JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Social Media Sentiment Analysis on COVID-19

Archana Bhusara*,
L.D.College of Engineering, Ahmedabad, India
archana.bhusara84@gmail.com

Prachi Pancholi
Assistant Professor , L.D.College of Engineering, Ahmedabad, India
prachipancholi@ldce.ac.in

*Corresponding Author: Archana Bhusara, archana.bhusara84@gmail.com ,

ORCID ID: 0000-0001-9705-9583

Abstract

We are living in the information era, where data is growing rapidly. During critical events, people share their opinion on social media. Nowadays, the world is confronting an infectious disease called COVID-19 or Coronavirus. As social distancing and lockdown are considered the most considerate way to stay away from disease. Therefore, the usage of social media is increased, and a survey shows that during lockdown 3.3 million tweets a day reported on COVID-19. People are using social media for help and to express their opinion on the global pandemic. There are many social networking sites people use to share their opinion, but Twitter has emerged as one of the most popular social media among all other social media. We have done sentiment analysis on big data created by this social media. In this study, we used Twitter to study sentiment analysis. Our dataset contains 1,88,880 tweets on COVID-19. On this dataset, we used different feature extraction methods with the Naïve Bayes machine learning approach and get 84.63% accuracy on big data.

Index Terms: Sentiment Analysis, Opinion Mining, social media, COVID-19, Machine Learning

I. Introduction

In this new era, technology connects people no matter how far they are. And social media such as Facebook, Twitter, Instagram, etc. plays an integral part in how we interact with the world. Social media's, the main purpose is to share and receive information, data as well as the communication system of people. Nowadays, people share their daily activity, their thoughts, their reviews, or opinion on a certain issue, event, or subject. And these reviews, opinions, or sentiments are useful in business. In the current COVID-19 pandemic situation, people use social media for education and to gain information. In such critical events, people use social media to share how they feel about the situation. And such opinions are used for sentiment analysis.

There are many social media platforms for sentiment analysis but a survey shows that 550 million tweets are posted each day. And all these social media user posts contain unstructured text data and those need to be classified properly to provide any useful information.

In this research, we analyze the opinions, thoughts, and sentiments of people about the current pandemic. We use worldwide twitter data for general opinion analysis and during this pandemic what are the other factors like a political party affecting the situation. We used the tweepy python library to collect the tweets with #COVID-19, #COVIDVIRUS, #CORONA, #PANDAMIC, etc. And also used some classical datasets on covid-19. We analyze the sentiment of people using this dataset. We also analyze the tweet based on dates so we can get daily tweet sentiment analysis. For sentiment analysis, we used the machine learning approach. The result shows the total number of positive, negative, and neutral tweets. And also validate the result using test and split dataset.

II. Related Work

Social Media Sentiment Analysis is the trending research area in the text mining and NLP field. And it is a popular method that is applied in various fields. Many Researchers worked in this area to find out the sentiment analysis in various areas using different classification methods.

Tianyi et al. [1] study on analyzing the opinion of the Covid-19 pandemic on social media Sina Weibo. For sentiment analysis, 2.4 million posts were used as a set of training and testing. And classification such as SVM, Naïve Bayes, BERT (Bidirectional Encoder Representation for Transformers) has been used in the experiments. In Garcia et al.'s [2] study, they collected 3,155,277 Portuguese tweets on Coronavirus. They used different machine learning methods to classify the classes into anger, sadness, and fear: Naïve Bayes, Random Forest, SVM, Logistic Regression, and Naïve Bayes gives the better result. Parsoon Gupta et al. [5] study on COVID-19 Lockdown in India using Twitter dataset. For the study, they used 12,741 tweets. For better accuracy, they used the SVM algorithm and get 84.4% of accuracy. Lapoz-chou et al. [3] analyzed 3000 Mexican tweets' opinions on natural disasters (earthquake). And classify using Machine learning methods Naïve Bayes, SVM, Decision Tree. Veny et al. [9] used Twitter social media to study sentiment analysis on the topic Anti-LGBT. For that, they collected 3744 Indonesian tweets on the LGBT issue in Indonesia. And for classification Naïve Bayes, Random Forest, and Decision Tree are used. And Naïve Bayes gives a better result. Meylan et al. [4] present a sentiment analysis on Twitter data written in the Indonesian language. The data were classified into positive, negative, and neutral classes for classification they applied K-NN, Naïve Bayes, and SVM. Naïve Bayes gives 80.90% accuracy. Table 1 shows the summary of some previous works. In[10] paper Dhanya et al. analyze public opinion on the implemented demonetization policy in India using Machine Learning methods. A Twitter dataset with 5000 tweets on Demonetization is used. For feature extraction, the n-gram method is used. And for classification SVM, Naïve Bayes (NB), Decision Tree (DT), Linear Discriminant Analysis (LDA) methods are used. Different parameters are used to assess the performance of classifiers such as precision, recall, accuracy, F1 score. And SVM shows the maximum accuracy comparing to other machine learning methods. Mrintyunay Singh et al. [28] discuss the impact of coronavirus using the BERT model. They used Twitter data set for analysis. They used 596,784 tweets for classification. Study shows the accuracy 89% of accuracy.

III. Methodology

Machine Learning approaches are mainly used in the sentiment analysis domain. For sentiment analysis following steps are essentials:

Step 1: Examination of currently existing methods in the sentiment analysis domain.

Step 2: Datasets are collected from different social media.

Step 3: Data pre-processing includes noise removal from the text.

Step 4: Various feature extraction methods are used for extracting useful features.

Step 5: The classification method classifies sentiments.

Step 6: The results have been analyzed.

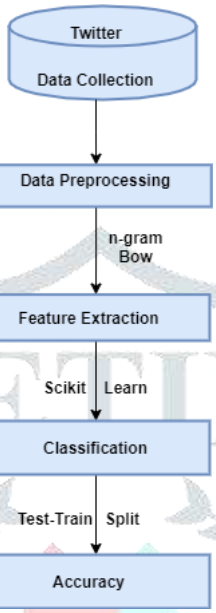

Figure 1 Workflow of Sentiment Analysis

Data Collection

For sentiment analysis on the current critical event, COVID-19 tweets are collected from Twitter. Tweets are collected using the Twitter developer account for user keys and used in tweepy the python library and save as CSV file. Tweepy library with authentication key in figure 2.

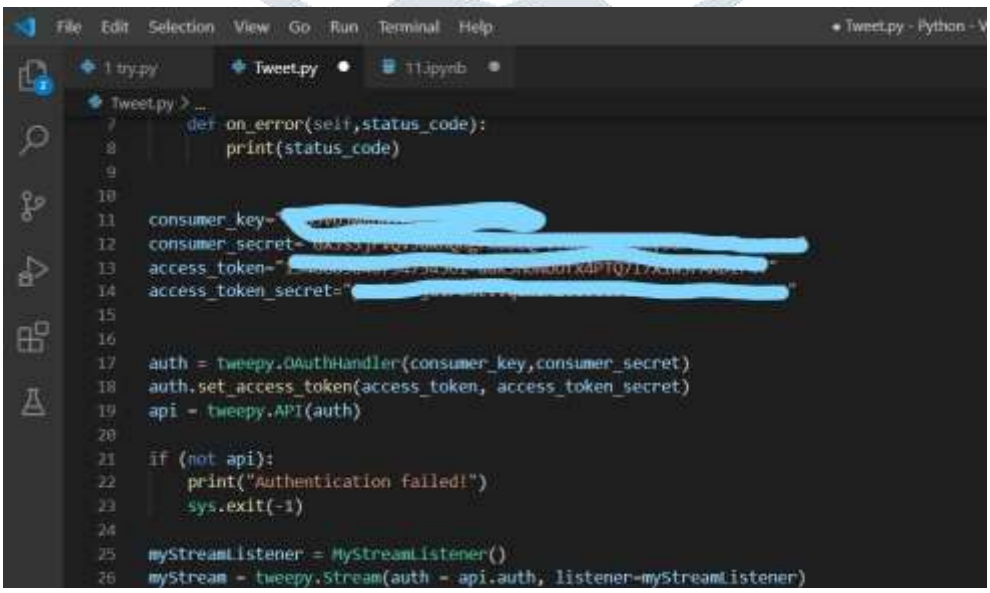

Figure 2 Tweepy- Tweet extraction

Preprocessing

In this study, we used the 188880-tweets dataset for sentiment analysis. After collecting data we have to clean the data for removing unwanted data fields. Data needs to be preprocessed. In this step unwanted noise, duplication of words, etc. are removed from the tweets. In preprocessing,

- All the tweets are converted into the lower caps lock.
- All the punctuation and URLs are removed from tweets.
- Stop words – most common words such as a, an, the, etc. are removed.
- Stemming is also done in preprocessing. Stemming is a process to extract the base form of the words by removing affixes from words. Figure 3 denotes the preprocessed text data.

Figure 3 Preprocessed data

Sentiment Score and Labelling

For Sentiment Analysis, we have to calculate the polarity or the sentiment score for labeling the tweet. And in this study, we used a big dataset so we can't label the tweets manually. Hence we used the sentiment analyzer polarity check function. We get four scores for each tweet. This score shows the positivity, negativity, and neutrality of the tweet, and the compound score is a combination of a positive, negative, and neutral score. This score describes the sentiment behind each tweet. Figure 4 shows the labeled tweet dataset.

Figure 4. Labeled data

Feature Extraction

We cannot perform any machine learning algorithms on raw text. We first need to convert the text into numbers or vector numbers.

- Bag of Words (BoW):

  This feature extraction model captures the frequencies of the word occurrences in text data. The BoW is not concern about the order in which words appear in the text, it only focuses on which words appear in the text.

  e.g., Text 1: Cats and dogs are cute.

  Text 2: Cats and dogs are not allowed.

  BoW creates a unique list of all words based on two texts.

  'cats', 'and' 'dogs', 'are ', 'cute', 'not', 'allowed'.

  So, each feature vector will be.

  Text 1: [1 1 1 1 0]

  Text 2: [1 1 1 1 0 0]

  The word present in the text is marked as 1 and the remaining as 0.

- N-gram:

  N-gram is a sequence of N-words in a sentence here n can be considered as a unigram, bigram, trigram. Unigram refers to n-gram of size 1, Bigram refers to n-gram of size 2, Trigram refers to n-gram of size 3. In N-gram ordered of the word is important.

- TF-IDF:

  TF-IDF stands for Term Frequency, Inverse Document Frequency. TF-IDF measures how important a particular word is to the entire text data set.

  TF- Term Frequency is the measure of the counts of each word in a text out of all same text datasets.

  TF(w)= (no. of times word w appears in a document) / (total no. of words in documents)

  IDF(w) = log (total amount of documents / number of documents with w in it)

  TF-IDF = TF*IDF (1)

Machine Learning Algorithm

When supervised machine learning algorithms are used for classification, the input dataset is desired to be labeled. In this study, we used Naïve Bayes and different feature extraction methods for better accuracy.

Naïve Bayes (NB) method:

Naïve Bayes is a probabilistic classifier method based on Bayes' theorem. In this study, we used the multinomial Naive Bayes classification technique. The multinomial model considers word frequency information in a document for analysis, where a document is considered to be an ordered sequence of words obtained from vocabulary 'V'. So in each document 'd_i' obtained from the multinomial distribution of word is independent of the length of d_i. N_{it} is the count of occurrence of w_t in document d_i. The probability of a document belonging to a class can be obtained using the following equation [21]:

$$P(d_i|c_j;\theta) = P(|di|)\ |d_i|!\ \prod_{t=1}^{|V|} \frac{P(W_t|C_j;\theta)^{N_{it}}}{N_{it}} \quad (2)$$

Where $P(d_i|c_j;\theta)$ refers to the probability of document 'd' belonging to the class 'c'. $P(|d_i|)$ is the probability of document d_i and $P(W_t|C_j;\theta)$ is the probability of occurrence of a word 'w' in class 'c'. After estimating the parameters calculated from the training document, the classification process is carried out on text data by calculating the posterior probability of each class and selecting the highest probable class.

IV. Experiments and Results

In this research, the experiments were conducted on large dataset using different feature extraction methods. The experimentations were conducted using the python programming language. Figure 5 shows the top source of tweets. Figure 6 shows top location of tweets. The results obtained from the experiments are shown in Table 1. Table 1 shows the comparison between the results obtained using different methods. Table 2 shows the Test and Train split validation.

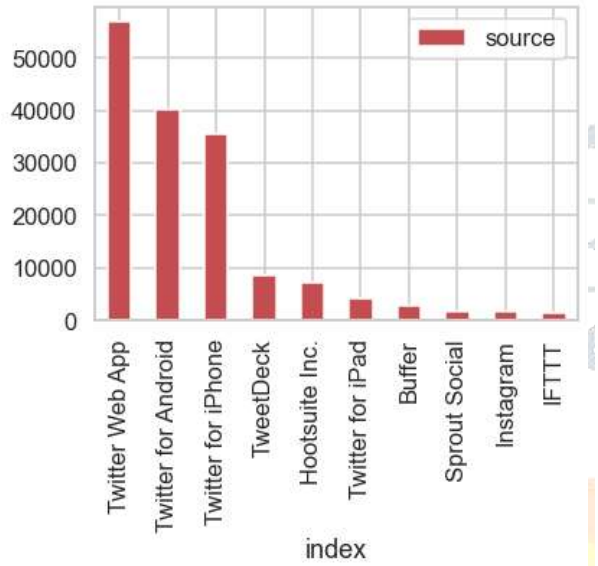

Figure 5 Sources of tweets

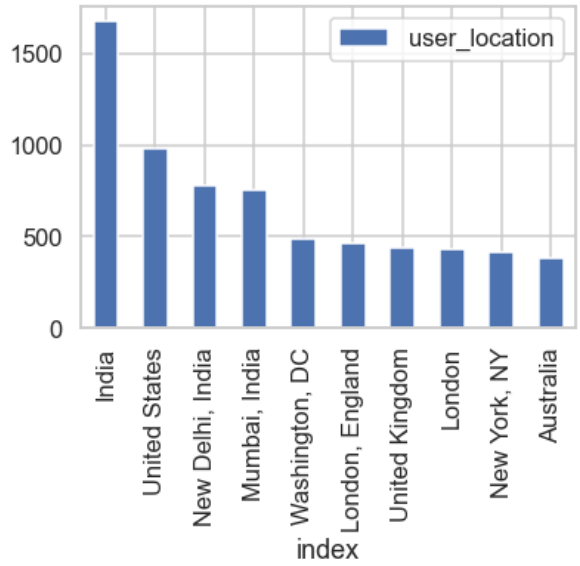

Figure 6 Location of tweets

Methods	Accuracy
Unigram+ NB	67.8%
Bigram+ NB	78.80%
Trigram+ NB	79.90%
Bag of Word+ NB	80.6%
TF-IDF+ NB	84.63%

Table 1 Comparison between different Feature Extraction methods

Train-Test Split	Naïve Bayes
80-20%	84.31%
75-25%	84.23%
50-50%	84.01%

Table 2 Test-Train Validation

V. Conclusion and Future Work

In this paper, the study focuses on the optimal extraction of features from the text document. For the research we used COVID-19 tweets which contains #COVID19, #CORONA, #CORONAVIRUS. We collect 188880 tweets for experiments. For feature extraction method we used BoW, n-gram and TFIDF and for classification we used Naïve Bayes algorithm. The Naïve Bayes (NB) supervised machine learning algorithm. The Naïve Bayes method gives the 84.63% of accuracy with TFIDF feature extraction method on large dataset. For future work we will compare Naïve Bayes algorithm with different machine learning method on large dataset.

References

[1] Tianyi wang, Ke Lu, Kam Pui chow, and Qing Zhu, "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model", IEEE-2020

[2] Klaifer Garcia, Lilian Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA", Applied Soft Computing Journal- 2020.

[3] Asdrúbal López-Chau, David Valle-Cruz, and Rodrigo Sandoval-Almazán, Sentiment Analysis of Twitter Data Through Machine Learning Techniques, Springer Switzerland,2020

[4] Meylan Wongkar, Apriandy Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter" – IEEE 2019

[5] Prasoon Gupta, Sanjay Kumar, R. R. Suman, and Vinay Kumar," Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter ", IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, 2020

[6] Mrityunjay Singh, Amit Kumar Jakhar, Shivam Pandey, " Sentiment analysis on the impact of coronavirus in social life using the BERT model", Social Network Analysis and Mining, Springer.

[7] Jim Samuel, G. G. Md. Nawaz Ali, Md. Mokhlesur Rahman, Ek Esawi and Yana Samuel, "COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification", Inf. 2020, DOI: https://doi.org/10.3390/info11060314

[8] Veny Amilia Fitri, Rachmadita Andreswari, Muhammad Azani Hasibuan, "Sentiment Analysis of Social Media Twitter with Case of AntiLGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm", Science Direct-2019

[9] Sayali Zipre, Bela Joglekar, "Polarity Shift Detection Approaches in Sentiment Analysis: A survey", IEEE-2017

[10] N. M. Dhanya, U. C. Harish," Sentiment Analysis of Twitter Data on Demonetization Using Machine Learning Techniques", Springer-2018.