Sentiment Analysis of Reddit & YouTube

A Data Collection & Analysis Pipeline

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

With the ever-growing usage of social media these days, a vast amount of data is being produced and consumed. Unfortunately, social media has increasingly become a breeding ground for negativity. Many users, whether anonymous or with public profiles, exhibit more negative and toxic behaviour when they encounter differing opinions. This issue has grown to significant proportions, and it necessitates a solution. We must identify and label negative content to ensure that every user adheres to the platform's rules. As a first step towards addressing this problem, we intend to collect data from Reddit and YouTube for our 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 automating the detection and labelling of negative comments on social media.

2 Data Collection

We are going to collect the data of both the Reddit and YouTube using the APIs provided by the Reddit and YouTube via their API documentation. We are going to make use of the APIs that are being exposed by them to collect our data on daily basis. We are going to use Java Http client libraries to make API calls and parse them into our desired format and store into the PostgreSQL thereafter.

2.1 Reddit Data

We are going to use the Reddit API to fetch all the comments under a subreddit. We are going to collect the data of specific subreddits, which we will be able to dynamically change. The source of subreddits can be changed from outside of the code via a file. Our crawler is going to fetch the comments under those listed subreddits.

We are planning to refer 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. We are planning to fetch the comments for various subreddits using keywords. The comments will be collected for further data analysis.

2.2 YouTube Data

We are going to use the Google APIs for fetching the data under random YouTube videos. We are going to collect the data using the HTTP GET method API provided by google.

https://www.googleapis.com/youtube/v3/commentThreads. Our plan is to get the comment threads and store in the PostgreSQL in desired format for further data analysis.

3 Data Analysis

Once the data has been successfully collected and stored in the database, our primary objective is to ensure that it conforms to the desired format through a data cleaning process. Subsequently, we plan to apply specific data analysis models using Python libraries.

The very initial thoughts on the analysis are that we are considering the utilization of the Text Blob library for sentiment analysis. However, the choice of our library is subject to change as we move further and explore various libraries that are relevant to our context and expertise. Basically, our plan is to assign polarity scores to the comment, indicating whether the text is positive, negative, or neutral. At later point, the data will be visualised using Matplotlib.

3.1 Estimated Data Collection

For Reddit Data Collection:
Maximum:
For every 5mins, we collect 100 comments.
For every 1hr, we collect 100*12 = 1200 comments.
For every 24hr, we collect 1200*24=28,800 comments.
For every 1 week, we collect 28,800*7= 201,600 comments.
For YouTube Data Collection:
Maximum:
For every 5mins, we collect 100 comments.
For every 1hr, we collect 100*12 = 1200 comments.
For every 24hr, we collect 1200*24=28,800 comments.
For every 1 week, we collect 28,800*7= 201,600 comments.
However, as we cannot guarantee the traffic on the social media platforms, the final numbers may vary. Since we are going to limit the count from our end, it could not be more than the maximum number as mentioned above. But there is a possibility that it can be lower than that depending on the traffic.

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-f8a1a790124a