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, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
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 DATA COLLECTION

For analyzing the twitter data ***Tweetstream*** module is used. This module will provide a framework where Twitter streaming API can be used. For scraping Reddit data - we used ***RedditAPI*** that is posted on Reddit can be accessed using this module. For the third data source we have choosen YOUTUBE and the data can be accessed using a google ***GoogleClientAPI***

2.1 COLLECTING TWITTER STREAMING API

Tweetstream is a python module that can be used to acquire tweets from the Twitter streaming API. Mainly there are two kinds of API's to obtain the tweets from the twitter - Samplestream and FilterStream. . In real time FilterStream is mainly used as it sends the tweets by considering a mere set of criteria and this criteria can filter the tweets by the following 3 keys:

- To search for certain specific keywords or can also track by a specific keywords.
- Based on the current user it can filter user specific tweets.
- This also can filter using specific locations but the tweets should be geo-tagged tweets.

We developed the twitter data collection using java. By accessing the various secret API keys and bearer tokens collected the data using the URI Builder ***tweets/sample/stream***. Once the data is collected we used python modules -***textblob*** and ***numpy*** by filtering the posts using regex. The following are the posts that have been collected. The overall data collection is **235197**.

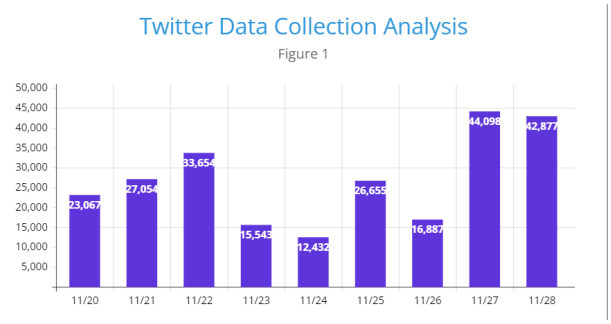

Figure 1. Figure 1a shows the overall collection of twitter for specific dates

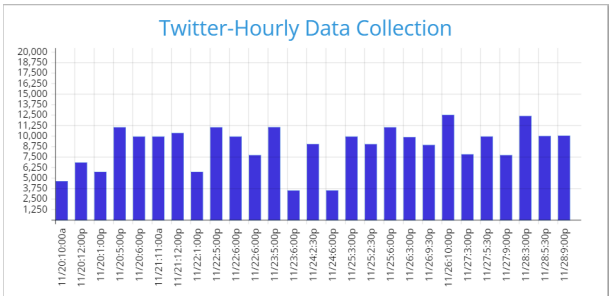

Figure 2. Figure 1b shows the hourly collection of twitter for specific dates

2.2 COLLECTING YOUTUBE STREAMING API

For extracting YOUTUBE comments we use GOOGLECLIENTAPI that is mainly designed for Python client-applications. By using this API it is very flexible and also easy to access Google API's. There two types of access - Simple or Authorized. To use one of there access methods we have to make use of a build() function that will generate a service object and then from here we can decide if we need Simple access or Authorized access. By taking random videos we will collect all the random comments and later we analyze them.

- Read only instance can be used to retrieve the post that are publicly available.
- Where as Authorized instance can be used to post, comment etc.

Once the data is collected and stored for a frequency of dates in mongodb collections the following is the final analysis of comments collected for a period of time. The overall data collection is **66603**.

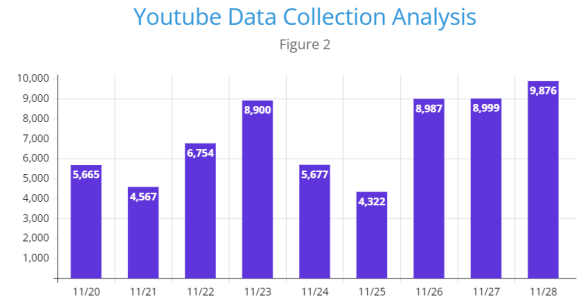

Figure 3. Figure 2a shows the overall collection of youtube comments for specific dates from a video with more than 60Million comments

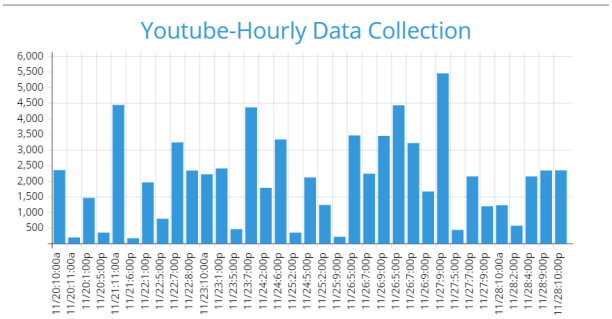

Figure 4. Figure 2b shows the hourly collection of youtube comments for specific dates from a video with more than 60Million comments

2.3 COLLECTING REDDIT STREAMING API

For extracting posts that are posted on Reddit we are using Reddit API where we directly access the data by using HTTP requests uisng ***www.reddit.com/dev/api/***. We are mainly using Python programming language to make the HTTP requests and store the extracted data in Mongodb using the module - pymongo. Also, we have created triggers where the data will be automatically collected and will store in the mongodb.

Once the data is collected and stored for a frequency of dates in mongodb collections the following is the final analysis of posts collected for a period of time. The overall data collection is **262312**.

3 DATA COLLECTION ESTIMATES

For twitter we are collecting approximately 1M tweets, for Youtube - approximately 6M comments and for Reddit we

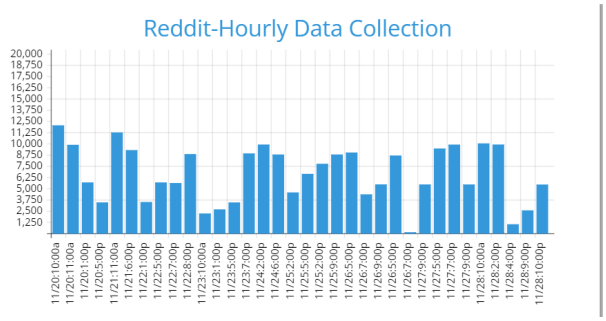

Figure 5. Figure 3a shows the overall collection of reddit posts for specific dates from the subreddit r/politics

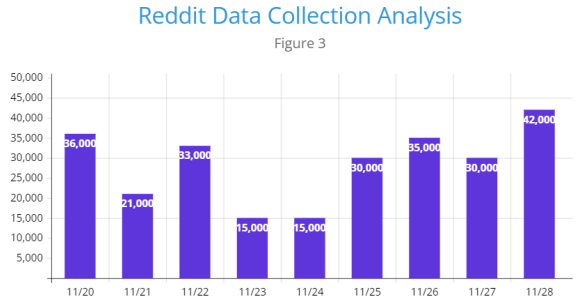

Figure 6. Figure 3b shows the hourly collection of reddit posts for specific dates from the subreddit r/politics

have an option of passing the parameter "LIMIT" and can set any value we wish to collect, so right now we are thinking of collecting approximately 100 posts per day. The data that is acquired is meaningful and we can use this data to analyze the emotions.

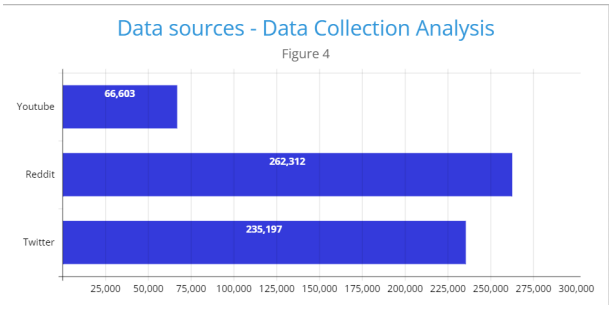

Figure 7. Figure 4 shows the overall collection of all the 3 data sources

4 METHODOLOGY

We used SampleStream to gather tweets from twitter, GoogleAPI-Client 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. For example, this function can be used to remove data that is not in english language. 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.

The following figure shows the final distribution of results of all the 3 data sources - Twitter/Reddit/Youtube:

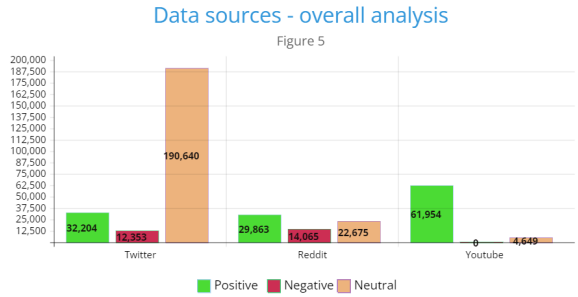

Figure 8. Figure 5 shows the overall distribution of data by resulting in positive/negative/neutral

5 ANALYZING DATA

For the data mining, we need to have raw text data which can be achieved by filtering the data and it simplifies the process of collecting the text's information by using machine learning algorithms. If we don't consider using any of the machine learning algorithms, the possibility of noisy and inconsistent data will increase significantly. So by using these algorithms, it removes inconsistent data, which won't help us understand the emotions behind tweets, posts and comments.

Once the data has been cleaned, we'll implement a classifier that accurately categorizes the negative tweets, posts and comments. In general, we'll implement either a contextual classifier or a general classifier. The first step is to identify if the tweet/ phrase is objective or subjective, which is a two-step classification method. And the next step is to determine whether the tweet is negative or not using machine learning techniques by executing a process. Once all the data has

been categorized,we analyze and thinking of representing the data on a bar graph.

6 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.

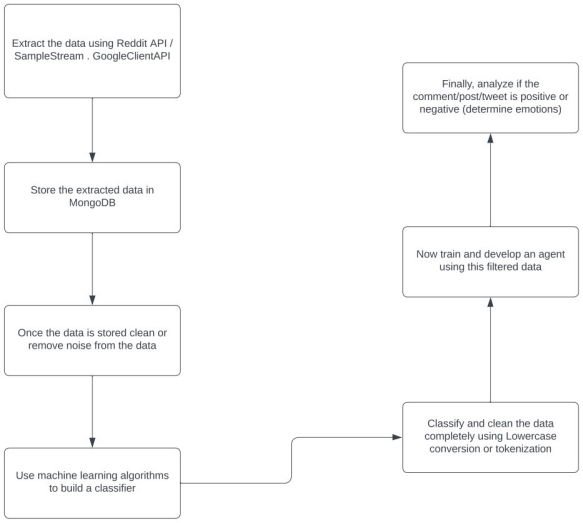

Figure 9. Figure 6 shows the overall distribution of data by resulting in positive/negative/neutral

7 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 collected data - what percent of data is NEGATIVE?
- From the collected data - what percent of data is NEUTRAL?

8 API METHODOLOGY

Once the data is collected for the 3 data sources and stored in mongodb. Each and every tweet/post/comment will be analysed individually by sending every single tweet/post/comment to analyze and get the polarity by which we can determine how much of the data is positive, negative or neutral.

The final results obtained and the data stored in mongo db is shown as follows:

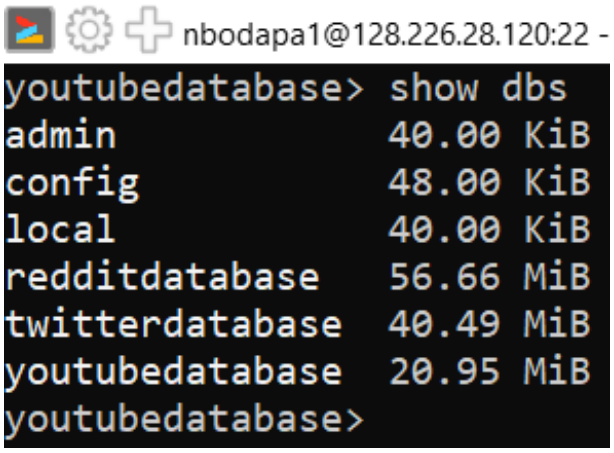

Figure 10. Figure 7 shows the data storage in mongo database

The final results obtained and the counts stored in mongo db is shown as follows:

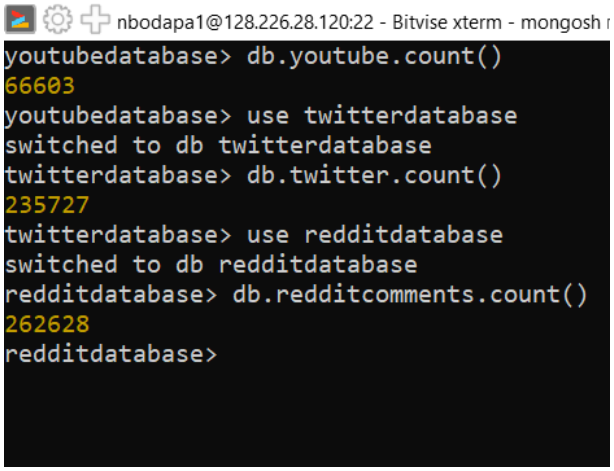

Figure 11. Figure 8 shows the counts obtained for each datasource in mongo database

The final results obtained and the complete analysis of positive/negative/neutral data obtained for each data source:

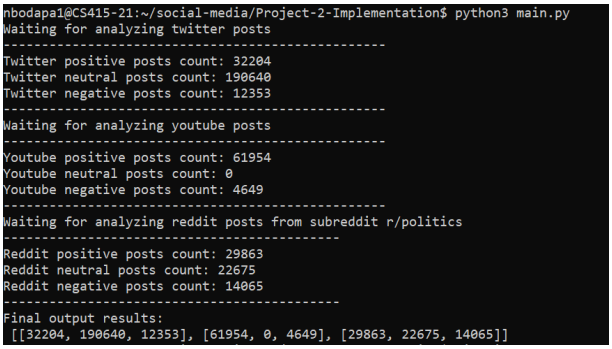

Figure 12. Figure 9 shows the final counts obtained for each datasource in mongo database

9 TABLE DATA ANALYSIS

The following figure represents the tabular form of data analysis representing the three data sources and the results obtained by performing the sentimental analysis

Total Count - Data Collection Analysis

Data Source	Total	Positive	Negative	Neutral
Twitter	235197	32204	12353	190640
Reddit	262312	32204	14065	22675
Youtube	66603	61954	4649	0

Figure 13. Figure 10 shows the table data obtained for each datasource

10 CONCLUSION

So, finally the main aim is to perform data analysis on the extracted data from Reddit, Youtube and Twitter. Using various tools we will represent the results and outcomes graphically.

11 ACKNOWLEDGEMENTS

We all would like to thank our Professor. Jeremy Blackburn for all the guidance and assistance provided to us for this project. We really appreciate the help and support given to us for the improvement and various opportunities to learn and always explore new technical stuff.

12 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/

5. https://www.natasshaselvaraj.com/twitter-sentiment-analysis-with-python/

6. https://www.reddit.com/r/programming/comments/s74adr/

7. https://towardsdatascience.com/automate-sentiment-analysis-process-for-reddit-post-textblob-and-vader-8a79c269522f