Project Report

Arjun Mahadkar
B00976832
amahadk1@binghamton.edu

Deepang Raval
B00924269
draval1@binghamton.edu

Rahul Verma
B00892091
rverma4@binghamton.edu

Sudeep Rawat
B00852066
srawat1@binghamton.edu

Yuraj Vartak
B00866245
yvartak1@binghamton.edu

ABSTRACT

Social media platforms are proven to be a good source of data for scientific analysis and for training various machine learning models. We present implementation of 3 data collection systems that collects real time tweets data from Twitter, comments on the posts from the specified subreddit and comments on all the videos from the specified YouTube channel. All the systems are implemented using Python programming language and designed to continuously collect the data and store it in a database.

1 IMPLEMENTATION

1.1 Twitter Tweets

(1) Access Twitter using Twitter Sample Stream API
(2) Use the keys and token generated on your developer account wtih the Twitter API
(3) We use a json request to get live data from Twitter
(4) The json is filtered with parameters which are language = 'English' and context related to 'Sports'
(5) We then push the response obtained in the database
(6) This process is continued until the program is stopped

Figure 1 displays the architecture of Twitter tweets data collection system.

1.2 Reddit Comments

(1) Add credentials for the API and names of the subreddits whose comments you want to extract in config file
(2) For each subreddit name present in config file we fetch its posts and append the post ids to a list in the config file if there is a new post
(3) Next we traverse through each post and make a request to search its comments
(4) We process the comment tree received using depth first search algorithm
(5) If the comment is of kind 't1' then we check whether or not it is already present in the list in config file. If not we add the comment id to the list in the config file and save the comment in the database
(6) If the kind is 'more' then we append the list of children ids to the more list
(7) Then we search for the comments using the ids present in the more list and the morechildren API endpoint
(8) We repeat step 4-6 for the response obtained from step 7
(9) We run the program periodically

Figure 2 displays the architecture of Reddit comments data collection system.

1.3 YouTube Comments

(1) Add credentials for the API and ids of the channels whose comments you want to extract in config file
(2) For each channel id present in config file we fetch its videos till 3 days before and append the video ids to a list in the config file if there is a new video

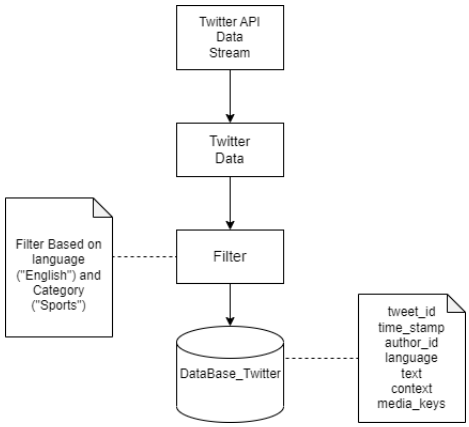

Figure 1: Twitter Data Collection System

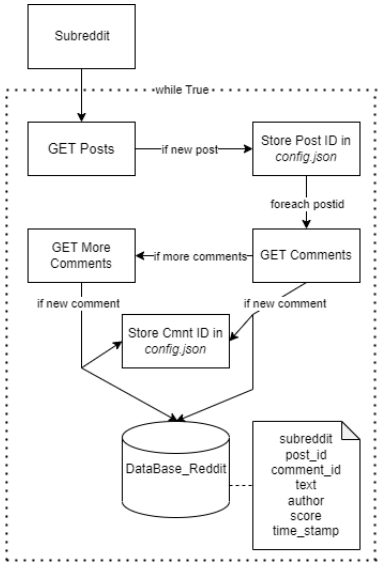

Figure 2: Reddit Data Collection System

(3) The search for videos continues till the response obtained contains 'next-page-id'. Since the response contains video ids in sorted order from newest to oldest, we can reduce the number of request we make by comparing it with the timestamp of the previous fetched video id
(4) Next we traverse through each video and make a request to search its comments

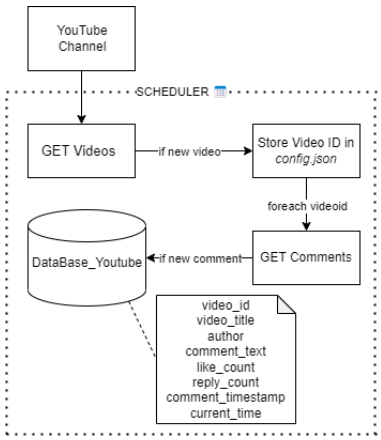

Figure 3: YouTube Data Collection System

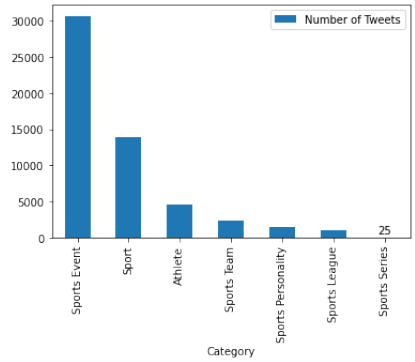

Figure 4: Category vs. Number of Tweets

(5) We store the comments and its replies if any in the database
(6) The search continues till we have 'next-page-id' and comments are also in sorted order, hence we fetch the newest comment compared by the timestamp
(7) We use the scheduler to run the program everyday

Figure 3 displays the architecture of YouTube comments data collection system.

2 PRELIMINARY DATA EXPLORATION

2.1 Twitter Tweets

(1) Figure 4 shows the total number of tweets per category between 2022/10/29 00:00:00 UTC and 2022/10/29 23:59:59 UTC
(2) Figure 5 shows the 10 authors with highest number of tweets between 2022/10/29 00:00:00 UTC and 2022/10/29 23:59:59 UTC

2.2 Reddit Comments

(1) Figure 6 shows the total number of comments per subreddit between 2022/10/29 00:00:00 UTC and 2022/10/29 23:59:59 UTC
(2) Figure 7 shows the 10 authors with highest number of comments between 2022/10/29 00:00:00 UTC and 2022/10/29 23:59:59 UTC discarding the deleted users

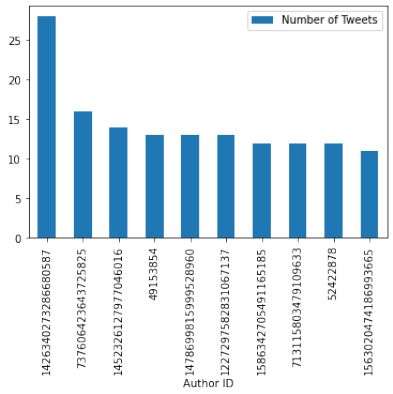

Figure 5: Author ID vs. Number of Tweets

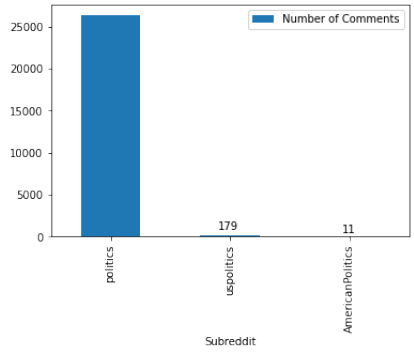

Figure 6: Subreddit vs. Number of Comments

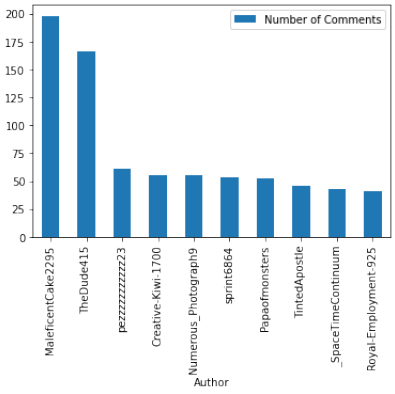

Figure 7: Authors vs. Number of Comments

2.3 YouTube Comments

(1) Figure 8 shows the 10 videos with highest number of comments between 2022/10/26 00:00:00 UTC and 2022/10/26 23:59:59 UTC
(2) Figure 9 shows the 10 authors with highest number of comments between 2022/10/26 00:00:00 UTC and 2022/10/26 23:59:59 UTC

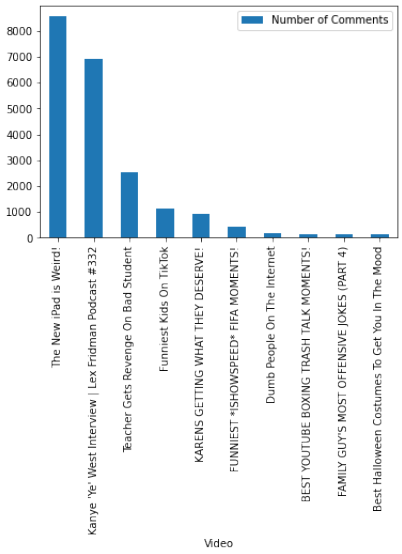

Figure 8: Videos vs. Number of Comments

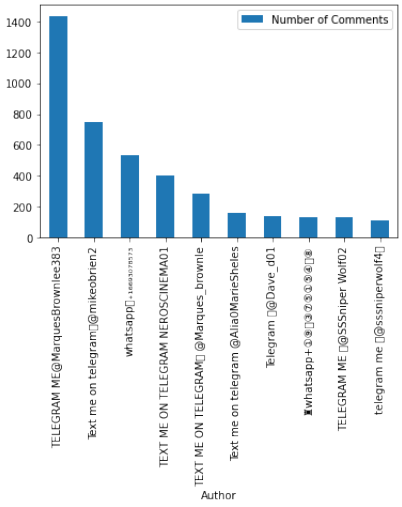

Figure 9: Authors vs. Number of Comments

3 API LIMITATIONS

(1) Listings provided by Reddit API are limited to 1000 items
(2) YouTube Data API allows only 10,000 requests per day
(3) YouTube Data API only supports replies to a top-level-comment and does not have to functionality to fetch reply of replies

4 DATA COLLECTION RATE

Figure 10 shows the trend of tweets grouped by each hour between 2022/10/29 00:00:00 UTC and 2022/10/29 23:59:59 UTC. Total count of tweets adds up to 53962.

Figure 11 shows the trend of Reddit comments grouped by each hour between 2022/10/29 00:00:00 UTC and 2022/10/29 23:59:59 UTC. Total count of comments adds up to 26523.

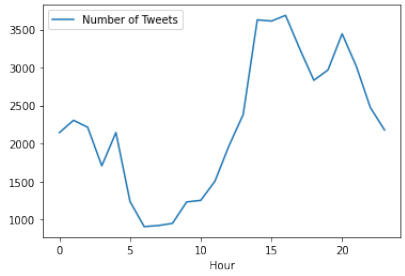

Figure 10: Hour vs. Number of Tweets

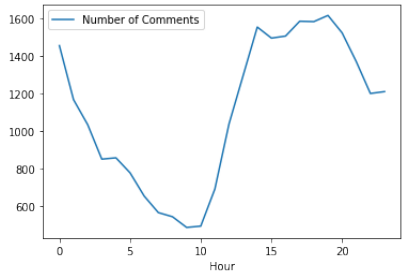

Figure 11: Hour vs. Number of Comments

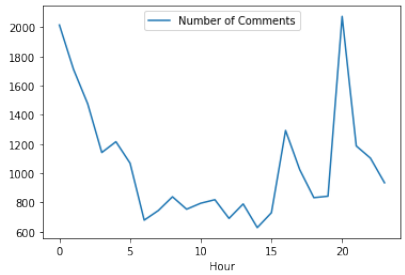

Figure 12: Hour vs. Number of Comments

Figure 12 shows the trend of YouTube video comments grouped by each hour between 2022/10/26 00:00:00 UTC and 2022/10/26 23:59:59 UTC. Total count of comments adds up to 25393.