As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 7(09): 13-18, 2021

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0709003
Available online at: http://www.ijmtst.com/vol7issue09.html

COVID19 Sentiment Analysis using Machine Learning Classification Algorithms

Kusumanchi Naga Sireesha[1], Padala Srinivasa Reddy[2]

[1]PG Scholar, Department of Computer Science, SVKP & Dr K S Raju Arts & Science College, Penugonda, W.G.Dt., A.P, India.
[2]Associate Professor in Computer Science, SVKP & Dr K S Raju Arts & Science College, Penugonda, W.G.Dt., A.P, India.

To Cite this Article
Kusumanchi Naga Sireesha and Padala Srinivasa Reddy. COVID19 Sentiment Analysis using Machine Learning Classification Algorithms. *International Journal for Modern Trends in Science and Technology* 2021, *7*, 0709007, pp. 13-18. https://doi.org/10.46501/IJMTST0709003

Article Info
Received: 09 August 2021; Accepted: 31 August 2021; Published: 01 September 2021

ABSTRACT

Along with the Coronavirus pandemic, another crisis has manifested itself in the form of mass fear and panic phenomena, fuelled by incomplete and often inaccurate information. There is therefore a tremendous need to address and better understand COVID-19's informational crisis. The diverse use of social networking sites, like Twitter, speeds up the process of sharing information and having views on community events and health crises COVID-19 has been one of Twitter's trending areas. The Twitter messages created via Twitter are named Tweets.

In this paper, we identify public sentiment associated with the pandemic using Coronavirus-specific Tweets and Python, along with its sentiment analysis packages. We provide an overview of two essential machine learning classification methods, in the context of textual analytics, and compare their effectiveness in classifying Coronavirus Tweets of varying lengths. This research provides insights into Coronavirus fear sentiment progression, associated methods, limitations, and different opportunities. In this project, we have designed a Sentiment analysis System that would identify the sentiment of a tweet and classify it into one of the five classes they include:"ExtremelyPositive","Positive","Neutral", "Negative" and "Extremely Negative".

KEYWORDS: covid, healthcare, nlp, machine learning, text data, tweeter, social media, sentiment analysis, text vectorization

1.INTRODUCTION

Nowadays, the Internet is becoming worldwide popular, and it is serving as a cost-effective platform for information carriers by the rapid enlargement of social media. Several social media platforms like blogs, reviews, posts, tweets are being processed for extracting the people's opinions about a particular product, organization, or situation. The attitude and feelings comprise an essential part in evaluating the behaviour of an individual that is known as sentiments. These sentiments can further be analyzed towards an entity, known as sentiment analysis or opinion mining. By using sentiment analysis, we can interpret the sentiments or emotions of others and classify them into different categories that help an organization to know people's emotions and act accordingly. This analysis depends on its expected outcomes, e.g., analyzing the text depending on its polarity and emotions, feedback about a particular feature, and analyzing the text in different languages require detection of the respective language.

It requires a large amount of data that may not be properly structured. Therefore, some preprocessing techniques are used to construct the final data set from the extracted data. Moreover, the real-time analysis helps us to look into the current scenario and make decisions to get better results.

The COVID-19 or Corona Virus has a major outbreak around the various parts of the world, and people are affected on a very large scale. The people have different views on the outbreak of the CoronaVirus. Therefore, our main focus is to do the sentiment analysis on COVID-19 to draw some conclusions on people's opinion. Recently, it has been observed that the number of people actively participated in social media like facebook, twitter, etc. However, this work uses the twitter, a social media platform, to collect the tweets with hashtag of covid19.

In the current scenario all over the world, globally, as on 1 July 2020, more than 10,268,839 confirmed cases of COVID-19 CORONAVIRUS disease including 506,064 deaths, confirmed WHO(World Health Organization) . From this data, we can conclude that this is one of the most natural virus outbreaks in the last few decades in the century. Many researchers and Health Staff analyzing these data of affected peoples from this we are getting new direction from this information and data on social networks websites. This can assist our researchers and government organizations to get advantages from these data like awareness of covid19 and be aware of thepeople's feedback and their feelings. In this, we have extracted tweets from twitter's social network platform using specified keywords which are #coronavirus, #COVID19, #covid19, #COVID19, #CORONAVIRUS, #StayHomeStaySafe, #StayHome, #StayHomeSaveLives, #Covid_19, #CovidPandemic, #covid19, #CORONA, #CoronaVirus, #Lockdown, #Quarantine, #quarantine, #CoronavirusOutbreak and #COVID. Pandas library function is used to extract the tweets from twitter. The Pandas is a library function in the python programming language. We used the jupyter Notebook which is inbuilt in the Anaconda tool.

System analysis is an important activity that takes place when we are building a new system or changing the existing one. Analysis helps to understand the existing system and the requirements necessary for building the new system. If there is no existing system then analysis defines only the requirements.

One of the most important factors in system analysis is to understand the system and its problems. A good understanding of the system enables designers to identify and correct problems. Based on the existing system the system is being planned. So the total definition problem is that the given problem has been analyzed.

2. LITERATURE REVIEW:

In this paper they have propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.[1]

Theories of children's developing understanding of mind tend to emphasize either individualistic processes of theory formation, maturation, or introspection, or the process of enculturation. However, such theories must be able to account for the accumulating evidence of the role of social interaction in the development of social understanding. We propose an alternative account, according to which the development of children's social understanding occurs within triadic interaction involving the child's experience of the world as well as communicative interaction with others about their experience and beliefs (Chapman 1991; 1999). It is through such triadic interaction that children gradually construct knowledge of the world as well as knowledge of other people. We contend that the extent and nature of the social interaction children experience will influence the development of children's social understanding. Increased opportunity to engage in cooperative social interaction and exposure to talk about mental states should facilitate the development of social understanding. We review evidence suggesting that children's understanding of mind develops gradually in the context of social interaction. Therefore,

we need a theory of development in this area that accords a fundamental role to social interaction, yet does not assume that children simply adopt socially available knowledge but rather that children construct an understanding of mind within social interaction.[2]

On 11th March 2020, World Health Organization announced COVID19 outbreak as a pandemic. Starting from China, this virus has infected and killed thousands of people from Italy, Spain, USA, Iran and other European countries as well. While this pandemic has continued to affect the lives of millions, a number of countries have resorted to complete lockdown. During this lockdown, people have taken social networks to express their feelings and find a way to calm themselves down. In this research work, country wise sentiment analysis of the tweets has been done. This research work has taken into account the tweets from twelve countries. These tweets have been gathered from 11th March 2020 to 31st March 2020, and are related to COVID19 in some or the other way. This analysis has been done to analyse how the citizens of different countries are dealing with the situation. The tweets have been collected, pre-processed, and then used for text mining and sentiment analysis. The results of the study concludes that while majority of the people throughout the world are taking a positive and hopeful approach, there are instances of fear, sadness and disgust exhibited worldwide. However, four countries, France, Switzerland, Netherland and United States of America have shown signs of distrust and anger on a bigger scale as compared to remaining eight countries.[4]

A study with a societal objective was carried out on people exchanging on social networks and more particularly on Twitter to observe their feelings on the COVID-19. A dataset of more than 600,000 tweets with hashtags like #COVID and #coronavirus posted between February 27, 2020 and March 25, 2020 was built. An exploratory treatment of the number of tweets posted by country, by language and other parameters revealed an overview of the apprehension of the pandemic around the world. A sentiment analysis was elaborated on the basis of the tweets posted in English because these constitute the great majority (USA, GB, India…). On the other hand, the FP-Growth algorithm was adapted to the tweets in order to discover the most frequent patterns and its derived association rules, in order to highlight the tweeters insights relatively to COVID-19.[6]

With the increasing availability of electronic documents and the rapid growth of the World Wide Web, the task of automatic categorization of documents became the key method for organizing the information and knowledge discovery. Proper classification of e-documents, online news, blogs, e-mails and digital libraries need text mining, machine learning and natural language processing techniques to get meaningful knowledge. The aim of this paper is to highlight the important techniques and methodologies that are employed in text documents classification, while at the same time making awareness of some of the interesting challenges that remain to be solved, focused mainly on text representation and machine learning techniques. This paper provides a review of the theory and methods of document classification and text mining, focusing on the existing literature.[7]

3. PROPOSED SYSTEM:

In this work, we use different types of classification algorithms to classify emotions on tweets. We developed models to analyse the emotional nature of various tweets, using the NLTK for emotional prediction, searching for connections between words, and marking them with positive or negative emotions .Where instead of simple positive and negative emotions, we have classified the various texts into a much more articulated class of emotional strength. They include positive, negative, strongly positive, strongly negative, and neutral. We also used different word embedding methods like bag-of-words. To overcome the drawbacks of the methods we have reviewed above, we propose a new model for sentiment analysis. In this model we combine many techniques to reach our final goal of emotion extraction. The steps for the process are documented below.

- Retrieval of data
- Pre processing
- Tweet correction

Advantages of Proposed System:

- We can predict the tweet sentiment more effectively.
- We can find not only positive or negative sentiments but also strongly positive/negative classes.

- Less computation power was required.
- Run time would be very small when compared to traditional models.
- works efficiently with large amounts of data

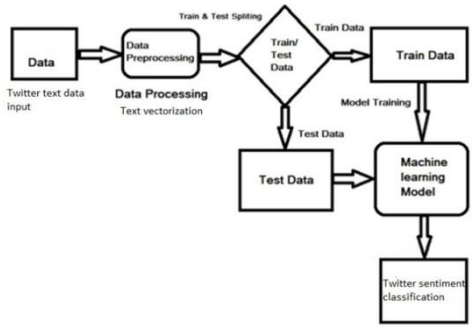

Fig1: Architecture of Proposed Work

4. METHODOLOGY:

Data Preprocessing:

Data Preprocessing is preparing the raw data and making it suitable for the machine learning model. The dataset we collected contains 82,314 tweets of Covid. The dataset contains six columns which include Username, Screen name, Location, Tweet At, Original tweet, label.

Username: The unique id of the user who had tweeted

Screen name: the unique id of the tweet

Location: The place from where it was tweeted

Tweet At: The data-month-year of the tweet

Original Tweet: the content of the tweet

Label: the sentiment of the tweet.

The pre-processing involved data cleaning, tokenization, Stemming, and removing stop words.

The data cleaning involved the following:

1. Removed @users

2. Removed HTTPS and URLs from the tweets

3. Removed punctuation, numbers, and special characters

4. Removed short words

Next, is Tokenization. This includes splitting up larger text into smaller words

. Next, it involves stemming. It is a process of reducing the inflected word to root form.

This is performed using Porter Stemmer And finally removing stop words using Nltk.corpus.

In our dataset, all six columns are not required for the model. So we only take Preprocessed Original tweets and label columns only

The dataset is divided into training and testing sets.

Training set:

It is a data set that is used to train your algorithm or model to accurately

Predict the outcome. Here we are taking 70% as training data. So the training data contains 57,619 rows.

Testing Data:

It is used to test the model after it has been trained on the initial training dataset. The testing data is 30%. So the testing set contains 24,695 rows.

Importing Libraries:

In our project, we have used three libraries. They include sklearn, seaborn, pandas, and NumPy.

Pandas: It is used for exploratory data analysis and also for reading the required data files.

Numpy: It is used for preprocessing the data

Seaborn: It is used for visualization the data (i.e tweets)

Sklearn: It is the most useful and robust library for machine learning. It is used here to import the necessary algorithms for our system.

TextVectorization using BagofWords:

Text vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers used to find word predictions and word similarities/semantics. Text Vectorization technique i.e Bag of words is a very popular choice for traditional machine learning algorithms that can help in converting text to numeric feature vectors. Each column of a vector represents a word. The values in each cell of a row show the number of occurrences of a word in a sentence. For this sklearnCountvectorizer will be used.

Tweets Example

Tweet number	Tweet
Tweet 1	Covid is dangerous
Tweet 2	Please wear a mask, when you go outside
Tweet 3	Don't panic, don't fear

Working:

1.The initial step is to find a vocabulary of unique words (ignoring the punctuation and cases). Vocabulary in the above example: [Covid , is, dangerous, please, wear, a, mask, when, you, go, outside, dont, panic, fear]

2.In our vocabulary, we have 14 unique words. Therefore, each Tweet is represented by a vector of 14 dimensions(each word representing a dimension).

3.The values corresponding to each word show the number of occurrences of a word in a Tweet.

Representation:

Tweet	Covid	is	Dangerous	please	wear	a	mask	when	you	go	outside	dont	panic	fear
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	1	1	1	1	1	1	1	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	2	1	1

Training the model:

Once vectorization is done. we need to train the model on our training dataset through which it will learn input and output patterns ..we have implemented two algorithms like logistic regression and Naive Bayes. These algorithms are imported from the sklearn library. The problem is a multi-classification problem so we have used multinomialNB(). The input for both the algorithms is a vector which is given by a bag of words. the fit() method takes training data as arguments .this takes tweet and their label as input and the model is trained on it.

Testing the model :

Now we need to test our model to know how correctly it is classifying the given inputs. Here we give the test dataset and see the output. Here we have used predict function. The predict function performs a prediction for each test instance and it usually takes a single input.

The accuracy of the model is measured using the accuracy_score() function**.** The training accuracy of the model can also be measured using the score() function. We have written a custom function that takes a tweet as an input, converts it into a vector, and then predicts the sentiment of the tweet.

Step1: We consider tweets dataset and take them as input.
Step2: We will perform Data Preprocessing on the collected dataset.
Step3: The dataset is split into training and testing dataset. According to our project, we have divided the training dataset as 70% and testing dataset as 30%.
Step4: We will use the 70% training dataset to test the 30% remaining dataset and it is given to the machine learning models Linear Regression and Naive Bayes to train itself.
Step5: After training the machine learning model will predict the new tweets sentiment as an output.

5. EXPERIMENTAL RESULTS:

The Classification algorithms are trained on the training data and whenever a new tweet is given to the model it predicts the tweet into the following five classes as Extremely Positive, Extremely Negative, Neutral, Positive, and Negative. The Logistic Regression classifier has got an accuracy of 87% and the Naive Bayes classifier has got an accuracy of 81%.

Testing Results on Logistic Regression

TWEET	RESULT
COVID Thanks for making more online shopping	Positive
HAPPY Friends know everyone uneasy about all that's going with the coronavirus but let's not	Extremely Positive
Thoughts impacts coronavirus food markets	Neutral
South Africans stock food basic goods coronavirus panic hits	Negative
Amid social distancing during COVID crisis Starbucks moves only	Extremely Negative

Testing Results on Naive Bayes

TWEET	RESULT
COVID Thanks for making more online shopping	Positive
HAPPY Friends know everyone uneasy about all that's going with the coronavirus but let's not	Extremely Positive
Thoughts impacts coronavirus food markets	Neutral
South Africans stock food basic goods coronavirus panic hits	Negative
Amid social distancing during COVID crisis Starbucks moves only	Extremely Negative

6. CONCLUSION:

A large number of the population rely on social media to update themselves, particularly on social media network platforms like Facebook, Twitter, Instagram, etc. The content circulated over social media regarding coronavirus has a direct impact on the lives of people. Sometimes it was handled positively by people and sometimes, it posted a negative impact on the daily routine. In this way, we can explore more from various textual data and tweets. We collected Twitter data based on hashtags keywords related to covid19. To maintain the credibility of data and also the ease of extracting tweets of users, the Twitter platform has been chosen for the study. For this study, the dataset of English tweets about COVID-19 was selected. 82,314 tweets were processed and analyzed. Our models will try to predict the various sentiments correctly. We have used various models for training our dataset but some models show greater accuracy while some do not. Classification Algorithms are suitable because we have to classify the tweet to predict the sentiment. Logistic Regression gave greater accuracy. Each tweet can map to one of five sentiments. Thus the model used will predict to which sentiment the tweet maps. From this, we can analyze the emotions of the people regarding the pandemic situation. Text can also represent the emotions of a person writing. From the analysis, we can infer that positive tweets are more than negative

In Future this project can be extended to understand emotions of people in a particular area or a country. For Example, How are people in London are feeling about this pandemic situation? So we can analyse such kind of questions in future by using and extending our system.

REFERENCES

[1] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean "Efficient Estimation of Word Representations in Vector Space". arXiv.org Wed, 16 Jan 2013

[2] J.I. M. Carpendale and C. Lewis ,"Constructing an understanding of mind: The development of children's social understanding within social interaction," Published online by Cambridge University Press, vol. 27, no. 1, pp. 79-96, 2004.

[3] D. S. Alehegn, "Document designed to create awereness for people covid 19," Jigdan college research and community service, Ethiopia, 2020.

[4] A. D. Dubey, "Twitter Sentiment Analysis during COVID-19 Outbreak," 2020.

[5] J. Samuel , G. G. .. N. Ali, M. M. Rahman , E. Esawi and Y. Samuel , "COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification,"2020.

[6] Y. Drias and H. Drias, "Mining Twitter Data on COVID-19 for Sentiment analysis and frequent patterns Discovery," medRxiv, 2020.

[7] Khan, B. Baharudin, L. H. Lee, and K. Khan, "A review ofmachine learning algorithms for text-documents classifica-tion,"Journal of Advances in Information Technology, vol. 1,pp. 4–20, 2010.

[8] A.Ahmad,C.Feng,A.Tahiretal.,"Anempiricalevaluationofmachine learning algorithms for identifying software re-quirements on Stack Overflow: initial Results," inProceedingsof the 10th IEEE International Conference on Software Engi-neering and Service Science (ICSESS 2019), Beijing, China,October 2019.

ABOUT AUTHORS:

Kusumanchi Naga Sireesha is currently pursuing MCA in SVKP & Dr K S Raju Arts & Science College, affiliated to Adikavi Nannaya University, Rajamahendravaram. Her research interests include Operation Research, Design and Analysis of Algorithm and Big Data Analytics.

P.Srinivasa Reddy is working as Associate Professor in SVKP & Dr K S Raju Arts & Science College, Penugonda , West Godavari District, A.P. He received Master's Degree in Computer Applications from Andhra University. His research interests include Operational Research, Probability and Statisitics, Design and Analysis of Algorithm, Big Data Analytics.