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

SENTIMENT ANALYSIS ON COVID 19 PANDEMIC USING SUPERVISED MACHINE LEARNING ALGORITHMS

[1]Dr Kommina Subhash Bhagavan, [2]Pasupuleti Mounika, [3]Nalamati Venkata Satya Pavan Sai Teja

[1]Associate Professor & HOD, [2]Student, [3]Student
Information Technology,
Sasi Institute of Technology and Engineering, Tadepalligudem, India

***Abstract*--**Social media is a source that produces large amounts of data on a large scale. Most of the people took social media platforms to share their emotions and opinions on the COVID-19 pandemic. In order to find the opinion of every people is a difficult task. Sentiment analysis is used to find the opinion in the text. In this paper we have taken the data set which contains the tweets posted by different person on this pandemic. Machine learning techniques are used to classify the sentiment of the text. Among all the techniques used in the paper. Logistic Regression classifier is best with an accuracy of 69% when compared to other models.

Keywords: Tweets, COVID-19, Sentimental Analysis, Twitter.

I. INTRODUCTION

The COVID-19 pandemic is an ongoing critical worldwide trouble all over the sector. The virus has rapidly spread over the world within a less period of time. The outbreak first got here to light in December 2019 in Wuhan, China. The (WHO)World Health Organization announced the outbreak a Public Health Emergency of International Concern on 30 January 2020, and a pandemic on 11 March 2020.COVID-19 virus infected on people and killed hundreds of thousands of people in the US (United States), Brazil, Russia, India and several other countries [1]. Since, this pandemic continues to affect millions of lives, and a number of countries have resorted to either partial or full lockdown. In today's world, the social media is everywhere, and everybody is available contact with it a day. The COVID-19 pandemic has made plenty of changes in health, business and education within the modern-day world. Now COVID-19 is increasing at very fast rate, especially in countries like the USA and India.COVID-19 has affected over 215 countries till 18th August 2020.Due to this virus, many countries have taken several decisions regarding social interactions. The COVID-19 not only disturbances their mental health, but also their physical health because of the decrease in their daily routine. Use of social websites, like Twitter, speeds up the procedure of sharing information and having views on community events and health crises [2]. Twitter data is beneficial in exposing public feelings about exciting issues and real knowledge of emerging pandemics. Within the ongoing COVID-19 pandemic, several government agencies around the worldwide use Twitter as one of the key means of contact to frequently exchange policy updates and news associated with COVID- 19 with the public [4]. For this study we are going to be considering only the Corona virus related tweets from Twitter. Analysis of all these tweets will give us a correct insight about the real emotions that the people should face during these COVID-19 times.

NLP (Natural Language Processing) is a major area to understand and analyze the human-readable text using some machine learning technique. NLTK (Natural Language Toolkit) is a most popular open-source package in Python. Rather than building all tools from scratch, NLTK provides all common NLP Tasks. The emotions were interpreted and trained using the different algorithm model to sort English and Filipino language tweets.

Before we proceed further, one should know what's mean by Sentimental Analysis. Sentimental Analysis is that the procedure of computationally identifying and categorizing opinions expressed in the form of text, especially to find out whether the actual topic is positive, neutral or negative. In this paper theclassified sentiment classes are given as fear, joy, angry, sad.

II. RELATED WORK

Rajput *et al.,* present a statistical evaluation of the tweets associated to COVID-19 posted since January 2020. They produce two types of empirical studies. The first one is related to word frequency and the second one related to the sentiments of the individual tweets. Unigram, bigram, and trigram frequencies had been modeled by a power-law distribution. The consequences had been proven with the aid of using the Sum of Square Error (SSE), R2, and Root Mean Square Error (RMSE). For good fit, model have High values of R2 and low values of SSE and RMSE. The results we obtained have the majority of the tweets related to positive polarity and only 15% of the tweets are negative.

Samuel *et al.,* point out the overall people's sentiment associated with the pandemic using COVID-19 connected Tweets using R statistical software and its sentiment analysis packages. Using descriptive analytics, authors reveal that the fear sentiment of

people about the COVID-19 reached the peak levels in the US. Additionally, they provide a summary of two primary machine learning (ML) classification algorithms-Naive Bayes and Logistic Regression, and compare the effectiveness in classifying tweets of various lengths. Accuracy of Naive Bayes is 0.9143 for shorter tweets and an accuracy of 0.5714 for extended tweets. Whereas, Accuracy of Logistic Regression is 0.7429 for shorter tweets and an accuracy of 0.52 for extended tweets.

Barkur *et al.*, analyze sentiments of Indians regarding lockdown announcements. The authors bring out the Tweets using two hashtags namely: #IndiaLockdown and #IndiafightsCorona between March 25th to March 28th, 2020. An overall of 24,000 tweets were calculated for the analysis done using R statistical software. Overall, the consequences show that Indians have taken the opposition againstCOVID-19 positively and therefore the majority are in correspondence with the govt for announcing the lockdown to flatten out the curve.

Jang *et al.*, investigate people's concerns and reactions about COVID-19 in North America, but it mainly focuses on Canada. They examine COVID-19 associated tweets using topic aspect-based sentiment analysis and modeling and explain the results with public health authority. They compare the topics discussed with the timing of execution of public health interference for COVID-19. They also examine people's sentiment about COVID-19 connected issues. Finally, they talk about how the outcomes frequently beneficial for public health agencies when scheming a policy for brand-new interventions.

Zhou *et al.*, researched the sentiment gesture of individual people surviving in the state of New South Wales (NSW)using Twitter tweets in Australia throughout the lockdown period. Additionally, this study examines the sentimental dynamics delivered by the trending topics on Twitter like lock-down, social- distancing, Australia's Job Keeper program.

III. PROPOSED MODEL

3.1 Data Collection

Dataset used in this paper is published by Kaggle under the name COVID-19 sentimental analysis. This data set consists of 4977 records of tweets posted by multiple users. The tweets have been collected in the days between dates 21st June 2020 and 20st July 2020. A dataset is nothing but a collection of data.

TABLE 1: Sample tweets in dataset.

	Unnamed	Sentiment	Text
0	3204	sad	Agree? the poor in India are treated badly
			their poors #corona….
1	1431	joy	-If only i could have spent the with this cutie #lockdown...
2	654	joy	Will nature conservation remain priority……..
3	2530	sad	Corona virus disappearing in Italy shows this to@stella…
4	2296	sad	#corona virus spread uk records lowest daily virus death toll since…

3.2 Data Preprocessing

It is a process of converting original data into structured data. In this dataset the unstructured data is nothing but usernames, links, Hashtags, URL, emoticons, Special characters, stop words, hyperlinks, etc. are removed. Then the normalization of tweets is done by using Stemming or Lemmatization.

3.2.1.Stemming just stems or removes the last characters of a word, often leads to incorrect meaning and spelling of the word. For example, in our dataset take the first text. After stemming the output will be like this:

['agre poor India treat adli poor seek live singapoor treat like citizen given free medic treatment given food daili sim card call home tell famili finr covid case treat foc hospit'] Which doesn't give proper meaning ofthe word?

3.2.2.Lemmatization is same as stemming, but the main difference is that lemmatization gives us proper meaning of the word. After lemmatization, the output will be:

['agree poor india treated badly poor seek living singapore treated like citizen given free medical treatment given food daily sim card call home tell family fine covid case treated for hospital']

By applying stemming and lemmatization the meaning of the tweet is understandable to machine which doesfurther processing.

3.3.Data Cleaning

The procedure of preparing data for analysis by removing or modifying incomplete data within a dataset is done. For this, we are using dropna function for removing the missing values and null values from the dataset. Finally, we get cleaned dataset. It is also called cleansed data.

3.4.Feature Extraction

Feature selection is based on word density. Count Vectorizer & TF-IDF Vectorizer are the two ways in which text can be converted into numbers.

3.4.1.Count Vectorizer is a way to transform a given set of strings into a frequency representation. It is very useful in understanding the kind of text by the frequency of words in it. However, it has some disadvantages like, It has lack of ability in identifying more important and less important words for analysis. It don't identify the similarity between words and the relationships between words.

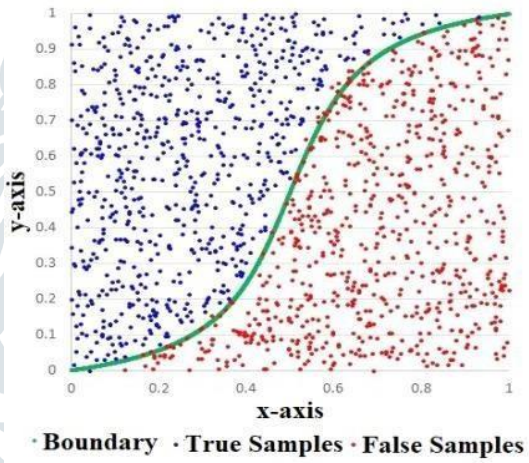
3.4.2.TF-IDF means Term Frequency-Inverse Document Frequency is a text vectorizer that covert the text into a usable vector. In TF-IDF the first technique is Term Frequency. The formula for term frequency is no of repetition of words in sentence by no of words in sentence. It tells about how important a word in a document. Inverse Document Frequency is another weight representing how common a word is across all documents. If a word is used in many documents, then the TF-IDF value would decrease. We can remove the words that are little important for analysis, hence making the model building less complex by decreasing the input dimensions.

IV. METHODOLOGY

Supervised learning requires a labelled dataset to train. In supervised learning generally contains at least two stages, i.e., train the model, and then you have to test your model [10]. During the training process the model has to train on the available trained data, where the model has to understand the patterns of the data and classify the tweets and then the model will be used to predict the class label for test data. Finally, the accuracy of the model is calculated based on the performance of model on the test data. It is of two types, namely classification and regression. Classification techniques help us find the suitable class labels which can forecast the positive, neutral and negative sentiments. The algorithms used in this methodology are Logistic Regression, Random Forest, Naive Bayes and Support Vector Machine.

4.1 Logistic Regression

Logistic regression is a supervised learning algorithm. It is used to predict the measure of a target variable. Logistic regression is used for classification. Logistic Regression basically a binary classification statement which gives us the probability of the



particular situation lies in between 0 and 1[13]. In logistic algorithm the first step is to train the data and the trained data is used to find the accuracy of the dataset by using logistic regression based on the equations.

4.2.Naive Bayes

Naive Bayes is basically a classification method. Naive Bayes is the first algorithm for solving text classification problems. It is not only known for simplicity, but also for effectiveness [14]. Using Naive Bayes algorithm you can build models fast and make quick predictions. It is a probabilistic classifier, because it predicts on the basis of the probability of an object using equation 4 [15,16].

4.3.Random Forest

Random forest is a supervised learning algorithm. It is used for both classification and regression. Random forest classifier or regressor is basically a bagging technique. In this bagging technique, decision tree is used as the base model. From the base dataset various decision tress are formed. By considering the subset of rows and columns, the decision trees get trained on the particular dataset. For aggregating the predictions of the decision trees, using the majority vote technique. So that the final decision depends on majority voting. Using multiple trees can improve the accuracy and reduce the risk of overfitting using equation (5).

$$mr(\mathbf{X}, Y) = P_{\Theta}(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} P_{\Theta}(h(\mathbf{X}, \Theta) = j) \quad (5)$$

4.4.Support Vector Machine

It is a supervised machine learning algorithm. It can be used for solving both regression and classification models[11]. However, it is mainly used for classification problems. Because in classification problems, we can easily separate two different class with a hyperplane. Apart from hyperplane, it also creates two margin lines. These two margin lines will have some distance so that it will

be linearly separable for both the classification points. This distance is called Marginal Distance using the equation (6). If the marginal distance is high, we get a more generalized model. Support vectors are nothing, but they are the points that is actually passing through the marginal plane [12].

$$bias + \sum_{s \varepsilon SV} w_s K(x^{test}, x^s) > 0 \quad \text{(6)}$$

For solving non-linear separable support vector machine uses support vector kernels. Using these kernels, it tries to convert low dimension into high dimension. It is preferred over other classification algorithms because it uses less computation and gives the best accuracy. Hence, it is used because it gives reliable results even if there is fewer data.

V. RESULTS

TABLE 1: ACCURACY PROPOSED ALGORITHMS

ALGORITHM	ACCURACY
Logistic Regression	0.695793
Naive Bayes	0.677994
Random Forest	0.631068
Support Vector Machine	0.619741

Fig. 1: Accuracy for different algorithms

TABLE II: CLASSIFICATION METRICS LR

Class	precision	recall	F1-Score
Anger	0.59	0.64	0.61
fear	0.58	0.63	0.60
Joy	0.85	0.71	0.77
Sad	0.77	0.80	0.79

Fig. 2: Precision, recall, f1-score for different algorithms

TABLE III: CLASSIFICATION METRICS NAIVE BAYES

	Precision	Recall	F1-Score
Anger	0.54	0.67	0.60
fear	0.66	0.57	0.61
joy	0.71	0.82	0.76
sad	0.81	0.69	0.74

Fig. 3: Precision, Recall, F1-score for Naive Bayes

TABLE III.CLASSIFICATION METRICS RANDOM FOREST

	Precision	Recall	F1-Score
Anger	0.64	0.57	0.60
fear	0.42	0.71	0.53
joy	0.81	0.54	0.65
sad	0.67	0.80	0.73

Fig.4: Precision, Recall, F1-score for Random Forest

TABLE IV: CLASSIFICATION METRICS SVM

	precision	recall	F1-score
Anger	0.53	0.61	0.57
Fear	0.49	0.62	0.55
Joy	0.77	0.54	0.63
Sad	0.70	0.73	0.72

Fig.5: Precision, Recall, F1-Score SVM

VI. CONCLUSION

Sentiment analysis deals with the emotions or feelings of the people towards a particular text. In this paper, a dataset is used, which contains tweets related to covid19 pandemic and classified them into four different classes named as Anger, fear, joy and sad. Supervised machine learning algorithms are implemented and trained on this dataset. Among all the models, Logistic Regression have better accuracy when compared to other machine learning algorithms like Naive Byes, Random Forest, Support Vector Machine.

VII. REFERENCES

[1]Kausar, Mohammad Abu, Arockiasamy Soosaimanickam, and Mohammad Nasar. "Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak."
[2]S. Banik, A. Ghosh, S. Banik and A. Mukherjee, "Classification of COVID19 Tweets based on Sentimental Analysis," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-7, doi: 10.1109/ICCCI50826.2021.9402540.
[3] Rufai, S. R., & Bunce, C. (2020). World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. Journal of Public Health.
[4] Rajput, Nikhil Kumar, Bhavya Ahuja Grover, and Vipin Kumar Rathi."Word frequency and sentiment analysis of twitter messages duringCoronavirus pandemic." arXiv preprint arXiv:2004.03925 (2020).
[5] samuel, Jim, et al. "Covid-19 public sentiment insights and machinelearning for tweets classification." Information 11.6 (2020): 314
[6] Barkur, Gopalkrishna, and Giridhar B. Kamath Vibha. "Sentiment anal-ysis of nationwide lockdown due to COVID 19 outbreak:Evidence fromIndia." Asian journal of psychiatry (2020).
[7] Jang, Hyeju, et al. "Exploratory analysis of covid-19 related tweetsin north america to inform public health institutes." arXiv preprintarXiv:2007.02452 (2020).
[8] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). Sentiment analysis of twitter data. In: Proceedings of the workshop on language in social media (LSM 2011) (pp. 30-38).
[9] Balahur, A. (2013). Sentiment analysis in social media texts. In: Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 120–128).
[10] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2015). Combining lexiconbased and learning- based methods for Twitter sentiment analysis. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), 89, 1–8.
[11] Chapelle, O.; Haffner, P.; Vapnik, V.N. Support Vector Machines for Histogram-Based Image Classification. IEEE Trans. Neural Netw. 2018, 10, 1055–1064.
[12] Vapnik, V. The Nature of Statistical Learning Theory; Springer: New York, NY, USA, 1995.
[13] Asri, H.; Mousannif, H.; Al Moatassime, H.; Noel, T. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Comput. Sci. 2016, 83, 1064–1069.
[14] Sisodia, D.; Sisodia, D.S. Prediction of Diabetes using Classification Algorithms. Procedia Comput. Sci. 2018, 132, 1578–1585.
[15] Rahman, A.S.; Shamrat, F.J.M.; Tasnim, Z.; Roy, J.; Hosain, S.A. A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms. Int. J. Sci. Technol. Res. 2019, 8, 419–422.
[16] Bansal, D.; Chhikara, R.; Khanna, K.; Goopta, P. Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia Detecting Dementia. Procedia Comput. Sci. 2018, 132, 1497–1502.