Analysis of Social Media Trends on Sports

Venkata Achyuth Kunchapu
SUNY Binghamton
Binghamton, USA
vkuncha1@binghamton.edu

Amitesh Dubey
SUNY Binghamton
Binghamton, USA
adubey1@binghamton.edu

Gowtham Surya Gunasekaran
SUNY Binghamton
Binghamton, USA
ggunase1@binghamton.edu

Shivani Harane
SUNY Binghamton
Binghamton, USA
sharane@binghamton.edu

Isha Harne
SUNY Binghamton
Binghamton, USA
iharne@binghamton.edu

ABSTRACT

Social media platforms like Reddit and 4chan serve as massive repositories of user-generated content, encouraging interactions and conversations about a wide range of topics, including sports. These platforms provide information about users' perspectives, disputes, preferences, and sports-related behaviors. However, the total volume and diversity of content makes it difficult to access, analyze, and derive significant insights from these discussions in a systematic manner. This data collection system can address this issue by using a methodical and systematic approach to gathering sports-related data from Reddit and 4chan via their respective APIs. This system's insights have the potential to inform content creation, increase user engagement, and provide strategic guidance to various entities in the sports domain, thereby addressing the prevalent issues associated with the exploration of sports-related discussions on social media platforms.

INTRODUCTION

In the realm of sports, the outburst of chants and cheers echos not only in stadiums but also across the vast universe of social media, generating an extensive variety of discussions, debates, and dialogues. APIs are like tools that facilitate communication and information sharing between various computer systems. We are able to access and use a variety of services and information that have been developed and shared on various online platforms using APIs. The Backend Daemon, a crucial component of our data pipeline, plays the important role, of managing the flawless gathering and processing of data from various different sources. As a result of this complex process, a wide range of structured and refined data has been produced that is ready to be broken down and analyzed. This has opened the door for insightful analytics that can shed light on the complex interactions and conflicts in online sports discussions and direct strategic interventions in the sports industry.

DATA SOURCES

Data will be gathered from the following sources:

- **Source:** Reddit
  - **API:** https://www.reddit.com/dev/api/
  - **Description:** A website of different communities where users can share and discuss about various topics.
- **Subreddits**
  - **Sports:** https://www.reddit.com/r/sports/
  - **Soccer:** https://www.reddit.com/r/soccer
  - **MMA:** https://www.reddit.com/r/mma
  - **Formula 1:** https://www.reddit.com/r/formula1
  - **NBA:** https://www.reddit.com/r/nba
  - **Football:** https://www.reddit.com/r/Collegefootballmemes/
  - **Boxing:** https://www.reddit.com/r/Boxing/
- **Source:** 4chan
  - **API:** https://a.4cdn.org/
  - **Description:** A website that is divided into boards dedicated to a variety of user interests and topics.
- **4chan Boards**
  - **Sports:** https://boards.4channel.org/sp/

DATA COLLECTION

Reddit API and 4chan API, two distinctive platforms illustrated as the starting points in our pipeline diagram (refer to Figure 1), are crucial for our data collection process. APIs, or Application Programming Interfaces, function as communication channels, enabling interaction between different software components, and in this scenario, they allow systematic access and retrieval of data from Reddit and 4chan platforms.

Reddit is a vast network of community forums or "subreddits," each dedicated to specific topics or themes. Via the Reddit API, shown on the left of the Daemon in Figure 1, we are enabled to systematically gather posts, comments, and other types of data, focusing particularly on sports-related content. This approach facilitates extensive analysis of user discussions and perspectives on Reddit.

Contrastingly, 4chan is segmented into various boards, each representing different interests or topics, including sports. The

4chan API, depicted below the Reddit API in Figure 1, enables the acquisition of threads and posts from these boards, thus allowing the collection of varied user opinions and debates.

The Backend Daemon, represented at the center of Figure 1, is the pivotal point in our data pipeline, serving as an integrator of data from the two diverse sources: Reddit and 4chan. It interacts with both APIs to secure sports-centric data. Following the data acquisition, the Daemon undertakes the task of organizing and processing the data before consigning it to a PostgreSQL database, as seen to the right of the Daemon in Figure 1. The choice of PostgreSQL ensures stability, scalability, and advanced data processing capabilities, thereby ensuring effective management of the amassed data.

The orchestrated and automated procedures of data collection and storage facilitated by the Backend Daemon ensure the acquisition of pertinent, high-quality sports data, paving the way for subsequent detailed analysis, as depicted in the final component of our pipeline in Figure 1.

SCHEDULING

The backend daemon is a crucial component of our system architecture, carrying out necessary functions and coordinating data flows. This daemon runs continuously on our server and is in charge of carrying out background tasks to maintain efficient and seamless system operations. We use Faktory to precisely manage and schedule these tasks. Operating on port 7420, Faktory offers features like job prioritization, retry mechanisms, and performance monitoring in addition to providing a stable framework for our daemon's operations. Most importantly, Faktory keeps thorough logs that show the number of commands executed, failed jobs, and queued jobs. These logs provide us with important information about system performance and possible areas for optimization.

DATA OVERVIEW

From Reddit, our data collection focuses on capturing a rich set of information to provide insights into user engagement and content dynamics. The following details are retrieved from each post:

- **Subreddit**: The specific community where the post originates.
- **Title**: The headline or main subject of the post.
- **Author**: The username of the individual who created the post.
- **Upvotes**: The number of positive votes the post has received.
- **Downvotes**: The number of negative votes the post has received.
- **Ratio**: The upvote to downvote ratio, indicating the overall reception of the post.
- **Number of Comments**: A count of how many comments the post has garnered.
- **Post ID**: A unique identifier for the post.
- **Comments**: The actual text content of user comments on the post.
- **Kind**: Specific category or type related to the post.

4chan Data Collection

Turning to 4chan, our data acquisition strategy is slightly different, reflective of the distinct nature of the platform. We collect:

- **Board**: The particular section or category of 4chan where the thread is located.
- **Thread Number**: A unique identifier for the thread.
- **Text**: The primary content or body of the thread.
- **Author**: The username or alias of the thread creator.
- **Comments**: The responses or replies to the main thread.
- **Title**: The subject or headline of the thread, if provided.

All of this data, from both platforms, is systematically stored in our PostgreSQL database, ensuring data integrity and ease of access for future analysis.

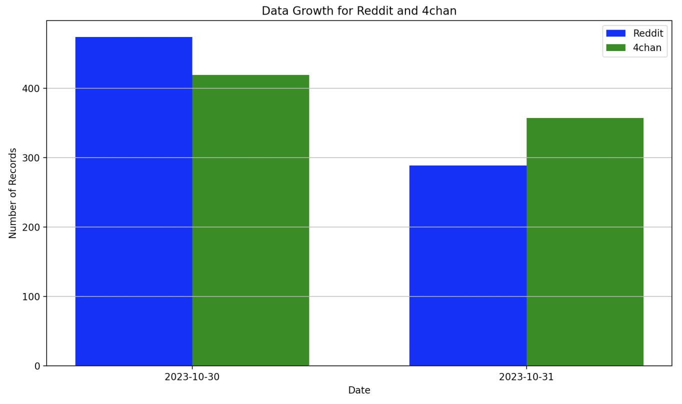

Figure 2. One Day Data Growth

QUANTITATIVE OVERVIEW OF DAILY DATA INGESTION

In a span of 24 hours, our system collected 348 records from 4chan and 289 records from Reddit. Given the average size of a Reddit post is approximately 50KB and that of a 4chan thread is about 30KB, the total data accumulation amounts to 23.1MB for Reddit and 16.7MB for 4chan each day, summing up to an approximate 39.8MB daily. Consequently, our PostgreSQL database is projected to see a monthly increase of around 1.19GB. With regards to processing, Faktory's system, which processes jobs every 10 minutes, may impose a slight overhead, utilizing a minor portion of the VM's memory. These calculations give us a basic understanding of our operational scale, but it's essential to note that real-world outcomes might slightly differ from these predictions.

Should we require more data in the future, we plan on expanding our data sources by including additional subreddits. We are confident that the storage capacity of the provided VM will suffice for our needs.

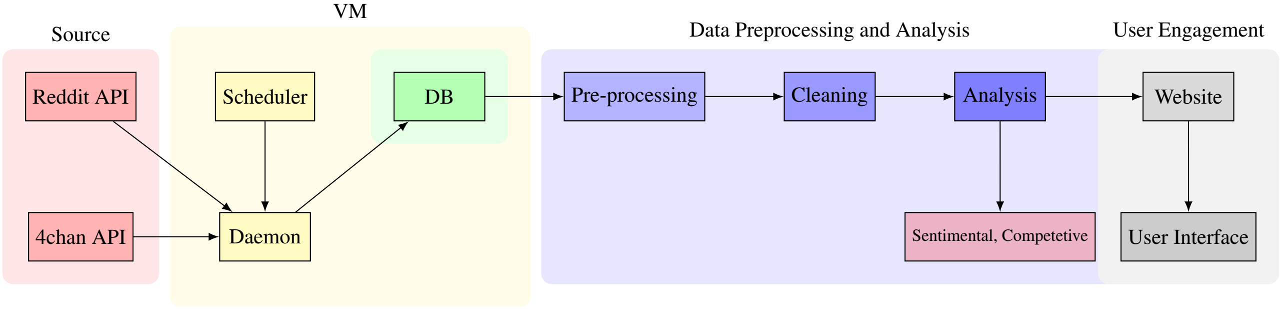

Figure 1. Data Collection and Analysis pipeline.

DATA ANALYSIS

- **Competitive Analysis:** This analysis involves comparing various sports, teams, or leagues to determine which ones have the most active and engaged communities on platforms like Reddit and 4chan.
  - **Data Collection** Collect data from the selected subreddits and 4chan board related to various sports, teams, and leagues. Ensure you have a substantial amount of data for each category.
  - **Data Pre-processing** We plan to clean and preprocess the data to remove noise, such as irrelevant posts, spam, and duplicates.After which we categorize the data into different groups, such as sports, individual teams, and leagues, based on keywords, tags, or subreddit/board categorizations.
  - **Engagement Metrics** Define engagement metrics to measure the activity and engagement within each category. Key metrics may include: Total number of posts and comments,Total upvotes and downvotes, Total number of unique users participating, Average comment length, Frequency of posts per day or week.

  In the end compare the engagement metrics across the different categories (sports, teams, and leagues). This comparison will help identify which ones have the most active and engaged communities. We plan to calculate the total number of posts and comments in each category. This will give you an idea of which categories are most active.Divide the total engagement metrics by the number of unique users participating in each category. This will help you understand the average engagement per user.Analyze the number of upvotes and comments per post in each category. This can indicate how popular and engaging the content is within each category.

- **Sentiment Analysis:** This analysis will be used for understanding how the community feels about different sports, teams, or events over the platforms like Reddit and 4chan.
  - **Data Collection** Collect data from the selected subreddits and 4chan board related to various sports, teams, and leagues. Ensure you have a substantial amount of data for each category.
  - **Data Pre-processing** It is a crucial stage in this analysis where we will be Removing special characters and symbols, Tokenizing the text into individual words or phrases,Removing stop words (common words like "and," "the," "is"), Lemmatizing or stemming words to reduce them to their base form.
  - **Sentiment Analysis** In this approach, you train a machine learning model (commonly using techniques like Natural Language Processing or Deep Learning) to predict sentiment based on labeled data. You'll need a labeled dataset with text samples and corresponding sentiment labels (e.g., positive, negative, neutral). Then apply the chosen sentiment analysis technique to the preprocessed data. Assign sentiment labels (e.g., positive, negative, neutral) to each post or comment based on their sentiment scores.

  In the end compare aggregate sentiment scores and labels to determine the overall sentiment of the community. You can calculate average sentiment scores for specific sports, teams, or events, or you can tally the number of positive, negative, and neutral comments.

FUTURE SCOPE

In the future implementation we are looking forward to clean and filter the data as per the analysis needs mentioned above and generate some insights using data visualization techniques.

CHALLENGES

During our data collection and scheduling endeavors, we encountered several challenges that necessitated tailored solutions. A prominent challenge arose when scheduling jobs using Faktory: whenever we approached the Reddit API rate limit, we would encounter errors, causing our scheduling process to halt. To address this, we incorporated exception handling in Python, allowing for more graceful error recovery and ensuring continuity in our data collection process.

In our initial stages of gathering data from 4chan, another hurdle presented itself. Instead of receiving consistent and relevant data, our system often fetched random, unrelated information each time it interfaced with the 4chan API. Our solution involved a reconsideration of our data collection logic. We refined our approach by considering the last updated date from our database, ensuring that we retrieved only the most recent and relevant threads.

Moreover, we occasionally faced connectivity challenges, particularly when attempting remote connections to our Virtual

Machine (VM). These instances, although infrequent, did disrupt our workflow. We are actively exploring more stable and reliable connection methods to mitigate such issues in the future.

REFERENCES

[1] 4chan API documentation:
https://github.com/4chan/4chan-API

[2] YouTube Data API documentation:
https://developers.google.com/youtube/v3/docs

[3] Official Reddit API documentation:
https://www.reddit.com/dev/api/

[4] Faktory Official Github Page:
https://github.com/contribsys/faktory