Sentiment Analysis Of Covid19 Tweets

Pinaki Dash (204101042) **Rajat Maheshwari (204161013)**
Stuti Priyambda (204101055) **Yashwant Patidar (204101062)**
Indian Institute of Technology Guwahati

Google Colab Link
https://colab.research.google.com/drive/17iYp9O5_uAdPHbhN5kAfyHlbDvfb0yIX?usp=sharing
Group No: 16
{pinaki_dash, mrajat, p.stuti, pyashwant}@iitg.ac.in

Abstract

The reach of social media platforms and its usage has increased many folds in the recent years. Social networks have become one of the most popular platforms to share one's ideas, opinions and reactions with the other users. This leads to easy access of peoples' opinions and sentiments towards many issues and events occurring around us. One such event is Coronavirus disease (COVID-19). In this project, we try to categorise public sentiments associated with the pandemic by collecting tweets regarding COVID-19 based on some specific hashtags, by making use of Twitter API and analyze the fetched tweets using NLP tools and techniques. We have used Twarc library and Naive Bayes classifier for carrying out the sentiment analysis operation. The results obtained after analysing are being shown in the form of graphical representation.

1 Introduction

Sentiment Analysis, also known as opinion mining, is a prime topic of research in Natural Language Processing which determines the attitude, opinions and emotions of a person within an online mention. This is highly used to find out how people feel about some particular topics. The rise in emphasis on NLP methods followed the tremendous increase in public dependencies on social media (e.g. Facebook, Instagram) for information instead of conventional news media. Social media is used extensively by many companies to advertise their products and services; researchers and practitioners to mine huge datasets to generate insights about people's behaviour, thoughts and emotions on a variety of social issues, etc. Consequently, social networks are becoming an important source of information for carrying out research about public sentiments.

1.1 Twitter sentiment analysis

Twitter allows only 280 characters per tweet. This restriction causes people to use phrases, words, abbreviations, emoticons to express their opinions about a topic. This makes the sentiment extraction process complex by using some existing NLP systems. Therefore some researchers have used deep learning and machine learning techniques to extract and mine polarity of tweets.

1.2 Project aim and scope

We strive to label relevant tweets by appropriate emotional polarity by using learning algorithms and classifiers. We aim at developing a model for predicting emotions by focusing on the relationship between words, thus labelling each of the relevant tweets as positive, neutral and negative. In this project, we follow a method to analyse data extracted from Twitter for identification of sentiment, key words association and trends to understand scenarios raised due to current COVID-19 phenomenon.

1.3 Challenges

Collection of data Getting access to Twitter API and thereafter making a dataset of the tweets regarding COVID-19.

Anaphora Resolution the problem of resolving what a pronoun, or a noun phrase refers to. ”We watched the movie and went to dinner; it was awful.” What does ”It” refer to?

Sarcasm If you don't know the author you have no idea whether 'bad' means bad or good.

structure of tweets abbreviations, lack of capitals, poor spelling, poor punctuation,poor grammar etc.

Training dataset Training the data set properly with labelled tweets as positive or negative relevant to COVID-19 to get appropriate accuracy while testing data.

2 Method

2.1 Data Extraction

At first, we tried to collect tweets by making use of Twitter API through Tweepy library but since we wanted the data set to include tweets generated during the month of April , 2020, we faced some issues. Later we found a well organised dataset for geo-tagged tweets from across the globe regarding COVID-19 in IEEE dataport. The tweets have been collected by an on-going project deployed at https://live.rlamsal.com.np The model monitors the real-time Twitter feed for coronavirus related tweets using 90+ different keywords and hashtags that are commonly used while referencing the pandemic. Some of the keywords and hashtags used are: ”corona”, ”corona”, ”coronavirus”, ”covid”, ”covid”, ”covid19”, ”covid19”, ”covid-19”, ”covid-19”, ”ncov”, ”ncov”, ”ncov2019”, ”ncov2019”, ”2019-ncov”, ”pandemic”, ”pandemic”, ”2019ncov”, ”2019ncov”, ”quarantine”, ”quarantine”, ”hand sanitizer”, ”handsanitizer”, ”lockdown”, ”lockdown”, ”social distancing”, ”socialdistancing”, ”work from home”, ”workfromhome”, ”working from home”, ”workingfromhome”, ”ppe”, ”n95”, ”ppe”, ”n95”, ”pneumonia”, ”pneumonia”, ”chinese virus”, ”chinesevirus”, ”wuhan virus”, ”wearamask”, ”wearamask”, ”corona vaccine”, ”corona vaccines”, ”coronavaccine”, ”coronavaccines”, ”faceshield”, ”faceshield”, ”face shields”, ”faceshields”, ”stayhomestaysafe”, ”coronaupdate”, ”frontlineheroes”, ”coronawarriors”, ”homeschool”, ”homeschooling”, ”hometasking”, ”masks4all”, ”wfh”, ”wash ur hands”, ”wash your hands”, ”washurhands”, ”washyourhands”, ”stayathome”, ”stayhome”, ”selfisolating”, ”self isolating”. This dataset contains around 140k geo-tagged tweets around the world and more than 310 million tweets in total. As per the need of the project, we filtered the tweets based on Indian origin from the portion of geo-tagged tweets and also filtered the non-geo tagged tweets by considering the [“place”] twitter object and by enabling boundary conditions. The above dataset contains only tweet IDs as sharing the whole text of tweets is not allowed as per policies. The tweets corresponding to the collected tweet IDs are fetched using Twarc library and Twitter API. The tweets thus fetched are stored in csv files for further processing. Next step is pre-processing of the tweets stored in the csv files.

2.2 Dividing the dataset

The tweets collected during the data extraction phase are used for training, validation and testing. For the training set, tweets are collected from the month of May, 2020 and are then preprocessed. These tweets are collected from a few days of each week of May, 2020 for having a fair distribution of words and sentiments. These tweet IDs are collected from global dataset and are filtered for Indian origin and the corresponding tweets are stored. Around 30000 tweets are used for building the training data set. For the validation set, tweets from the month of August, 2020 are being used. These tweets are also extracted from global dataset and subsequently filtered for Indian origin the corresponding tweets are extracted. For the testing set, we are considering only geo-tagged tweet ID from April and September. After getting a list of such IDs, we extract the corresponding tweets for testing the model.

2.3 Preparing the training data set

For designing the corpus for training data set, we collected tweet IDs from a few days of each week of May, 2020 for having a fair distribution of words and sentiments. After extracting the tweets associated with these tweet IDs, next step is to label each and every tweet. We have used TextBlob library’s built in polarity tool to assign polarity to each and every tweet. The polarity lies between -1 to 1 (both inclusive). We have set the threshold polarity to be 0 for declaring a tweet to carry positive sentiment. If polarity of a tweet ¿ 0, it is labelled as positive. If polarity ¡ 0, it is labelled as negative, otherwise neutral. So each row of the corpus for training data contains tweet ID, tweet and label of the tweet decided by its polarity obtained using TextBlob.

2.4 Data preprocessing phase

We need to preprocess our data so that we can save a lot of memory and reduce the computational process. After the extraction of tweets about covid19 we are going to perform preprocessing of tweets. Preprocessing involves a series of techniques which should improve the next phases of elaboration, in order to achieve better performances.

tweets usually contain lots of noise and uninformative parts, such as URL, mentions and hashtags. This increases the dimensionality of the problem and makes the classification process more difficult. The algorithms which are most used to polish and prepare data that comes from Twitter include the removal of punctuation and symbols, tokenization, removal of stopwords, stemming, and replacement of negations etc.

Basic cleaning Converting every character into lower case and Removal of punctuations and symbols.

Removal of repeating characters people are often not strictly grammatical. We do write words like “caaaaaaar”, “happppppppy” in order to more emphasize on the word. We are converting such words into it’s correct form i.e.. “caaaaaaar” will become ”car”, “happppppppy” converts to ‘happy’.

Removal of HashTag,URL and Emoticons

- Removing hashtag from the tweet such as #COVID-19 to COVID-19, #SSRwarriors to SSRwarriors.
- Removing URL (starting from https and www) and any tagging of a particular account (@username).
- As emojis and emoticons play a significant role in expressing the sentiments we need to replace them with the expression they represent in plain English so we convert emoticons and emojis into words emot library

Stopword’s Stopword’s removal enhances the system because it removes words which are useless for the classification phase. As a common example, an article does not express a sentiment but it is very present in the sentences.

Contractions Contractions are words or combinations of words that are shortened by dropping letters and replacing them by an apostrophe.Removing contractions contributes to text standardization. In Contraction we are converting “you’re” into ‘you are’, “it’s” converted to ‘it is’ and many more. For expanding these contractions we are using ”contractions” library.

2.5 Naive Baye’s Classifier

We will be having corpora of tweets with positive, neutral and negative sentiments.

Positive tweets Example ‘I am happy because I can focus on online learning’ , ‘I am happy that trump will be our PM again’

Neutral tweets Example ‘I am OK with whatever type of learning be it online or offline’ , ‘I am OK with whoever becomes our Prime Minister.’

Negative tweets Example ‘I am sad, I am not able to learn in online mode,’ , ‘I am sad, Modi will not be our Prime Minister.’

Naive Bayes Classifier depends on Bayes’ Theorem. The first step is about building the vocabulary. It will contain all the tokens present in the preprocessed training data set. The list word_feature is used to store the distinct tokens with the count of occurrence of such tokens acting as key.
The second step is about matching the list word_feature word by word against the tweet at hand and associating 1/0 with the word. 1(true) is assigned if the word in vocabulary is present in the tweet and 0 (false) is assigned if the word in vocabulary is not present in the tweet.
The third step is about building the feature vector.To perform feature extraction from the above list, we have used apply_features() function of nltk.

The fourth step is about training the classifier. We have used the built in Naive Bayes Classifier of nltk. In order to train the classifier, we call train( ) function and pass the feature vector obtained in the above step to this function.

3 Contribution

Yashwant and Pinaki started data collection by registering an application with Twitter as it is the only way to get authentication credentials. After receiving the credentials, they authenticated the python script with the API using the credentials. They tried to fetch the tweets using Tweepy library that were generated during April, 2020 but faced problems in retrieving the tweets. Then they came across the covid tweets dataset hosted in IEEE dataport and decided to use the same since it suited the project’s need.
Stuti and Rajat did extensive research on existing systems on tweet sentiment analysis and covered many models that can be implemented as a part of this project given the time constraint without losing much on accuracy with satisfiable output.

They found out that, for time being, using Naive Bayes classification model will cater our purpose of tweet sentiment analysis. They also proposed to use another model and check the accuracy of the new model with the dataset fetched but this part is kept as a future scope for now.
Rajat and Stuti has worked for preprocessing of the tweets. Together We have completed the Feature Extraction And Training the Model. Final Result (finding the accuracy of the Model and graph plotting) is done by Yash and Rajat. Stuti and Pinaki has worked on writing Report.

4 Result

After training the classifier on training set (word_feature list), it was run on validation set (labelled tweets) and the accuracy of the classifier is estimated through Naive Bayes algorithm. The output of the validation is matched with the already labelled tags of the tweets obtained by using TextBlob library. The accuracy result which came out was approximately 72% for the classifier model. We have also shown the graphical representation in which we have plotted a pie chart showing the percentages of number of tweets which have positive, negative and neutral sentiment respectively for the month of April and September.

5 Future Scope

We would like to work on other classifier models to obtain better accuracy of the data and get better results in the form of more accurate representation of public sentiments. Instead of classifying a tweet into 3 categories like positive, negative or neutral, we can work further to categorise the tweets into more classes of expression like sad, angry, happy, etc.

6 References

- Harrag, Fouzi Alsalman, Abdulmalik Alqahtani, Alaa. (2019). Prediction of Reviews Rating: A Survey of Methods, Techniques and Hybrid Architectures. Journal of Digital Information Management. 17. 164. 10.6025/jdim/2019/17/3/164-178.

- P. P. Surya, L. V. Seetha and B. Subbulakshmi, ”Analysis of user emotions and opinion using Multinomial Naive Bayes Classifier,” 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 410-415, doi: 10.1109/ICECA.2019.8822096.

- Bao, Y., Quan, C., Wang, L., Ren, F.: The role of pre-processing in twitter sentiment analysis. In: International Conference on Intelligent Computing. pp. 615–624.Springer (2014)

- https://doi.org/10.1007/s10489-020-02029-z