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

NLP AND ML BASED FRAMEWORK FOR COVID-19 STATISTICAL ANALYSIS

[1]Shweta D. Mahajan, [2]Dr. Dayanand Ingle
Computer Engineering
Bharati Vidyapeeth College of Engineering
Navi Mumbai
[1]Shwetam613@gmail.com, [2]dringleus@gmail.com

Abstract: Today internet has a big number of social media sites for online learning, developing new ideas, and sharing their ideas, thoughts, opinions, views, and feelings about anything with a huge amount of people throughout the world. Nowadays it is very free to discuss your ideas and establish own business using social media platforms such as Facebook, Twitter, websites and others, and to raise one's own voice and share their opinions on a regular basis. Sentiment analysis from Twitter has recently among the largest attractive academic disciplines. For the aim of developing such systems, it blends natural language processing techniques with data mining technologies. I presented an effective technique for analysing Twitter sentiment in this research. To compare accuracy, several machine learning classification methods such as naive bayes, SVM, KNN, decision tree, logistic regression were developed for recognising positive and negative tweets.

Index Terms: **Twitter, Covid 19 vaccines, Polarity, Sentiment, Machine learning.**

I. INTRODUCTION

Twitter, Facebook, and Instagram are examples of online social media that allow users to communicate with people all over the world. They express their own thoughts about items or share personal experiences, and they even have the ability to influence politics and businesses. For example, practically every large corporation has a Twitter account to keep an eye on client comments on their services or products [3].

Sentiment Analysis is the study of people's thoughts, feelings, and attitudes, as well as their behavioural responses to certain events or episodes, using written language [18]. Surveys, reviews of various topics and arts, forums debates about prevent issues, blogs and micro-blogs about opinions and information exchange, and twitter discourse about trending are all examples of sentiment analysis. Sentiment analysis is rife with emotion or the expectation of passionate responses, and it is a form of judgement. Sentiment mining is a task that requires natural language processing (NLP) [18] and information extraction techniques to scan a large number of archives in order to compile the sentiments of various authors' remarks [1]. Computational etymology and information retrieval are among the tools used in this process. The basic concept is to take a single tweet from Twitter, calculate the polarity sentiment, and then give the value of polarity using the sentiment value, which indicates the level of the sentiment [19]. A sentiment's polarity might be classified either "positive," "negative," nor "neutral." It's important to recognize that emotion mining can be done on three different levels as follows [2, 16]:

- Document level Sentiment classifies: The polarity of the entire document is decided by whether it belongs to the "positive," "negative," or "neutral" class.
- Sentence level classifies: Each sentence is ranked according to its sentiment value for a certain topic is belongs to the "positive," "negative," or "neutral" class.
- Aspect level classifies: The feelings are based on the feelings or documents of each aspect or feature for a specific issue is corresponds to the "positive," "negative," or "neutral" class.

Sentiment analysis can be done by Machine Learning Methods. In Machine Learning there are Supervised, Semi-supervised and Unsupervised approach. I'm utilising Nave Bayes, Decision Trees, Support Vector Machines (SVM), Logistic Regression, and K-Nearest Neighbors (KNN) as supervised machine learning approaches for categorization [8, 16]

In this paper, I used sentiment analysis to categorise specific English tweets regarding covaxin or covishield using a python library called tweepy, scikit learn, and the framework flask to test as per users wishes on live tweets and analyse sentiments, as well as evaluate accuracy and performance measures as per classification methodologies, and compare them. The main objective of this paper is Sentiment analysis and opinion mining utilising user tweets can be done in a variety of ways.
The rest of the chapter is structured out as follows: section II of the overview of literature survey is presents that have taken place. Section III briefs about the existing system architecture. Section IV presented my proposed system architecture. Section V talk

about the 50000 tweets collected from twitter API dataset used for this paper. Section VI describes the proposed methodology, including the pre- processing required, feature extraction and presents the results obtained from classification techniques. Section VII paper is the concluded by what and how work can be done in future.

II. OVERVIEW OF LITERATURE SURVEY

Abdullah [1] applied naïve bayes, maximum entropy and SVM algorithm. Performed level wise document level, sentence level. In detail explain about sentence level in various approaches machine learning, ensemble methods, lexicon based and hybrid.

Radhi [2] focus on movie tweets using sentiment analysis and she is using fuzzy c-means with SVM classification worked on up to 3-grams.

Sahar A El [3] Using several testing matrices such as cross validation and f-score, data is collected on two subjects: McDonald's and KFC, in order to determine which restaurant is more popular.

Yogesh C [4] using three learning machine learning, voting based classification and deep learning classification and polarity-based classification. Those classification are compared and check accuracy. They conclude that deep learning model is better than other classification technique

M.Trupthi, S Pabboj, G.Narasimha [5] are using naïve bayes algorithm. This paper based on three level of sentiment analysis document, sentence, entity level. They are displayed two to three examples of entity level sentiment analysis using donut chart and they explain shows limitation also.

K. Chakraborty, S. Bhattacharyya [6] In this paper researcher presents a detailed survey of social networks and its related term how to apply sentiment analysis method after accumulating data from social media.

Anees Ul H, J. Hussain, M. Hussain, M. Sadiq, S. Lee [7] using naïve bayes, SVM, maximum entropy classification technique. Using sentiment ideas, machine learning, and natural language processing methods, a person's depression level may be determined by observing and extracting.

Y. Garg, N. Chatterjee [10] used naïve bayes as well as maximum entropy classifier. They explain about application in various domain and pre-processing steps by step and explain feature set in detailed.

Nikhil Yadav and team [16] naïve bayes, decision tree, SVM, XGBoost, random forest classification techniques are used. They explain pre-processing and feature extraction set using three levels document, sentence and feature level

III. EXISTING SYSTEM ARCHITECTURE

They show how to use emotion theories, machine learning techniques, and natural language processing techniques to determine a person's depression status by monitoring and extracting emotions from text. They use a binary classification technique and a multi-class sentiment classification technique for sentiment classification. Pre-processing, feature extraction, meta learning, and training data are the four main elements of technique, as displayed in Fig 1. They compared SVM, NB, and ME classifiers for phrase level sentiment analysis for depression measurement in this work [7]. They used a vote system and a feature selection method. They looked at two datasets: the twitter dataset and the 20newsgroups dataset. The results of the experiment reveal that SVM outperforms Nave Bayes and Maximum Entropy classifiers. They discovered that SVM has a 91 percent accuracy, Nave base has an 83 percent accuracy, and Maximum Entropy has an 80 percent accuracy.

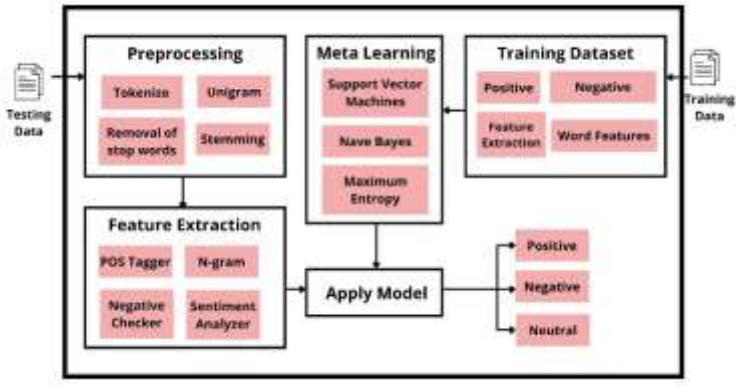

Fig 1: Existing System Architectures

IV. PROPOSED SYSTEM ARCHITECTURE

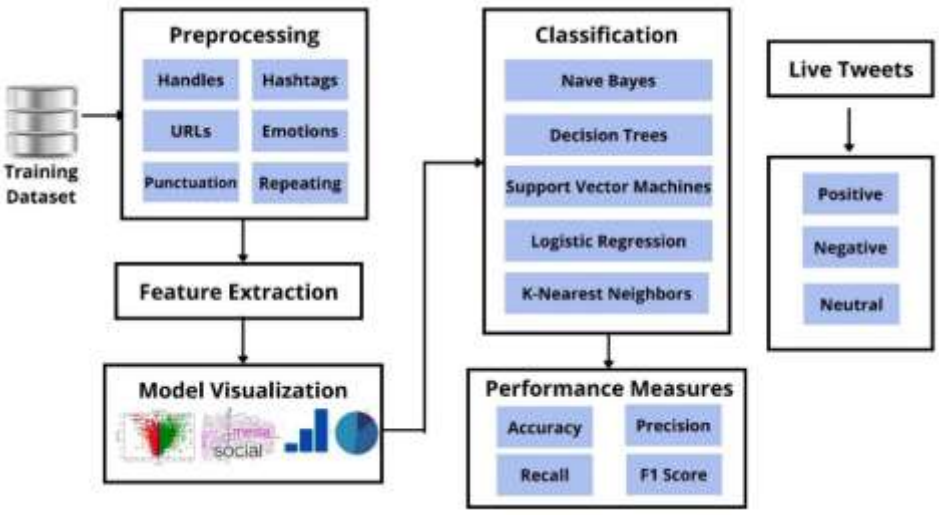

Fig 2: Proposed System Architecture

The proposed system architecture of five modules: Training Data, Pre-processing, feature exaction, classification approaches in Nave Bayes, Decision Trees, Support Vector Machines (SVM), Logistic Regression, and K Nearest Neighbours (KNN), model visualisation and check performance measures in accuracy, precision, recall and F1 score. By using a classification method, it is possible to eliminate irrelevant data and improve accuracy.

V. DATA RETRIEVING METHODOLOGY

Collecting a dataset is one of the most difficult aspects of Twitter sentiment analysis. Twitter streaming API is used to obtain the training dataset, which is used to analyse public sentiment on the two most popular in pandemic Covid 19 is a combination of the vaccines Covaxin and Covishield [10]. This is a primary data set. The tweets of the two vaccines are collected using the Twitter API and saved in a CSV file. I used the Twitter Python library to download data such as tweet content, date, and time to enhance this data. The first step is to comprehend the data, which requires collecting, describing, and exploring the data, as well as ensuring the data's quality. Because the data was acquired, it contains noise, thus I need to eliminate that noise, and data pre-processing is needed [15]. After that, I began cleaning up tweets by converting them to lowercase, removing URLS, hashtags, handles, punctuation, and repeating tweets with alphanumeric characters. Determine polarity and subjectivity for sentiment score after completing all stages. The sentiment is "Negative" if the polarity is less than zero. If the polarity is 0, the sentiment is "Neutral," otherwise it is "Positive." Then I obtained a sentiment analysis training data set.

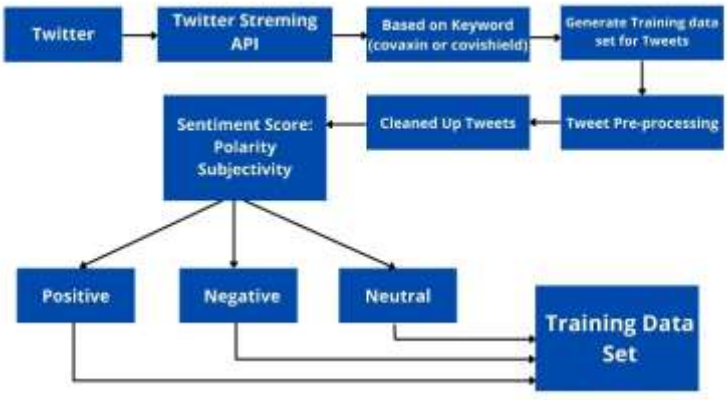

Fig 3: Data retrieving Methodology

VI. METHODOLOGY

The goal of this study is to look at tweets from the Twitter dataset in order to predict the assessment direction that those tweets will take. Fig. 3 depicts the methodology used to organise the tweets, starting with the Tweet gathering stage (dataset), Fig. 2 depicting the pre-processing phases (extraction and classification), and finally the evaluation step. The goal of this analysis is on mining English-language tweets. Taking a look at how people feel about something and what they think about it based on their tweets The dataset is created by collecting tweets and is divided into two sections: testing and training. The test data is used to construct the model's performance metric, whereas the training data is utilised to train the model. The model is able to classify newly generated tweets received from the Twitter API into positive or negative categories as a result of this. Created a web-based application for Twitter that allows users to search for live tweets on the internet and assess their sentiment.

A. Text Pre-processing:

Tweets are commonly written in colloquial language and include emoticons, hashtags, and repeated tweets, among other things Table 1. We used a number of pre-processing techniques to reduce the size of the feature set so that it suitable learning algorithms. [10, 16, 17]

a. Hashtags. A hashtag is a word or phrase that commences with the hash sign (#)
b. Handles. Every Twitter user has a distinct handle. Anything directed towards that user can be specified by using @ before their username.
c. URLs. In their tweets, users frequently include hyperlinks.
d. Emoticons: Clients frequently use emojis in their tweets to convey varied emotions. Because the number of emojis being used on social media sites is always growing, it's impossible to keep track of them all.
e. Punctuations. Although not all punctuations are significant in terms of classification, some, like as the comma and exclamation mark, might convey information about the text's thoughts.
f. Change the tweet's case to lowercase.
g. Swap two or more dots with spaces.

Tweet Before Cleaning
[22:54:09] 122001, 07-06-2021, 91 dose1 at Apollo Spectra Sheetla Hos., #COVISHIELD Sheetla Hospital Near Dronachar… https://t.co/IUKZUju6Hq
Tweet Before Cleaning
dose at apollo spectra sheetla hos covishield sheetla hospital near dronachar

Table 1: Tweet Before and After Cleaning

B. Feature Extraction:

The pre-processed dataset has a number of unique characteristics. We extract characteristics from the processed dataset using the feature extraction method [14]. Later on, this feature is utilised to compute the positive and negative polarity of a sentence, which is important for determining individual opinions using models such as unigram and bigram [9]. For machine learning algorithms to work, the key properties of text or documents must be represented. These fundamental characteristics are referred to as feature vectors, and they are used in the classification process.

C. Model Visualization:

The sentiment model was used to analyse the actions of users on social media, which is displayed in the pie chart, bar chart, and scatter plot in Fig 4. The graph and the user's scores are included in this report. The pie chart illustrates the users' overall engagements. Positive, Negative, and Neutral emotions are all present on the platform. The bar chart depicts the how much tweets are in which sentiment Fig 5. Scatter plot shows polarity and subjectivity with sentiment Fig. 6. Scatter plot shows polarity and subjectivity with sentiment. In addition, to gain a better awareness of the data, I produced word clouds using Python's "wordcloud" library. Generating a word cloud from text provides a better knowledge of the most commonly used words in tweets regarding a certain topic. Figure 7.

Fig. 7:Wordcloud for covaxin or covishield

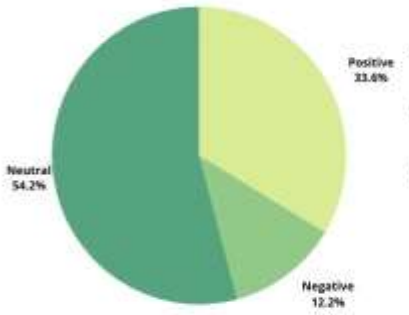

Fig 4: Pie Chart for covaxin or covishield

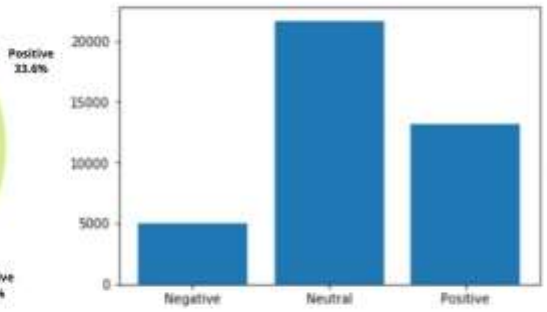

Fig 5: Bar Chart for covaxin or covishield

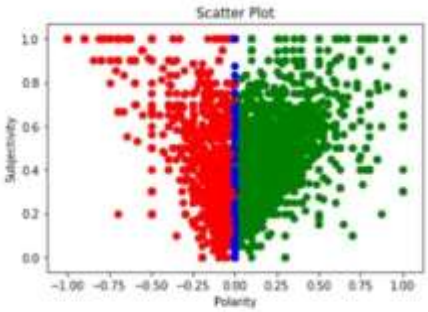

Fig 6:Scatter Plot for sentiment model

D. Classification:

Multiple supervised learning algorithms applied for the purpose of training: Naive Bayes, support vector machine (SVM), decision tree, Logistic Regression, and K-Nearest Neighbors (KNN).

i. **Naïve Bayes:** This is a classification method that is based on Bayes' Theorem and assumes high independence between features [13]. The proximity of one member in a class is expected to be unrelated to the closeness of other components, according to a Naive Bayes classifier. When it comes to categorising texts, the Naive Bayes algorithm is commonly used.
ii. **support vector machine (SVM):** SVM analyses data, characterises decision limits, and performs calculations in the input space using features. The machine then locates the boundary between the two classes, which is not visible in any of the training dataset [1]. The distance determines the classifier's margin; increasing the margin reduces indecisive decisions.
iii. **Decision Tree:** A Decision Tree is a model in which each node correspond to check on a data set attribute, and its branches represent the results. The leaf node refers to the data set's final classes. It is a supervised classifier that prepares the decision tree using data with labels and then applies the model to the test data [16].
iv. **Logistic Regression:** Logistic regression belongs to a group of classifiers known as exponential or log-linear classifiers. The probability values lie between 0 and 1 in logistic regression, which is used to predict the categorical dependent variable using a collection of independent factors [20]. Using continuous and discrete datasets, it can generate probabilities and categorise new data.
v. **K-Nearest Neighbors (KNN):** he K-NN approach saves all available data and categorises new data points depending on how similar they are to the current data. This means that utilising the K-NN approach [12], fresh data can be swiftly sorted into a well-defined category.

E. Performance Measures*:*

I used precision, recall and F1 score, support as the performance measure and compare with naïve Bayes, Decision Trees, Support Vector Machines (SVM), Logistic Regression, and K-Nearest Neighbors (KNN) Fig. 7 Precision is the ratio of the number of correct results to number of total results. The number of correct results based on number of correct results that should have been returned is known as recall. F1 score is function of precision as well as recall [3, 17].

Classifier Name	Accuracy	Precision	Recall	F1 Score
Naïve Bayes	86%	66%	80%	72%
SVM	73%	100%	24%	38%
Logistic Regression	95%	92%	77%	84%
Decision Trees	99%	85%	80%	82%
KNN	91%	90%	70%	79%

Table 2: Performance Evolution Table

VII. CONCLUSION

In this paper, the data from tweets is collected using the Twitter API and then given through machine learning classifiers. I do a survey and comparison of existing opinion mining approaches, including machine learning. According to the findings, machine learning algorithms such as nave bayes, KNN, SVM, decision trees, logistic regression, and their many combinations outperform other machine learning algorithms. For test data in the given domain, the Decision tree model achieved a very high percentage accuracy. Future improvements to this work will be limited to the English language. It may be possible to create a system that works for all languages. I used machine learning here, but I may employ Deep Learning or Hadoop in the future.

VIII. ACKNOWLEDGMENT

I would like to thank Bharati Vidyapeeth College of Engineering for providing me with the opportunity to work on this project. I had the chance to learn about sentiment analysis. I'd want to express my gratitude to my HOD, Dr. Dayanand Ingle, for obtaining the necessary approvals and helping us through the project.

REFERENCES

[1] Abdullah Alsaeedi, Mohammad Zubair khan, "A Study on sentiment analysis techniques of twitter data", International Journal of Advanced computer Science and application vol.10. No.2,2019.

[2] Radhi D. Desai, "Sentiment Analysis of Twitter Data", International Conference on Intelligent Computing and Control Systems (ICICCS 2018) IEEE Xplore Compliant Part Number: CFP18K74-ART; ISBN:978-1-5386-2842-3

[3] Sahar A. El_Rahman, Feddah Alhumaidi AlOtaibi, Wejdan Abdullah AlShehri, "Sentiment Analysis of Twitter Data", 978-1-5386-8125-1/19/$31.00 ©2019 IEEE

[4] Yogesh Chandra, Antoreep Jana, "Sentiment analysis using machine learning and deep learning", 978-93-80544-38-0 /20/$31.00_c 2020 IEEE

[5] M.Trupthi, Suresh Pabboju, G.Narasimha, "SENTIMENT ANALYSIS ON TWITTER USING STREAMING API", 2017 IEEE 7th International Advance Computing Conference 978-1-5090-1560-3/17 $31.00 © 2017 IEEE

[6] Koyel Chakraborty, Siddhartha Bhattacharyya, Rajib Bag, "A Survey of Sentiment Analysis from Social Media Data", IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS2329-924X © 2020 IEEE.

[7] Anees Ul Hassan, Jamil Hussain, Musarrat Hussain, Muhammad Sadiq, Sungyoung Lee, "Sentiment Analysis of Social Networking Sites (SNS) Data using Machine Learning Approach for the Measurement of Depression", 978-1-5090-4032-2/17/$31.00 ©2017 IEEE

[8] Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau, "Sentiment Analysis of Twitter Data", Anthology ID:W11-0705

[9] Vishal A. Kharde, S.S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016

[10] Yogesh Garg, Niladri Chatterjee, "Sentiment Analysis of Twitter Feeds", S. Srinivasa and S. Mehta (Eds.): BDA 2014, LNCS 8883, pp. 33–52, 2014. c_ Springer International Publishing Switzerland 2014

[11] Dilesh Tanna, Manasi Dudhane, Amrut Sardar," Sentiment Analysis on Social Media", 978-1-5386-8125-1/19/ ©2019 IEEE

[12] Onam Bharti, Mrs. Monika Malhotra," SENTIMENT ANALYSIS ON TWITTER DATA", International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 5, Issue. 6, June 2016, pg.601 – 609

[13] Varsha Sahayak, Vijaya Shete, Apashabi Pathan, "Sentiment Analysis on Twitter Data", International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 1, Volume 2 (January 2015)

[14] Kiran Shriniwas Doddi, Dr. Mrs. Y. V. Haribhakta, Dr. Parag Kulkarni, "Sentiment Classification of News Articles", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4621-4623 ISSN:0975-9646

[15] Ankita Sharma, Udayan Ghose, "Sentimental Analysis of Twitter Data with respect to General Elections in India", International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020, Procedia Computer Science 173 (2020) 325–334

[16] Nikhil Yadav, Omkar Kudale, Aditi Rao, Srishti Gupta, Prof. Ajitkumar Shitole, "Twitter Sentiment Analysis Using Machine Learning For Product Evaluation", Proceedings of the Fifth International Conference on Inventive Computation Technologies (ICICT-2020) IEEE Xplore Part Number:CFP20F70-ART; ISBN:978-1-7281-4685-0

[17] Bac Le, Huy Nguyen, "Twitter Sentiment Analysis Using Machine Learning Techniques", c_ Springer International Publishing Switzerland 2015 279 H.A. Le Thi et al. (eds.), Advanced Computational Methods for Knowledge Engineering, Advances in Intelligent Systems and Computing 358, DOI: 10.1007/978-3-319-17996-4_25

[18] Aliza Sarlan, Chayanit Nadam, Shuib Basri, "Twitter Sentiment Analysis", 2014 International Conference on Information Technology and Multimedia (ICIMU), 978-1-4799-5423-0/14/$31.00 ©2014 IEEE

[19] Mayur Wankhade, A Chandra Sekhara Rao, Suresh Dara, Baijnath Kaushik, "A Sentiment Analysis of Food Review using Logistic Regression", International Conference on Machine Learning and Computational Intelligence-2017, ISSN : 2456-3307

[20] N.Lalithamani1, Leela Sravanthi Thati2, Rakesh Adhikesavan3, "Sentence-Level Sentiment Polarity Calculation For Customer Reviews By Considering Complex Sentential Structures", International Journal of Research in Engineering and Technology, eISSN: 2319-1163 | pISSN: 2321-7308