International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 45-51

Sentiment Analysis of Tweets in India during covid-19pandemic lockdown

Manish M Krishna[1], Akshay R Hegde[2], C R Arun[3], Dattaguru Hegde[4],Malini M Patil[5]

[1,2,3,4](Department of ISE, JSS Academy of Technical Education, Bengaluru, India) [5](Associate Professor, Department of ISE, JSS Academy of Technical Education, Bengaluru,India)

Abstract: *Sentiment analysis is a Natural Language processing technique which deals with identifying and also classifying the sentiments or opinions expressed in the text. Many Social networking sites are accumulating a large amount of sentiment data with every passing day in the form of tweets, blog posts, status updates etc. Sentiment analysis of these data which is generated by users is very useful in mining the opinions of mass crowd. Sentiment analysis of twitter data is generally difficult than normal sentiment analysis because of the presence of slangs in tweets, misspelings and also 140 is the maximum character limit in Twitter. Machine learning approach and another approach that is Knowledge base approach are the two strategies that are used for sentiment analysis from the text data. In this paper, we analyse the twitter posts which are tweeted during Covid 19 lockdown period in India by using the Machine Learning(ML) approach. Also, by using sentiment analysis for a particular domain , it is possible to find the effects of the domain information in the sentiment classification . We use the feature vector for classification of the tweets into positive, neutral or negative and then extract Indian twitter users opinions during covid lockdown in India.*

Keywords – *NLP, Sentiment Analysis , polarity , covid-19 , tweets*

I. I.Introduction

Sentiment analysis has emerged as a hot topic in the domain of natural language processing as the use of the Internet has grown rapidly (NLP). The text's implied emotion can be mined efficiently for various situations using sentiment analysis. Social media is being used by people to receive and send information during COVID-19 and disseminate various forms of data on a vast scale. Using such content to analyse people's emotions can help make key decisions about how to maintain control of the situation. The purpose of this research is to get into the feelings of Indian Twitter users during the government's nationwide lockdown to combat the spread of Coronavirus in India. In this study, NLP and machine learning classifiers will be used to analyse the sentiment of tweets produced by Indian twitter users. Information will be collected using the Tweepy API from Twitter, labeled with the TextBlob and VADER lexicons . We examine whether the Indian twitter users were positive , negative or neutral during the covid-19 pandemic lockdown . The method of identifying the emotions behind the words is known as sentiment analysis .

Steps in NLP are Tokenization , lemmatization ,stemming , POS(Part of Speech tagging) , Named Entity recognition and finally Chunking .

The possibility of analyzing public opinion during covid-19 lockdown using tweets is the motivation for this paper .

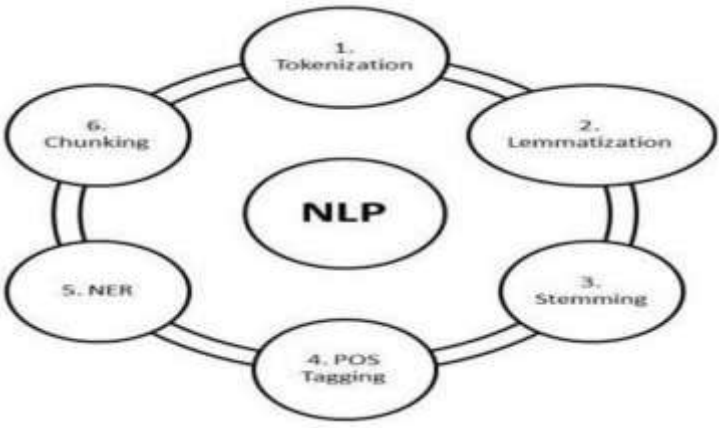

Fig 1: Steps in Natural Language processing

II. Literature Survey

One of the newest academic disciplines is sentiment analysis using social media generated data. It is more important since it plays a significant role in the health emergencies like Coronavirus pandemic. Though much research on sentiment categorization and NLP from diverse perspectives is still ongoing, the following are some of the finished works.

Wu used data on infected people transported from Wuhan from 30th December 2019 to January 29, 2020to calculate the total cases in Wuhan, China from Dec 2019 to Jan 26, 2020. The Cases that were exported inside the country was then forecasted. They used Airlines booking data and covid-19 positive patients who flew by flight to predict COVID cases across the country, and thus predicted the national and also international spread of COVID19 after finding the impact of the Wuhan and surrounding cities metropolitan wide quarantine, which began on January 23–24 in China .

Medford gathered COVID-19-related tweets and analyzed the prevalence of phrases like prevention of disease, vaccination, and racial bias. To determine sentimental valence and dominating emotions, they used sentiment analysis. They used topic modelling to identify and investigate popular debate themes over time. They culled 126 049 tweets from 53 196 distinct accounts. From January 21, 2020, the number of COVID-19-related tweets grew dramatically. Roughly half of all posts (49.5 percent) indicated dread, while about 30 percent expressedastonishment. The frequency of new occurrences of COVID-19 positives was closely matched by the quantity of racist postings. The most often discussed themes were the COVID-19's financial and political implications.

Li gathered and analyzed Weibo social networking site posts of around 17864 users through ecological recognition based on different ML prognostic models. Features like frequency of words, sentiment indicator scores like anxiety, anger and depression , and also cognitive indicators were examined in the recovered postings (such as social risk assessment and life fulfilment). To analyze differences within the same group, they employed opinion mining along with paired sample t-test for before and after the ratification of COVID-19 on 20th Jan 2020. According to their findings, sensitivity to social hazards and negative feelings have grown, while positive sentiments and life pleasure ratings have declined.

Pandey bridged the information gap and reduced the danger of disinformation by creating a lifelong learning approach that gives accurate information in local languages like Hindi, India's most widely spoken language. Using ML and NLP , they linked the sources of legitimate and providers of genuine information, like WHO news. They found that the top performing combination had a Cohen's Kappa of .54 and was used in their application.

III. Methodology

This part covers the specifics of data set explanation, sentiment labelling tools and procedures, and sentiment score values.

Data set description: This repository's data collection is part of the covid-19 data set. The table below has a full description of the data. The total number of instances is 649 thousand, with six variables, all of which are English sentences.

VADER Tool: Valence Aware Dictionary for Sentiment Reasoning is a sentiment analysis tool that can predict emotional state with high accuracy. C.J. Hutto invented the Vader lexicon, which is based on rule-based techniques. Because it is built on a lexicon with a standard sentiment library, it does not require any training. Textblob is used exclusively by researchers for context-based sentiment analysis of social media datasets that use photographs as context sources. Textblob is a library based on NLTK and patterns that is used in the Python programming languageto perform standard text processing functions.

TextBlob library: It is Natural Language Processing (NLP) package in python . Natural Language Toolkit is used extensively by Textblob to achieve its objectives efficiently . NLTK is a library which allows the users to work withcategorisation, clasification, and a veriety of other tasks by providing easy access to a huge number of lexical resources. Textblob is a package which allows for extensive textual data analysis and operations. A sentiment is determined by its semantic directions and also the intensity of each word in the phrase in the lexicon based techniques. Hence this necessitates the use of a pre-defined vocabulary that categorises negative and positive words. A text message is often expressed by a collection of words. Following the assignment of individual scores to all of the words, the final sentiment is derived using a pooling technique such as the average of all of the sentiments.

Experimental Procedure: Input: dataset: Covid-19 related tweetsResult: Sentiment labels [as positive, negative and neutral]
Step1: Input the data set
Data source is in the tab separated value format of table with six attributes, a method of data frame is used in'pandas' package to handle the table structure as frames.
Step2: Metadata validation: Metadata of table form is validated as per data dictionary.
Step3: Pre-processing of the text data is done by using NLTK modules which includes punkt libraries for stop words and also neatText libraries were also used. The intermediate results are stored as .csv file for further processing of tasks of classification.
Step4: Installations: All the necessary required package libraries compatible with python programme should be installed like TextBlob, NLTK, Scikit.
Step5: Computation: The rules are computed for calculating polarity score of the sentence. The text blob uses themodel for classification which uses support vector classifier with rbf kernel.

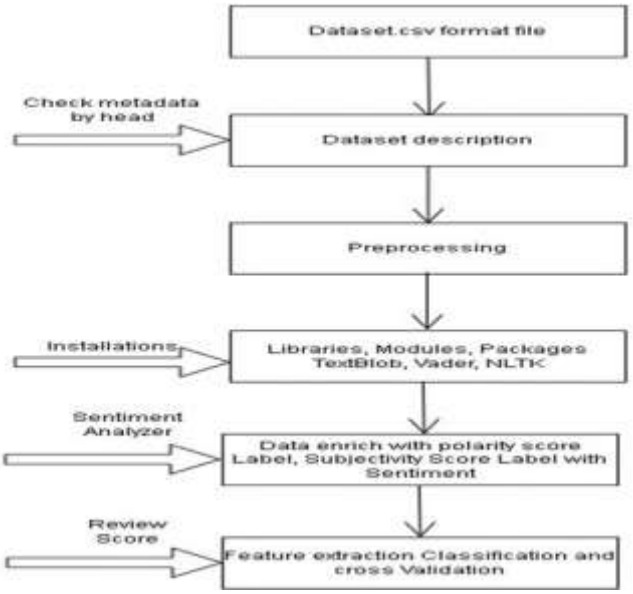

Fig 2 : Steps involved in methodology

Dataset Description:

Index	Text_id	Text	Date	Location	Sentiments
1	1266955718772559873	For the first time perhaps, someone took time to acknowledge and appreciate the efforts of the government instead of pure critisism. Good people exist! https://t.co/86qWzMFRv7	Sun May31 04:52:36 +0000 2020	India	0.4464285
2	1266955725936287746	RT @hvgoenka: 70 year Kamalamma offered an NGO in Bangalore Rs 500 from the Rs 600 monthly pension she received. She said "It is a small am…	Sun May31 04:52:38 +0000 2020	Nagpur, India	-0.25
3	1266955735537266688	Odisha state reported 129 new positive #COVID cases; taking the total cases count to 1948. Active cases stand at 889: Health Department of State	Sun May31 04:52:40 +0000 2020	New Delhi,India	0.0575757
4	1266955733297397761	@nidhiindiatv Happy rainy day .. Hope rain will wash some negative vibes of covid 19 As old saying World is on hope..	Sun May31 04:52:40 +0000 2020	Muzaffarnagar,India	0.2
5	1266955733255536640	RT @nsui: Kerala Student Union activists did Sanitization to safeguard the people from this Corona pandemic in Aluva, Ernakulam district. #…	Sun May31 04:52:40 +0000 2020	New Delhi,India	0.0

Sample sentiments:

Sl. No	**Sentence**	Polar ity	Subjectivity	Sentiment Label
1	RT Shaheen Bagh is still on Mosques are open MuIIahs are saying Coronawont harm if you read qalma Tiktokiye are m…	0.00	0.5000	Neutral
2	That son of a top official in West Bengal amp Kanika Kapoor both dodged mandatory corona screening This is all you need to know..	0.5	0.50	Positive
3	Corona has proven that Indias bigger problem is notilliteracy it is stupidity of literate people	-0.3	0.7500	Negative

IV. Result

The following results are obtained by performing the technique according to the methodology stated above using Python programming language and its libraries .

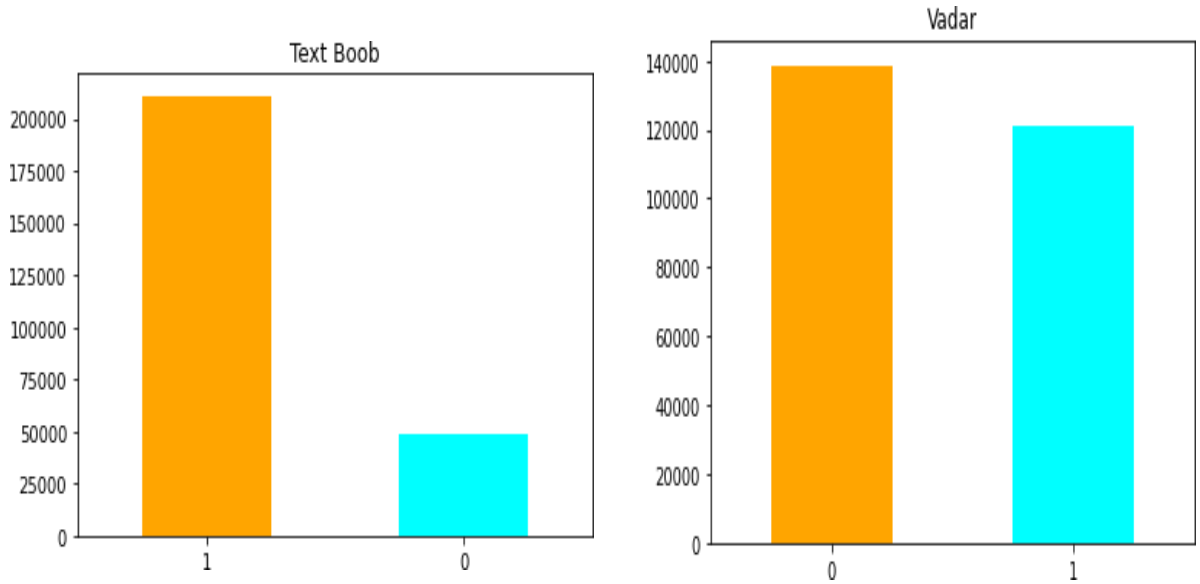

Fig 3 : Comparison of sentiment labelling using Textblob and Vader.

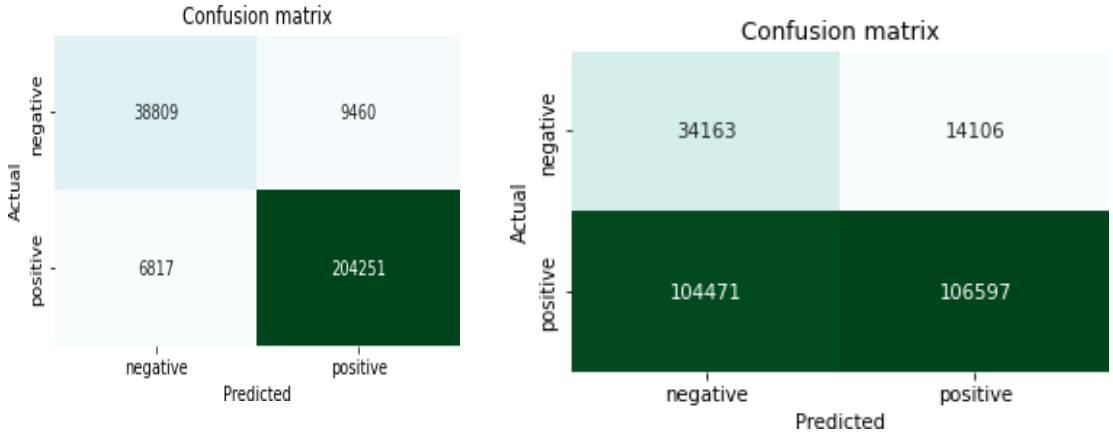

Fig 4 : Comparison of Confusion Matrix generated using textblob and Vader respectively

==========Textblob==========				
	precision	recall	f1-score	support
negative	0.85	0.80	0.83	48269
positive	0.96	0.97	0.96	211068
accuracy			0.94	259337
macro avg	0.90	0.89	0.89	259337
weighted avg	0.94	0.94	0.94	259337

Fig 5 : Classification report for Textblob

==========Vader==========				
	precision	recall	f1-score	support
negative	0.25	0.71	0.37	48269
positive	0.88	0.51	0.64	211068
accuracy			0.54	259337
macro avg	0.56	0.61	0.50	259337
weighted avg	0.76	0.54	0.59	259337

Fig 6:Classification report for Vader

Following are some of our observations:
Positive recall means that the Vader approach selects 51 percent of positive labels, but Textblob selects it 97 percent of the time it should have been selected.
Negative recall means that negative labels are selected 71% of the time, although they should have been selected 80% of the time according to Textblob.
This model is quite effective in classifying error; however, it appears to be weaker in the Vader approach of classifying positive classes, with a recall percentage of 51% compared to 97% in text blob.

Model category with confusion matrix:
The experimentation is modelled with the use of common classifier methods such as Logistic Regression, Nave Bayes, and SVM. Linear SVM and Logistic Regression have higher accuracy values.

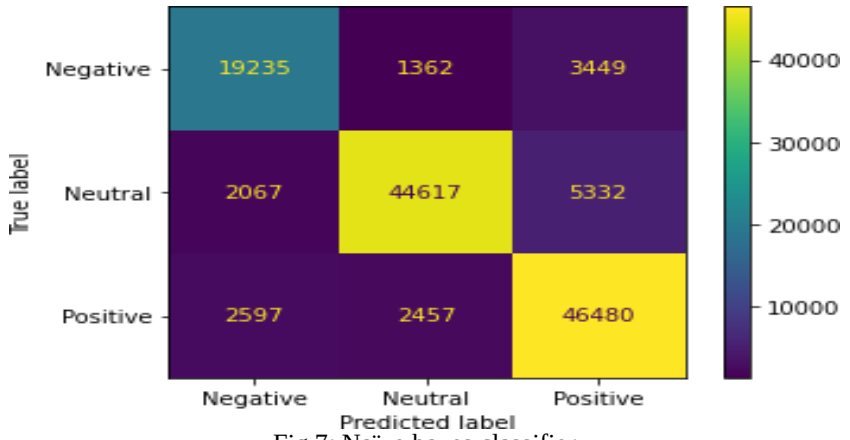

Fig 7: Naïve bayes classifier

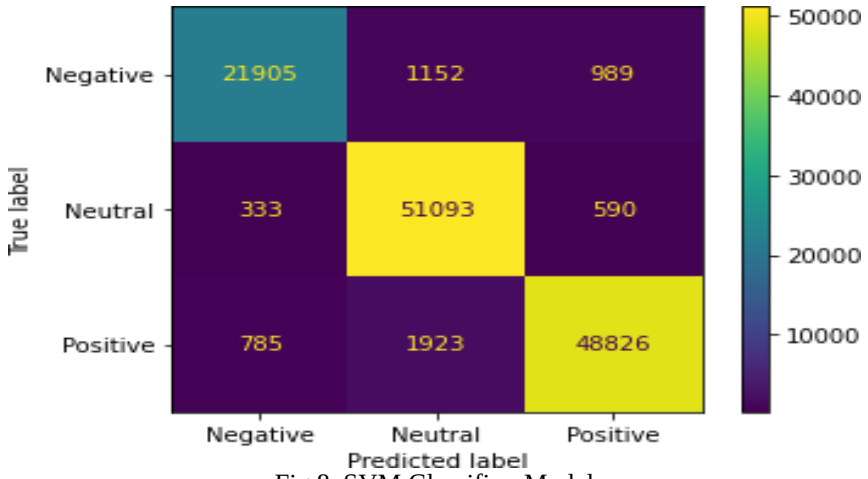

Fig 8: SVM Classifier Model

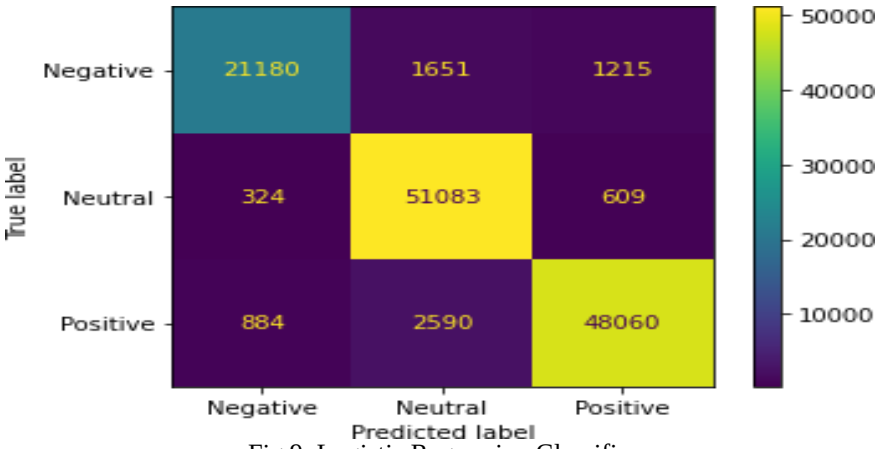

Fig 9: Logistic Regression Classifier

Accuracy score obtained using Naïve Bayes Classifier is 0.865 and Support vector machine(SVM) is 0.955 whereas by using Logistic Regression the score obtained is 0.943 .

V. Conclusion and Future Work

The number of social media users are rising exponentially every day. Nowadays People prefer to convey their emotions on social media platforms rather than sharing personally with someone else. In this paper , we have observed the general public's response during the Indian government's adoption of lockdown due to the increase in spread of COVID-19 by collecting the tweets from Twitter. We have captured tweets during any phase of India's lockdown and preprocessed it . After annotation and preprocessing, we have applied the obtained data to the supervised ML approaches with different text. The performance is consolidated by calculating aspects like precision, F1-Score, and also cross-validation for all combinations, and finally analyzing which classifier produces the best results.

Presently the model accepts only text input . This model can be further developed to accept even voice based input .

References

[1] Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet, 395(10225), 689-697.

[2] Medford, R. J., Saleh, S. N., Sumarsono, A., Perl, T. M., & Lehmann, C. U. (2020). An "Infodemic": leveraging high-volume Twitterdata to understand public sentiment for the COVID-19 outbreak. medRxiv. Preprint posted online April 7.

[3] Li, S., Wang, Y., Xue, J., Zhao, N., & Zhu, T. (2020). The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users. International journal of environmental research and public health, 17(6), 2032.

[4] Jung, S. M., Kinoshita, R., Thompson, R. N., Linton, N. M., Yang, Y., Akhmetzhanov, A. R., & Nishiura, H. (2020). Epidemiological Identification of A Novel Pathogen in Real Time: Analysis of the Atypical Pneumonia Outbreak in Wuhan, China, 2019–2020. Journal of clinical medicine, 9(3), 637.

[5] Pandey, R., Gautam, V., Pal, R., Bandhey, H., Dhingra, L. S., Misra, V., ... & Sethi, T. (2022). A machine learning application for raising wash awareness in the times of covid-19 pandemic. Scientific reports, 12(1), 1-10.

[6] Kayes, A. S. M., Islam, M. S., Watters, P. A., Ng, A., & Kayesh, H. (2020). Automated measurement of attitudes towards social distancing using social media: a COVID-19 case study.

[7] Pastor, C. K. L. (2020). Sentiment analysis on synchronous online delivery of instruction due to extreme community quarantine in thePhilippines caused by COVID-19 pandemic. Asian Journal of Multidisciplinary Studies, 3(1), 1-6.

[8] Dubey, A. D. (2020). Decoding the Twitter Sentiments towards the Leadership in the times of COVID-19: A Case of USA and India. Available at SSRN 3588623.

[9] Chen, L., Lyu, H., Yang, T., Wang, Y., & Luo, J. (2020). In the eyes of the beholder: analyzing social media use of neutral and controversial terms for COVID-19. arXiv preprint arXiv:2004.10225.

[10] Barkur, G., & Vibha, G. B. K. (2020). Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India.Asian journal of psychiatry, 51, 102089.

[11] Alhajji, M., Al Khalifah, A., Aljubran, M., & Alkhalifah, M. (2020). Sentiment analysis of tweets in Saudi Arabia regarding governmental preventive measures to contain COVID-19.

[12] Samuel, J., Ali, G. G., Rahman, M., Esawi, E., & Samuel, Y. (2020). Covid-19 public sentiment insights and machine learning for tweets classification. Information, 11(6), 314.

[13] Liu, R., Shi, Y., Ji, C., & Jia, M. (2019). A survey of sentiment analysis based ontransfer learning. IEEE Access, 7, 85401-85412.

[14] N. Kaka et al., "Digital India: Technology to transform a connected nation," McKinsey Global Inst., India, Tech. Rep., Mar. 2019.

[15] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the COVID-19 pandemic: Infoveillance study," J. Med. Internet Res., vol. 22, no. 4, Apr. 2020, Art. no. e19016.