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. It not only hurts one's self-esteem or deters people from having meaningful discussions, but also provokes people to such sinister acts as recent capital riots at US Congress and attacks on farmers for protesting in India.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

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*** python module is used. This module will provide a framework where Twitter streaming API can be used. For scraping Reddit data - we used ***RedditAPI*** scraper where data 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 library - ***GOOGLEAPICLIENT***.

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.

On the other hand SampleStream will provide random sample or short tweets that are posted in real-time. In this project, we will make use of SampleStream to acquire twitter data.

2.2 COLLECTING REDDIT API

For extracting posts that are posted on Reddit we are using Reddit API where we directly access the data by using HTTP requests. 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.

2.3 COLLECTING YOUTUBE 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.

3 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.

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

4 PROJECT WORKFLOW

The data is retrieved from three main data sources Reddit, YouTube and twitter. We use Python modules, Tweet Stream, Reddit API 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.

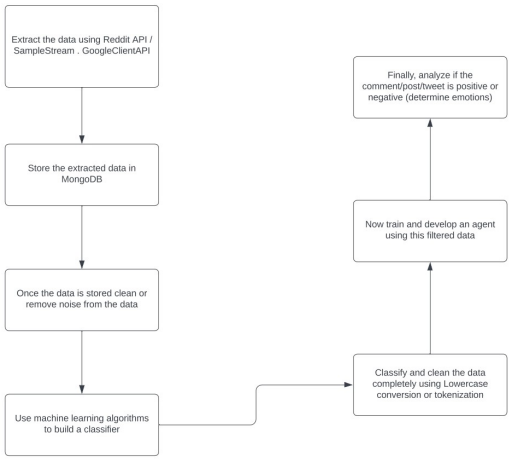

Figure 1. Project Flow

5 DATA COLLECTION ESTIMATES

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

6 Updates made from Proposal to Report

For data extraction of Twitter and Youtube we have created and developed using Java and Spirngboot. For reddit data extraction we have implemented in python programming language by directly requesting the HTTP API's and not using PRAW.

7 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 Post Graduate - Emotions Analysis on Reddit News Headlines with Python's Natural Language Toolkit: Article.