IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SENTIMENT ANALYSIS OF MONKEYFOX VIRUS USING TWITTERDATA

[1]Ankita Nathe , [2]Swaroop Wajagi ,[3]Rutik Ambhore, [4]Bhushan Satpute,[5]Rohit Wasnik
[1]Assistant Professor, [2-5]Projecties
Department Of Computer Technology
Kavikulguru Institute of Technology and Science, Ramtek,
India

Abstract: Twitter is one of the most popular social media sites, and it had a tremendous surge in tweets on the monkey pox virus, including good, negative, and neutral messages, in a short amount of time. Due to the heterogeneous nature of tweets, the researchers go on to conduct sentiment analysis and assess the various feelings of the public regarding Monkeypox. In a next step, our proposed technique in this research studies Monkeypox by focusing on Twitter users who express their ideas on this social media networking site. The suggested method evaluates the emotions of collected tweets for sentiment classification by utilizing multiple feature sets and classifiers. Early identification of Monkeypox attitudes in tweets allows for a better understanding and management of the infections. Tweets are divided into three emotion categories: positive, negative, and neutral. Our suggested technique is built on three distinct machine learning models, including the Nave Bayes classifier, the Random Forest classifier, and the support vector machine classifier. Because these three classifiers have different benefits, the proposed technique effectively classifies tweets from Twitter. The results will be graphed using the matplotlib package provided by Python.

***Keywords*—** **tweets, Monkeypox, Sentiments, classifier, learning models.**

I. INTRODUCTION

Sentiment analysis is a kind of text analysis it use machinery to characterize text feelings either as positive, negative, or neutral. Utilizing Data sets for performing arts sentiment analysis can help business acquire qualitative insights into what people are talking with on the whole. Twitter has nearly 30 million active users and an average of 500 million tweets per day, make it among the most important social media platforms for news, data, and interaction with brands and important people across the world. Hence, it should come as no surprise that businesses view this micro - blogging platform as a critical tool for their marketing efforts and customer support. Twitter enables businesses to reach a large audience and connect with consumers directly. Monitoring Twitter enables businesses to understand its audience, stay on top of what is being said about the sector asa whole and its competitors, and spot emerging issues. When it comes to analyzing Twitter data, quantitative indicators like the number of mentions or retweets don't seem to be enough to get a whole picture of a situation. Understanding the significance of these mentions is what really matters. Are they exclusively or negatively discussing a particular product or subject? And sentiment analysis makes that exact determination. It gives qualitative insights on the subject or overall being discussed.

Need of sentimental analysis

1) Market Evolution: In contrast to the set of all the unstructured data, just the usable amount is needed in the industry. Nonetheless, the sentiment analysis performed is helpful for removing the key feature from the data that would be required exclusively for industry. Emotional Analysis will give businesses in various sectors a wonderful chance to grow their brands and audiences. This will be advantageous for all business-to-consumer sectors, whether they are in the restaurant, entertainment, hotel, mobile customer, retail, or travel sectors. Page Layout.

2)Studying Demand: The demand for research in evaluation, appraisal, opinion, and classification is another significant factordriving SA's expansion. Also, the research topic will be based on well- known computer science disciplines including text mining, machine learning, natural language processing, artificial intelligence, voting advice applications, automated content analysis, etc.

II. LITERATURE SURVEY

G. P. Zhang’s theory (2000)- Classification is one of the most active research and application fields for neural networks, according to G.P. Zhang's (2000) theory. Literature is abundant and expanding. The most significant advancements in neural network categorization research are outlined in this summary. in particular, the relationship between neural and conventional classifiers, the trade-off between learning and generalization, the choice of the feature variable, and the impact of misclassification costs are all studied. The goal of this review is to present an overview of the literature in this field and to pique readers' curiosity in the suggested areas of research. Machine learning, lexicon-based, and hybrid methods are the three categories under which sentiment categorization techniques are categorized. The number of followers/friends, the number of likes/shares/RTs per post, and more complicated metrics like the engagement rate, the reaction rate, and other composite metrics are typically taken into account when evaluating the buzz that has been generated. The ability to assess user opinion, however, is not a simple task. Following the acquisition of the necessary dataset, the system does sentiment analysis. It claims that there are two main sentiment analysis methodologies: machine learning and lexicon-based approach.

D.Can, S.Narayana (2012)- It was these scholars who suggested a mechanism for real-time analysis of public responses. They gather the responses from the microblogging website Twitter. Twitter is one of the social media platforms where users may express their views, ideas, and opinions on any hot topic. Twitter comments from US election candidates generated a significant amount of data that was used to gauge public opinion of each contender and predict who would win.The reactions people post on Twitter and the entire election cycle are connected in terms of feelings. They investigate sentiment analysis' impact on these public events as well. Also, they demonstrate how quick this live sentiment analysis is compared to traditional content analysis, which can take days or even weeks to complete. The technology that they six presented analyses the sentiment of all the Twitter data regarding the election, candidates, promotions, etc. and produces results continuously.In terms of feelings. They investigate sentiment analysis' impact on these public events as well. Also, they demonstrate how quick this live sentiment analysis is compared to traditional content analysis, which can take days or even weeks to complete. The technology that they six presented analyses the sentiment of all the Twitter data regarding the election, candidates, promotions, etc. and produces results continuously.

O. Almatrafi, S. Parack, B. Chavan, et al. (2014) -Authors suggested a method based on location. They claim that Sentiment Analysis involves the extraction of a sentiment from a text unit that is from a certain location using Natural Language Processing (NLP) and machine learning methods. They investigate numerous applications of location- based sentiment analysis utilizing a data source that makes it simple to gather data from multiple locations. A script may simply access the tweet location feature in Twitter, allowing data (tweets) from a specific location to be gathered for the purpose of analyzing trends and patterns. They study the 2014 general elections in India as part of their research. On 600,000 tweets that were gathered over the course of seven days for two political parties, they perform mining. They use supervised machine learning techniques, such as the Naive- Bayes algorithm, to create a classifier that can categorise tweets as positive or negative. They employ a Python module to use the views and attitudes of users towards these two political parties in various places, and they plot their findings on a map of India.

Dr. Ratnadeep R. Deshmukh (2014))-The system developed by authors to undertake a wide range of analyses of twitter datafaces new obstacles because the data distribution is inherently sparse as a result of the numerous messages that are posted every day using a diverse vocabulary. The goal of feature selection methods is to choose pertinent terms from the text that will be used for sentiment analysis. Machine learning, lexicon-based, and hybrid methods are the three categories under which sentiment categorization techniques are categorized. The system developed by author to undertake a wide range of analysis of twitter data faces new obstacles because the data distribution is inherently sparse as a result of the numerous messages that are posted every day using a diverse vocabulary. The goal of feature selection methods is to choose pertinent terms from the text that will be used for sentiment analysis. Machine learning, lexicon-based, and hybrid methods are the three categories under which sentiment categorization techniques are categorized. The number of followers/friends, the number of likes/shares/RTs per post, and more complicated metrics like the engagement rate, reaction rate, and other composite metrics are typically taken into account when evaluating the buzz that has been generated. On the other hand, it takes work to be able to analyses user opinion. It is the procedure of analyzing and separating the thoughts or feelings expressed in a specific text. The system requires keys and access to the Twitter API.

B. Sun, V. Ng, et al. (2016)- Panda Library contains data structures and tools for working with structured data sets used in many different domains, including statistics, finance, the social sciences, and many others. The library offers integrated, simple procedures for common data operations. Analysis and manipulation of large data collections. It intends to serve as the basic framework for Python's statistical computing in the future. In addition to implementing and expanding the kinds of data manipulation features present in other statistical programming languages like R, it acts as a good complement to the current scientific Python stack. Along with describing its layout and panda traits. A library called Natural Language Toolkit (NLTK) is made up of a number of software modules, a sizable collection of structured files, tutorials, problem sets, several statistics functions, classifiers for machine learning that are ready to use, computational linguistics courseware, etc.The primary objective of NLTK is to do natural language processing, or to analyses data from human language. Corpora are made available by NLTK and areused to train classifiers.

III. METHODOLOGY

A. *Overview*

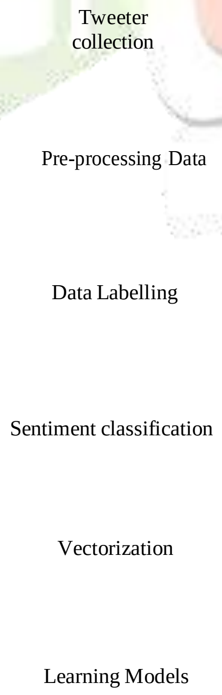
Our experimental approach as shown in figure 1.0 started with data gathering, translation, and preparation. We deleted retweets, punctuation marks, hashtags, user tags, stop words, digits, and repetitive words during pre-processing, and emoji's were converted to text.

B. *Data Collection*

Building a dataset to be processed and information derived from it requires accessing comments from social media platforms. Suggest to use the python library Tweepy which contains the API to extract the Via authenticating connection with the server. Remove retweets while gathering tweets.

C. *Data Pre-processing*

Data pre-processing is a component of natural language processing (NLP) that aids in the reconstruction of raw text into a comprehensible format. In this investigation, we employed a number of tools and processes for pre- processing. Because our dataset is multilingual, we must ensure that all of the data is in the same language. To do this, we utilized Python's built-in neattext package to translate all non-English tweets to English. Following that, all retweets, punctuation,



Hashtags, stop words, tokenization, stop words and repetitive words are eliminated. We will go over each job in the following order:

1) *User Tag Removal:* Duplicates are formed when tweets are shared, which can have a severe effect on model training and accuracy, hence retweets have to be removed. A "RT" denotes retweet, whereas a "@Someone" denotes a user tag. They were also left out.
2) *Emoji's and text conversion:* Emoji's are little digital graphics and icons that people use to communicate their thoughts and feelings. To optimize our model training, we transformed these photos into their matching textual format. To avoid double recognition of the same term, all dataset texts were changed to lowercase.
3) *Hashtag, numeral and punctuation removal: A* hashtag is a phrase used to look for and save similar information on social media. The hash sign (#) usually comes before the term, and its a strong tool in social networking. Nevertheless, it was unneeded for learning models, thus it was deleted from the dataset. Regular expressions were used to eliminate numerical, repetitive phrases, and punctuation (RegEx). This lowered memory utilization while also speeding up the learning process.
4) *Stop word removal:* Stop words are words that contribute little significant information to a phrase, such as 'to',"me,"my", 'ours,' and so on. To avoid noise in our dataset, they were deleted using the Python package library stop word.
5) *Tokenization:* Tokens are constructed by utilizing the natural language toolkit to break text into smaller parts of individual words. To facilitate feature extraction in sentiment analysis, tokenization is required.

D. *Data Labelling*

In TextBlob, we determined the label using the polarity score. There were three possible labels: positive, negative, and neutral. TextBlob is another lexicon-based sentiment analyzer (Rule-based sentiment analyzer) that we used in our study. We created a Python loop that looped over all of the rows in our datasets, and the polarity and subjectivity were retrieved using the textblob () function. A polarity score is a float between 0 and 1, and its subjectivity is also between 0 and 1. In this study, we are interested in the polarity score, which has been transformed to a label, as stated in following Equation.

$$\text{Label} = \begin{cases} \textit{Positive} \quad \textit{if} \quad 0 < score < 1 \\ \textit{Neutral} \quad \textit{if } score = = 0 \\ \textit{Negative if} \; -1 < score < 0 \end{cases}$$

After labelling the dataset, the percentage of neutral tweets is38.63%, positive tweets are 35.81%, and negative tweets are 25.54%. Its graphical representation is shown in figure 2.0.

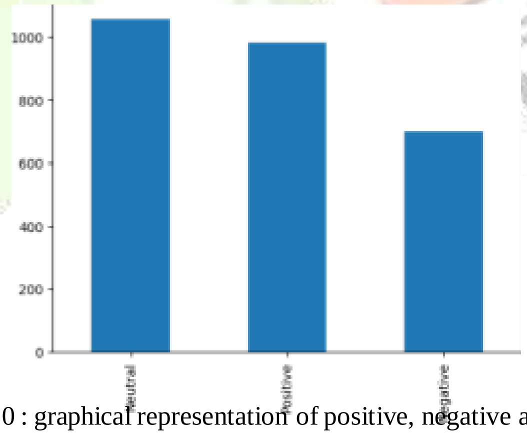


Fig 2.0 : graphical representation of positive, negative and neutral tweets

E. *Vectorization-*

Word embedding, also known as text vectorization, is a method of natural language processing that converts words or phrases from sentences into a vector of real numbers, which may then be used to identify word semantics, similarity, and prediction. It is simpler to train and extract features in machine learning when word embedding is used. in order to do sentiment analysis using vectorization, TF-IDF is employed.

1) *TF-IDF:* Term frequency (TF) and Inverse document frequency make up this method (IDF).IDF focuses on how the frequency of a word is calculated, whereas TF concentrates on the document's overall word count. The TF formulation may be seen in Equation 1.

$$TF\,(t,\,d) = \frac{\text{Frequency of term (t) in the document (d)}}{\text{Total word in the document (d)}} \qquad (1)$$

The goal of IDF is to determine how informative each word is in a document. We require IDF because it increases the effect of uncommon phrases while reducing the weight of often occurring terms. Equation 2 may be used to calculate IDF.

$$IDF\,(t) = \log_2 \left(\frac{\text{Total Documents (N)}}{\text{1+Total Documents with term (df(t))}}\right) \qquad (2)$$

TF-IDF expression on Equation 3 is the aftermath ofcombining Equations 1 & 2

$$TF - IDF = tf\,.idf\,(t,\,d,N) = tf\,(t,\,f\,).idf\,(t,N) \qquad (3)$$

F. *Learning Models-*

A number of machine learning approaches are utilised in this work to generate, develop, and evaluate numerous models. The remaining 80% of the datasets were utilised for training, whereas only 20% were used for validation. The accuracy, precision, recall, and F1 score of each model are used to evaluate its performance here. The default hyper-parameter values from sklearn were used for each of the learning algorithms. The actual algorithms are discussed further down.

1) *Random Forest:* We utilized the Random Forest model. Itis an algorithm in ensemble machine learning classification. This algorithm generates an excessive amount of categorisation decision trees, which implies that the majority of the trees choose a class category. This strategy collects the results of the tree prediction and randomises them. Node size, number of trees, and number of features sampled are the minimum three hyper parameters that must be met for random forest to function properly. The bagging ensemble technique, also known as bootstrap aggregation, is used, which generates a unique subset of training data using sample training data. The outcome is determined by the rate of preference.

2) *Support Vector Machine:* Data is sorted into one of the available categories using the robust SVM model, which provides boundaries between classes. A hyper plane is a necessary component of SVM decision boundaries (separator lines) between classes. Three hyper planes are present. Positive, negative, and ideal hyper planes are the three types of hyper planes. Equation1-3 represents these hyper planes mathematically:

$$\vec{w}.\vec{x} + b = 1 \quad \text{for Positive hyperplane} \quad (1)$$

$$\vec{w}.\vec{x} + b = -1 \quad \text{for Negative hyperplane} \quad (2)$$

$$\vec{w}.\vec{x} + b = 0 \quad \text{for optimal hyperplane} \quad (3)$$

The width of the margin is w, the bias is b, and the features are x. For the model to have the most ideal hyperplane, the margin width must be maximised. In non-linear problems, the method is good enough to solve them using the kernel, which mounts into higher dimensions, allowing them to be separated. SVM includes kernels such as polynomial, Gaussian, and Gaussian radial basis functions (RBF).

1) *Naïve Bayes:* Another method used for classification isthe Nave Bayes. It is a probabilistic classifier that determines the class likelihood of its input using conditional probability. The computation of probability for each class is defined inEquation 4.

$$\frac{\text{Conditional probability * prior probability}}{\text{Evidence}} \qquad (4)$$

Mathematically,

$$P\left(y/X\right) = \frac{P\left(X/_{y}\right)P(y)}{P(X)}$$

where:

P(y): Prior Probability
P (X/y): Likelihood probability
P (y/X): Posterior Probability
P (X): Marginal probability (Evidence).

There are many other varieties of nave Bayes; however, in this investigation, we used multinomial nave Bayes, a model that focuses on classification issues in document processing.

IV. RESULT AND DISCUSSION

This section presents the experimental steps used to evaluate the performance of the proposed models. In our study, we designed, developed, and evaluated 3 models based on the labelling, vectorization, and normalization method. Each model was trained and assessed using three machine learning algorithms: Random Forest, Support Vector Machine (SVM), and Naive Bayes Classifier.

TABLE I

Comparisons of Model's Performance

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.8946	0.8851	0.8958	0.8694
Random Forest	0.8778	0.8806	0.8648	0.87752
Naïve Bayes	0.7421	0.7928	0.7632	0.7594

Table 1 displays the results of models created by labelling with textblob and applying the TF-IDF vectorizer. According to our trial among the three algorithms. Support vector machine has the best performance, with an accuracy of 0.8946.

V. CONCLUSION

Sentimental analysis and opinion mining are now popular topics in machine learning. The purpose of sentimental analysis is to categorize texts according to the sentiments they contain. Sentimental analysis is a new field of study in computational linguistics and text mining that has received a lot of attention recently. In this project, we discuss a fundamental technique for categorising tweets into a positive or negative category using machine learning and Python. By continuing to extract more features from the tweets, we could further enhance our classifier. The Twitter API is exceptionally useful for data processing applications and may give you a lot of information about what the general public thinks.

VI. FUTURE SCOPE

Further methods and approaches for word embedding (example: doc2Vec) and text labelling (example: Azure Machine Learning) will be incorporated in future work to increase the model's performance. Furthermore, we want to apply deep learning and transformer algorithms to improve sentiment analysis and emotion prediction.

VII. REFERENCES

1) G.P Zhang (2000). "Determining the sentiment of opinions", *Proceedings of the 20th international conference on Computational Linguistics, page 1367 Association for Computational Linguistics, Stroudsburg,PA, USA.*
2) H. Wang, D. Can, F. Bar, S. Narayana et al (2012). "Twitter sentiment analysis and opinion mining", *proceeding of the Workshop on Information Extraction and Entity Analytics on Social Media Data. United Kingdom: Knowledge Media Institute*,2011
3) . O. Almatrafi, S. Parack, B. Chavan et al (2014). "Twitter as a corpus for sentiment analysis and opinion mining.", *Proceedings of the Seventh conference on International Language Resources and Evaluation. European Languages Resources Association, Valletta, Malta*
4) Manisha Mishra and Monika Srivastav (2014). "Artificial intelligence", *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL'04.
5) Dr Ratnadeep R Deshmukh (2014). "Opinion spam and analysis", *Proceedings of the 2008 International Conference on, Web Search and Data Mining*, WSDM 'ACM, New York, NY, USA, 219–230.
6) Sumneet Kaur, Aman Puri, Yashi Jain (2019). "Sentence- level sentiment polarity classification using a linguistic approachreviews", *Proceedings of the 21st, International Conference on World Wide Web, WWW* '12. ACM, New York, NY,USA, 191– 200.
7) P. Pang, L. Lee (2000). "An information theoretic approach to sentiment polarity classification". *Proceedings of the 2Nd Joint WICOW/AIRWeb Workshop on Web Quality, Web Quality* '12. ACM, New York, NY, USA, 35–40.
8) Wilson T, Wiebe J, Hoffmann P (2018). "Recognizing contextual polarity in phrase-level sentiment analysis". *Proceedings of the conference on human language technology and empirical methods in natural language processing.*

Association for Computational Linguistics, Stroudsburg, PA, USA. pp 347– 354.

9) Staphord bengesi , Timothy oladunni, Ruth olusegun andHalima audu(2023), "A Machine Learning□ SentimentAnalysis on Monkeypox Outbreak: An Extensive Data set to Show the Polarity of Public Opinion From Twitter Tweets",*https://ieeexplore.ieee.org/stamp/stamp.jsp?arnu mbe r=10036414*, volume – 11.

10) Babacar Gaye, Aziguli Wulamu (2019), "Sentimental Analysis for Online Reviews using Machine learning Algorithms", *International Journal of advance research in computer science and softwware engineering ICPT*, 34(33),46-55.

11) C. Sitaula, A. Basnet, A. Mainali, and T. B. Shahi, ''Deep learning-based methods for sentiment analysis on nepali COVID-19-related Tweets,'' *Comput. Intell. Neurosci., vol.* 2021, pp. 1–11, Nov.2021.