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.

Analysis reveals that discussions surrounding extreme diets and body image are associated with higher toxicity levels, offering insights into contentious fitness topics

INTRODUCTION

Our project analyzes fitness and health discussions on online platforms such as Reddit and 4chan, with a focus on topics like workout routines, nutrition, and diet plans. This helps us understand the overall sentiment and tone of fitness-related discussions.

The integration of toxicity analysis allows us to not only identify popular fitness trends but also highlight controversial or toxic topics within these communities. Ultimately, this project aims to provide insights into both positive and negative behaviors in online fitness discussions, contributing to a better understanding of how these conversations evolve.

The data collected from these platforms not only reveals trends in fitness discussions but also helps stakeholders in the wellness industry understand community engagement and address potentially harmful or toxic conversations

DATA COLLECTION

Our data collection strategy uses two primary sources: Reddit and 4chan. Fitness-related discussions from these platforms provide a rich dataset for analysis.

Reddit

Reddit is a popular platform where users engage in discussions on various topics, and it provides real-time data through its API. For this project, we focused on collecting data from fitness-specific subreddits, including /r/fitness, /r/bodyweightfitness, /r/HealthyFood, /r/exercise and /r/nutrition. The data retrieved from these subreddits includes posts and comments related to workout routines, diet plans, and general health.

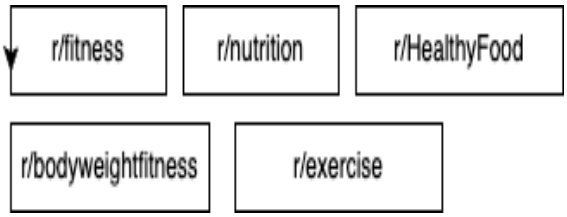

Figure 1: List of subreddits we are using to fetch data

To facilitate data storage, we created a PostgreSQL database that stores the data retrieved from the Reddit API. The process begins with sending an authentication request to Reddit to obtain an access token, which is then used to make API requests. The API endpoint https://www.reddit.com/api/v1/access_token is used to generate this token.

Once authenticated, we iterated over a predefined list of subreddits to fetch the most recent post data and their associated comments. The comments were accessed using the API endpoint https://oauth.reddit.com/r/{subreddit}/comments/{post_id}/{post_title}, which requires the subreddit name, post ID, and post title as parameters. We focused on these subreddits due to their large active user base and relevance to fitness topics, ensuring a comprehensive dataset on health discussions.

4chan

Data from 4chan's /fit/ board, dedicated to fitness discussions, was collected using the 4chan API. We extracted threads and comments on topics like workout routines and supplements and stored them in a PostgreSQL database for analysis. The **/fit/** board was selected for its dedicated fitness community, offering unique perspectives on workout routines and supplements.

Data Collection

The API functions get_threads() and get_catalog() were used to retrieve thread and post data. get_threads() fetched thread metadata

such as ID and content, while get_catalog() monitored new discussions in real-time. The data was continuously collected and stored for further analysis.

Data Preprocessing

A clean_comment() function was used to preprocess the comments by removing HTML tags, unescaping special characters, and filtering out irrelevant data. This step ensured the dataset was cleaned and structured for toxicity analysis.

4. Preliminary Results

In the preliminary analysis of the data, we examined both the frequency of fitness-related discussions and the associated toxicity levels. These results provide insights into which topics are driving the most engagement and where toxicity is more prevalent.

We observed that **/r/fitness** had the highest level of engagement, while discussions on supplements and extreme diets, especially on 4chan, showed higher toxicity scores.

Topic Analysis

- Workout Routines: Discussions around different workout routines, especially High-Intensity Interval Training (HIIT) and strength training, were the most frequently mentioned. These topics spanned across both platforms, with Reddit showing a higher volume of posts compared to 4chan.
- Diet and Nutrition: Dietary discussions were common, especially regarding specific diets like keto and intermittent fasting. However, posts related to extreme diets saw higher toxicity scores, indicating these are more controversial topics within the fitness community.

Data Collection Trend

The data collection for Reddit and 4chan was automated using Python scripts to fetch new posts and comments at regular intervals, storing the data in a PostgreSQL database.

Reddit Data Collection

The **reddit_crawler.py** script used the Reddit API to collect data (post ID, title, comments) from subreddits like **/r/fitness** and **/r/nutrition**, running every 10 minutes for continuous updates.

4chan Data Collection

The chan_crawler.py script used the 4chan API to retrieve data from the /fit/ board, with the get_catalog() function monitoring real-time discussions.

Trends Observed

- **Reddit**: Higher data volume, with spikes in discussions on workout routines and diets like intermittent fasting.
- **4chan**: Fewer posts but higher toxicity, especially on controversial topics like extreme diets and body image.

The time-series plot shows the growth in posts over time, with peaks reflecting increased discussions during trending fitness challenges.

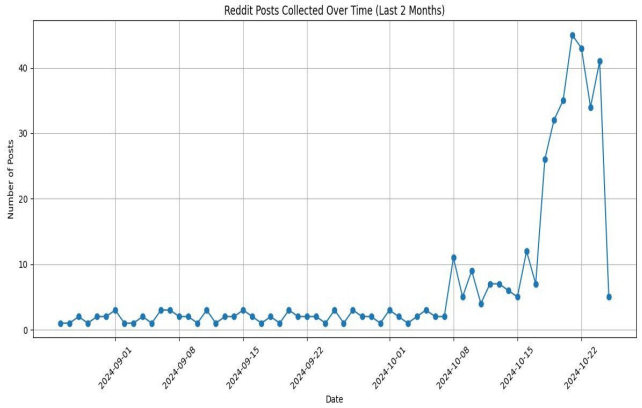

Figure 2: Chart showing Data Collection over time

ARCHITECTURE

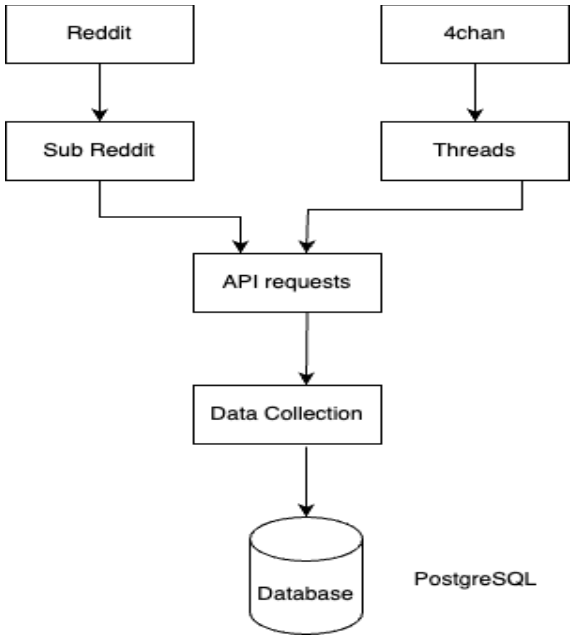

Figure 3: Architecture of Data Collection System

Figure 3 shows the overall architecture of the data collection system, including how API requests are made, processed, and stored in a PostgreSQL database for further analysis. The system is designed to handle real-time data collection, processing, and toxicity measurement**.**

CHALLENGES

Reddit: Reddit enforces API rate limits, so we added a one-second delay between requests to avoid exceeding the limit while collecting data efficiently.

4chan: The unstructured data from 4chan included slang and HTML tags, which complicated analysis. We resolved this by using a cleaning function to preprocess the data.

REFERENCES

1]**Reddit** -https://www.reddit.com/dev/api

2]**4chan-**https://github.com/4chan/4chan-API

3]**PostgreSQL-** https://www.postgresql.org/docs/