Sentiment Analysis of Reddit & YouTube

Data Collection Report

Venkata Purna Vishnu Vardhan Kolleboyena
Computer Science
Binghamton University
Binghamton, New York, USA
kvenkatapurn@binghamton.edu

Avinash Kunamneni
Computer Science
Binghamton University
Binghamton, New York, USA
akunamneni@binghamton.edu

Shashank Reddy Karnati
Computer Science
Binghamton University
Binghamton, New York, USA
skarnati@binghamton.edu

ABSTRACT

In today's world, social media has seen an exponential increase in data production and consumption. Unfortunately, social media platforms have increasingly become hotspots for negativity. Many users, whether using anonymous profiles or public ones, often exhibit more negative and toxic behavior when confronted with differing opinions. This issue has grown to substantial proportions and requires a solution. It's crucial that we identify and categorize content based on sentiment to ensure all users adhere to the platform's guidelines. As an initial step in tackling this problem, we aim to gather data from Reddit and YouTube for sentiment analysis.

KEYWORDS

RedditAPI, YouTubeAPI, Java, Http Client, Python, Matplotlib, PostgreSQL, Text Blob, Data Analysis, Data Collection, Data Visualisation.

ACM Reference format:

Venkata Purna Vishnu Vardhan Kolleboyena, Avinash Kunamneni and Shashank Reddy Karnati. 2018. Sentiment Analysis of Reddit and YouTube: A Data collection and analysis pipeline. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 INTRODUCTION

Social media usage has become an integral part of our daily lives, with individuals spending approximately 4 to 5 hours each day scrolling through social media platforms, actively engaging with users worldwide, and sharing their opinions on various topics. Amidst this wealth of knowledge and information exchange, we frequently encounter comments that diverge from the ongoing discussions, occasionally crossing the boundaries of lawful discourse. Hate speech & unlawful criticism are prevalent issues, and many comments fail to comply with social media platform rules.

The manual identification and labelling of such comments would be an exceedingly arduous task for anyone. Therefore, as a key component of our project, we aim to automate this labelling process to the greatest extent possible. Our initial focus will be on establishing a data collection system to gather content from specific subreddits and YouTube video comments. Reddit and YouTube, being prolific real-time data generators, are our chosen platforms for the first phase of our sentiment analysis on social media.

2 Implementing the Data Collection

We are collecting data from Reddit and Youtube. We have selected few Subreddits and a youtube videoId for our data collection system. We have used Java HTTP client libraries to make HTTP requests. We have used JSON libraries to parse the JSON responses to fetch the data we needed for our analysis.

2.1 Reddit Data

We've established a database table to maintain a record of the chosen subreddits. Our code is designed to retrieve the subreddit list from this database and initiate HTTP requests to gather the relevant data from these subreddits. This approach enables us to dynamically change the source of subreddits.

We are referring to one of the reddit provided api's https://oauth.reddit.com/r/subreddit/comments to fetch the comments under specific subreddits. The API method is GET. Here's the list of subreddits that we are currently collecting the data from: worldnews, news, conspiracy, TrueReddit, offbeat, soccer, science,

movies, breakingbad, gameofthrones, gaming, announcements. These are few of the subreddits that seems to be active. We are fetching the following data from the JSON response: subreddit, submissionId, commentId, commentText, comment creation time UTC and inserting into the database table.

2.2 YouTube Data

We are using the google provided APIs to fetch the comments data using the videoId. We are storing the videoId in a database table. Our code is designed to retrieve that videoId and perform the HTTP calls. We can dynamically add/delete videoId from the database table. We are only fetching the latest videoId added into the database.

https://www.googleapis.com/youtube/v3/commentThreads. We are using HTTP client libraries to make the http calls and using JSON libraries to parse and fetch the data we needed for our analysis. We are collecting the following data: commentId, targetVideoId, commentText and comment creation time UTC.

3 System Design Architecture

Our project involves the development of a Java application within our integrated development environment (IDE), comprising two separate Maven projects. The first project is dedicated to collecting comments from Reddit, while the second focuses on gathering comments from YouTube videos. To achieve the desired automation and data collection efficiency, we are using scheduled executor service.

The scheduled executor service in Java is a component of Java from concurrency framework that allows us to automate the execution of specific tasks at predefined intervals. In our case, we have configured the scheduler to run our data collection and processing tasks at regular 5-minute intervals. This automation not only ensures the consistent and timely retrieval of comments from Reddit and YouTube but also minimizes the need for manual intervention.

For this project, we employed HTTP client libraries for making HTTP calls, a JSON library to parse JSON data, and PostgreSQL driver libraries to facilitate interactions with the database. Our code is hosted on GitHub. Within our virtual machine (VM) environment, Maven is configured to handle packaging. Once the code is successfully packaged, we are running the resulting JAR file in the background using the **nohup** command. This approach ensures that our session remains active, allowing our code to run uninterrupted.

Please find the system architecture diagram in Figure 1 that explains the details of our project.

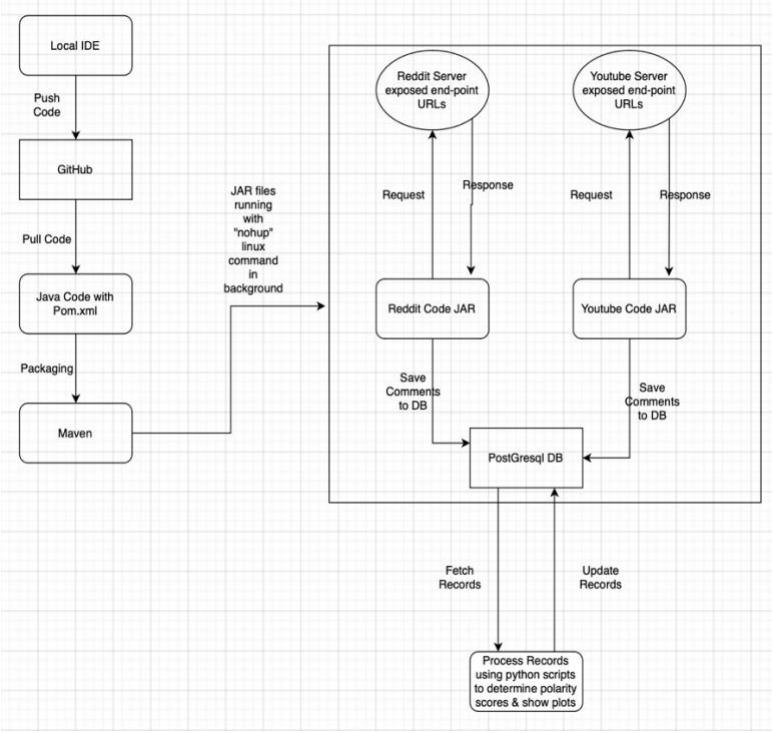


Figure 1

As part of our initial project, we are collecting the data. Once the data is collected, as shown in the figure 1, we are going to run the Pythons scripts to process the data for sentiment analysis.

3.1 Preliminary Exploration

While implementing our data collection system, we have adhered closely to our project proposal, with only one notable deviation. Initially, we had proposed collecting a maximum of 100 comments per API request. However, we have since modified our approach to collect data from approximately 12 subreddits, which may yield a maximum of 300 comments per request. It can still vary depending on the traffic.

Our preliminary exploration of the data has revealed that the collected data can be notably erratic. Despite aiming for 300 comments per request, the actual number of comments retrieved can exhibit significant fluctuations. We attribute this volatility to the dynamic nature of online platforms, especially evident in sources like Reddit. Additionally, during our exploration, we observed that a portion of the data includes comments in languages other than English. Since most machine learning models and Python libraries primarily support English, we are strategizing to focus our sentiment analysis on English comments and generate visualizations based on this subset of data.

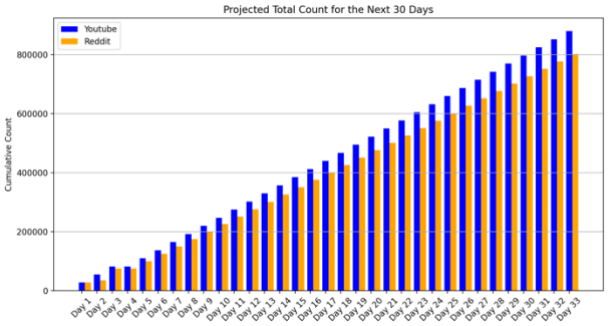
3.2 Challenges

The primary challenge we faced was ensuring our code continues running within our virtual machine (VM) even when the session is not active. However, we came up with the **nohup** command which helped us to overcome that challenge.

Furthermore, estimating data volumes poses another significant challenge due to the highly variable and unpredictable nature of social media data. However, since we are configuring the data source dynamically, we plan to meet our expected collection range.

3.3 Data Collection Projection

We plotted down based on our initial 72-hour collection using a python script. Below are the data collection projections of our project for 30 days. This is the amount of data being collected over the time.



However, as we cannot guarantee the traffic on social media, the final number may vary. But we are planning to collect nearly 1M to 1.2M comments for each data source. i.e. Reddit & YouTube.

For our data sources, since we have configured to change the videoId and subreddits dynamically, our plan is to collect between 1M to 1.2M comments each. We will change the source accordingly to collect the data in that range towards the end of the project.

4 REFERENCES

Noman Ashraf, Arkaitz Zubiaga, and Alexander Gelbukh - Abusive language detection in youtube comments leveraging replies as conversational context - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8507480/

https://www.ijraset.com/research-paper/analysing-sentiments-for-youtube-comments

https://medium.com/@kiddojazz/reddit-sentiment-analysis-f8a1a790124