Analyzing Emotions - Twitter / Reddit / Youtube

Nagalakshmi Prasanna Pujita Bodapati
Computer Science
Binghamton University
Binghamton, New York, USA
nbodapa1@binghamton.edu

Saisuraj Aitha
Computer Science
Binghamton University
Binghamton, New York, USA
saitha2@binghamton.edu

kshamitha Gandu
Computer Science
Binghamton University
Binghamton, New York, USA
kgandu1@binghamton.edu

Mohith kumar Sopparam
Computer Science
Binghamton University
Binghamton, New York, USA
msoppar1@binghamton.edu

Anshul Upadhyay
Computer Science
Binghamton University
Binghamton, New York, USA
aupadhy5@binghamton.edu

Abstract

The comments sections of social media platform have become the new playground for online bullying.The impact of toxic comments is much more catastrophic than we think. Therefore, having a solid toxicity flagging system in place is important if we want to maintain a civilized environment on social media platforms to effectively facilitate conversations.But, deciding if we have to flag a comment or not is very time taking. If we have an automated process where we can automatically detect abusive keywords in the comments or posts can save the time of website moderators and also it will have a great impact on improving the discussion rather than focusing on using harsh words. So, we are mainly focusing on posts/comments in Reddit, Youtube and Twitter that contains toxic keywords and flag them as negative emotions by taking into consideration of data we have obtained in a period of time.

Keywords: YouTube API, Tweet-Stream, data analysis, Reddit API, GoogleAPIClient, Python, MongoDB, Twitter API, Reddit API

ACM Reference Format:
Nagalakshmi Prasanna Pujita Bodapati, Saisuraj Aitha, kshamitha Gandu, Mohith kumar Sopparam, and Anshul Upadhyay. 2022. Analyzing Emotions - Twitter / Reddit / Youtube. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recently, people have started using online forums frequently for discussions. Along with the discussions on online forums, trolls and spammers have become more common. It is a time-consuming and tedious job to moderate the comments and posts on these forums, so the organizations have to rely on external people to handle them. We are trying to differentiate between negative and positive comments and tweets on Twitter, Reddit, and YouTube. To achieve this, we would get real-world data from the above mentioned platforms. We also intend to visualize Twitter, Reddit, and YouTube's data.

2 DATASET AND DATA COLLECTIONS

We have chosen the 3 datasets - Twitter / Youtube / Reddit. From all the 3 data sources we have collected the data and stored it in MongoDB. When collecting data we haven't used any filters and collected all the data that is required and analyzed for emotions and performed sentimental analysis.

3 DATASOURCES

As mentioned above, the main 3 data sources we have chosen are: ***Twitter*** : For retrieving the Twitter data (tweets), we have used the TweetStream module that has 2 sources to obtain APIs - Sample Stream and FilterStream.

But, In our project, we have made use of Sample Stream and retrieved the data from Twitter. Once the data is collected and stored in the database, we performed sentimental analysis on the module to determine the final results - Positive / Negative and Neutral.

Youtube : For retrieving the comments from youtube, we have used the GoogleAPIClient module. In this project we directly tool the "textOriginal" form of comments to perform our sentimental analysis on every single comment and determined the final results.

Reddit : The third datasource is Reddit, the subreddit we have chosen to retrieve the posts is "r/political" subreddit. The module we have chosen is taken from the Reddit API itself where we have directly hit the API to retrieve the posts

Conference'17, July 2017, Washington, DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and store it in the mongoDB. Once the data is stored, we have looped through every single post and performed the sentimental analysis to get the final result.

4 METHODOLOGY

We used SampleStream to gather tweets from twitter, GoogleAPIClient to extract random comments for a particular video and used direct Reddit oauth authentication and directly hit the API to collect data on subreddits by generating a query and collected the posts from these subreddits. Once, the complete data is collected we stored in the MongoDb by creating separate database and collection for each datasource we have chosen.

We have collected an average of 235197 twitter posts and approximately 66603 YouTube comments and around 262312 reddit posts over the past seven days. The data will be then extracted from MongoDB and cleaned by implementing a regex function. By using this function, we will now remove the noise from the data collected. Now we use this updated clean data to perform the analysis to find emotions whether a particular tweet/comment/post is positive, Negative or neutral.

The process of final analysis or result can be obtained using Textblob which is python module. By check every comment/post that is collected and sent to textblob the result will be in form of a polarity where we have 3 options - -1, 0, 1 which shows Negative / neutral / Positive.

5 PROJECT WORKFLOW

The data is retrieved from three main data sources Reddit, YouTube and twitter. We use Python modules, Tweet Stream, PRAW and GoogleAPIClient. The data is stored in MongoDB, and it is cleaned in order to remove all the inconsistent data. Machine learning algorithms are implemented to generate contextual data once the clean data is obtained. Different methods like Tokenization and converting strings to lowercase techniques are used in order to obtain features from MongoDB. Once the data is collected and stored, we analyze this data by using the python modules and by considering various keywords we detect the emotions using the Textblob python module.

6 RESEARCH OBJECTIVES

The main aim of the project is to understand better how the people use the social media sites. The decisions made by these individuals may come from the situations or individual persons attitude and mindset. Its not unusual to come accross hateful comments or posts that are made by large number of persons.

The main research question we are trying to find is:

- From the collected data - what percent of data is POSITIVE? - From the data collection part and the performance of sentimental analysis for all the 3 data sources, we found that for Twitter, the positive percentage is **13.69**, and for Youtube, the positive percentage is **93.01**, and for Reddit, the positive percentage is **12.27**
- From the collected data - what percent of data is NEGATIVE? - From the data collection part and the performance of sentimental analysis for all the 3 data sources, we found that for Twitter, the negative percentage is **5.25**, and for Youtube, the negative percentage is **6.98** and for Reddit, the negative percentage is **5.36**
- From the collected data - what percent of data is NEUTRAL? - From the data collection part and the performance of sentimental analysis for all the 3 data sources, we found that for Twitter, the neutral percentage is **5.25**, and for Youtube, the neutral percentage is **0**, and for Reddit, the neutral percentage is **8.64**

For the reddit dataset - the unwanted or filtered comments percent is **73.73**

7 DASHBOARD AND TOOLS ANALYSIS

We are thinking of implementing the dashboard using Pure HTML and for the styling part, we are using Bootstrap. By obtaining the results by performing the sentimental analysis we will graphically represent our result in a web page.

The visualization of the graphical representation will have date pickers where you can select the dates and once the submit button is clicked the corresponding graphs will be shown in the UI.

We will show the 3 data sources by comparing the emotional analysis for each graph.

For each data source - graphs with the sentiments will be shown - Positive / Negative / Neutral.

8 CONCLUSION

So, finally, the main aim is to analyze the extracted data from Reddit, Youtube and Twitter. Using various tools, we will graph the results and outcomes on a web page by giving filters to select the date range, and the results will be shown for each data source with sentiments shown as bar graphs.

9 REFERENCES

1. Suvrat Arora — Published On July 7, 2022 and Last Modified On July 12th, 2022 - Beginner Machine Learning NLP Python. Emotions Analysis of a Youtube: Blog. 2. Brendan Martin - Founder of LearnDataSci, Nikos Koufos - CS Engineering Post Graduate - Emotions Analysis on Reddit News Headlines with Python's Natural Language Toolkit: Article. 3. Yalin Yener - Data Engineer | GIS Analyst | Data Analyst | Data Scientist - Twitter Emotions Analysis in Python: Blog. 4. https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners.