Sentiment Analysis of Reddit & YouTube

A Dataset Measurement & Analysis Proposal

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 the second phase of our project, we continue to collect and analyse data from Reddit and YouTube comments to address the issue of negativity and toxic behaviour on social media platforms. Our focus is on conducting sentiment analysis and applying modern hate speech detection APIs to understand the emotions and toxicity levels within these online communities. We will use data visualizations to compare Reddit and YouTube, providing insights into the prevalence of toxic content and the impact of differing opinions on these platforms. This research aims to contribute to a more positive and respectful online environment for all users.

KEYWORDS

RedditAPI, YouTubeAPI, Java, Http Client, Python, Matplotlib, PostgreSQL, Text Blob, Data Analysis, ModernHateSpeech API, 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

In the second phase of our project, we are committed to tackling the growing issue of negativity and toxic behaviour within the realms of social media platforms, particularly Reddit and YouTube. Our strategy involves a two-fold approach, employing both sentiment analysis and the ModernHateSpeech API to delve deeper into the intricacies of these online communities.

First, we will harness the power of TextBlob to perform sentiment analysis on the vast corpus of comments generated by users. This analysis will provide us with valuable insights into the emotional and attitudinal underpinnings of these interactions. By categorizing the sentiments as positive, negative, or neutral, we can unveil patterns and trends in user behaviour.

In parallel, we will leverage the ModernHateSpeech API to detect and quantify the presence of toxic content within these platforms. This specialized tool will help us identify and categorize harmful content, including hate speech, insults, and offensive language, ultimately contributing to the creation of a more respectful and positive online environment.

By employing these two distinct yet complementary approaches, we aim to gain a comprehensive understanding of the emotional landscape and the extent of toxicity within these online communities. Through our research and data visualizations, we hope to shed light on the impact of differing opinions and provide insights that will pave the way for a more constructive and harmonious online space for all users.

2 System Design for Data Analysis

As we are collecting the data continuously, we are simultaneously going to run a python script which processes the data records in batches to evaluate and update the sentiment score and hate speech detection.

Our plan is to retrieve the records that are being collected in batches and then at first, we are going to perform the sentiment analysis. Subsequently, for the same data we are going to perform the API call to the modern hate speech API which will return us the results. We are going to store those two fields back into the data base against those data records. You can find a simple overview of what we are going to perform in the below

figure 1. The figure assumes entire Project 1 work as a Data collection System being in place.

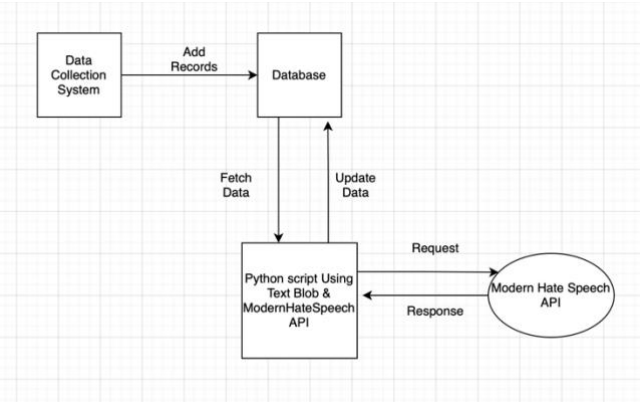

Figure 1

3 Data Description

Once we start evaluating the sentiment analysis and hate speech content for the comments that we are collecting in both YouTube and Reddit, we intend to obtain certain plots that can help us explain the data.

For example, we are going to plot down the count of the comments that is being collected over the span of 30 days for both the data sets. We are going to plot how is the trend in sentiment analysis across various subreddits and how is the sentiment trend in the YouTube videos. We are especially collecting data from English pop singer's music videos. We are going to plot down, how much of the overall content is hateful/non hateful and what are the trends.

We are maintaining separate database tables for collecting "r/politics" for the given time. This will help us easily distinguish politics from the other list of subreddits that we are already collecting in parallel. We are going to plot down the number of submissions and comments made in "r/politics" between the given time. Similarly, we are going to plot down every subreddit's data collection count trend on X-axis and Y-axis graph.

On a whole, we are going to describe our data in plots as mentioned above and we are planning to address the below three research questions as well in our further work.

1. Is there a corelation between the sentiment analysis and the hate speech detection in the comments?
2. How is the overall trend in the hate speech across various subreddits and YouTube videos?
3. Text Blob vs Modern hate speech API. How effective are these two when compared against manual labelling for a small sample size?

4 Additional Information & Validation

As part of our data collection system, we are collecting the data from both Reddit and YouTube. Our projections as part of our Project 1 have shown that we will be collecting nearly one million comments from both the Reddit and YouTube and we believe that the data would be suffice for us to carry out our analysis. Hence, no further additional data apart from the mentioned subreddits and YouTube is required for our analysis.

5 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

Understanding the Behaviors of Toxic Accounts on Reddit - https://dl.acm.org/doi/abs/10.1145/3543507.35835