data in JSON format, allowing for seamless integration into our data pipeline. The collected posts will be stored in a data frame, where we will continuously loop through and update content based on relevance and frequency, ensuring real-time analysis of fitness trends and discussions from the /fit/ board.

API: https://a.4cdn.org/

2 Data Collection

For our data collection process, we will develop two crawlers: one for 4chan and one for Reddit. Using a custom client class to manage API requests, we will extract fitness-related posts based on keywords such as "workout," "strength training," "nutrition," and "fat loss." The collected data will be stored in a PostgreSQL database, ensuring structured management and efficient querying. A set of positive and negative sentiment keywords will also be applied to analyze posts, allowing us to evaluate the balance of positive versus negative discussions within the fitness community. This approach ensures scalable and organized data collection for deeper analysis.

3 Data Analysis

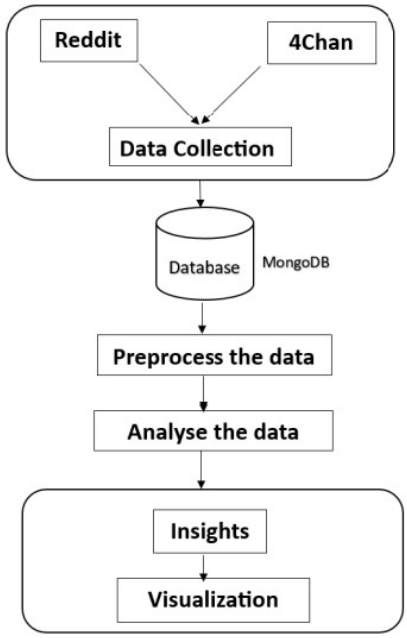

Figure 1: Figure 1: System Architecture

We will categorize data using fitness-related keywords such as "HIIT," "strength training," "cardio," "bulking," "cutting," and "calisthenics." This categorization will structure conversations and posts about specific fitness trends, allowing for targeted analysis. Additionally, statistical analysis will be performed to identify patterns in the frequency of discussions, user engagement, and the correlation between different fitness trends. Metrics such as post counts, comment activity, and sentiment distribution will help uncover the most popular fitness routines and emerging trends within these communities.

4 Napkin Math Estimates

For our data collection from the 4chan /fit/ board, we estimate a daily data volume of approximately 900 KB, resulting in a weekly total of about 6.3 MB. In terms of data entries, we project around 750 rows will be collected each day, leading to a cumulative total of

approximately 5,250 rows over the course of a week. In comparison, our data collection from the Reddit fitness subreddits is expected to yield around 6,402 KB daily, culminating in a weekly total of approximately 44.81 MB. We anticipate collecting around 6,060 rows per day from these subreddits, resulting in a total of approximately 42,420 rows collected weekly. These estimates provide us with a clear understanding of the expected data volume and entry counts for our fitness project.