IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning Based Real Time Twitter Sentiment Analysis

1Mr. Mayur Doifode, 2Dr. Vidya Dhamdhere

1Student, 2Guide

INTRODUCTION

The emergence and spread of infectious diseases often lead to widespread concern and discussion on social media platforms like Twitter. Understanding public sentiment during outbreaks is crucial for health authorities and policymakers to gauge public perception, address misinformation, and implement appropriate interventions. In the case of the Monkeypox outbreak, the analysis of Twitter data using machine learning techniques for sentiment analysis can provide valuable insights into public sentiment and opinions. This study aims to conduct sentiment analysis on a comprehensive dataset collected from Twitter during the Monkeypox outbreak. By leveraging machine learning algorithms, we seek to classify tweets into different sentiment categories, such as positive, negative, or neutral, to discern the general public opinion regarding the outbreak. The sentiment analysis will involve natural language processing (NLP) techniques to understand the sentiments expressed in tweets and quantify the overall polarity of opinions. The dataset comprises tweerts collected during the outbreak period, containing keywords related to Monkeypox. Leveraging machine learning models like Support Vector Machines (SVM), Naive Bayes, or Recurrent Neural Networks (RNNs), we will preprocess the text data, perform feature extraction, and train the models to classify tweets based on sentiment.Understanding the sentiment of the public towards the Monkeypox outbreak from Twitter data can provide valuable insights into community perceptions, concerns, and the overall impact of the outbreak on social media discourse. The findings of this analysis could assist health authorities in devising targeted communication strategies, addressing public concerns, and managing the outbreak more effectively

LITERATURE SURVEY

Paper title with year: Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. 2023 [1].
Summary: In this study examines 3 sentiment computation methods (Azure Machine Learning, VADER, and TextBlob) to analyze COVID-19 vaccine hesitancy. Five learning algorithms (Random Forest, Logistics Regression, Decision Tree, LinearSVC, and Naïve Bayes) with different combination of three vectorization methods (Doc2Vec, CountVectorizer, and TF-IDF) were deployed. Vocabulary normalization was threefold; potter stemming, lemmatization, and potter stemming with lemmatization. For each vocabulary normalization strategy, we designed, developed, and evaluated 42 models. The study shows that Covid-19 vaccine hesitancy slowly decreases over time; suggesting that the public gradually feels warm and optimistic about COVID-19 vaccination.
Paper title with year:COVID-19 vaccine sentiment analysis using public opinions on Twitter. 2022 [2]
Summary: In recent months, there has been a growth in unfavorable opinion about both domestic and export vaccines in all over the globe, which is concerning. Because health authorities want to improve the adoption of COVID-19 immunizations in order to stop the epidemic, they might use sites such as twitter to spread good messages and reduce negative ones.
Paper title with year: Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm 2021 [3].
Summary: This study illustrates general peoples' sentiment towards the Pfizer, Moderna, and AstraZeneca vaccines made to fight COVIS-19. During the pandemic, people under lockdown have been expressing their

feeling on social media like Twitter about COVID-19 and its vaccines. Therefore Twitter has become an important source of information. Extracting such tweets, authors analyzed the sentiments of general people towards the vaccines. Using NLP to preprocess the raw tweets and KNN Classification Algorithm to classify the processed data, it is seen that general people have higher positive sentiment towards Pfizer and Moderna vaccine with the rate of 47.29 and 46.16 respectively compare to AstraZeneca vaccine with a rate of 40.08. This analysis can help the authority interact with the people and provide them the vaccine they trust and peacefully control the pandemic.

Paper title with year: Deep Learning-Based Methods for Sentiment Analysis on Nepali COVID-19-Related Tweets 2021 [4].

Summary: In this paper, we have proposed three CNN models to classify the Nepali COVID-19-related tweets into three sentiment classes (positive, negative, and neutral). CNN models show stable and robust performance. Also, we have proposed to use three different kinds of feature extraction methods for the representation of tweets during classification. We have validated our proposed features' extraction methods using traditional machine learning algorithms, which show that our proposed features can discriminate the complex COVID-19 tweets in most cases.

Paper title with year:A Hybrid Feature Extraction Method for Nepali COVID-19-Related Tweets Classification 2022 [5].

Summary: Analyze peoples' sentiment using both kinds of information (syntactical and semantic) on the COVID-19-related twitter dataset available in the Nepali language. For this, we, first, use two widely used text representation methods: TF-IDF and FastText and then combine them to achieve the hybrid features to capture the highly discriminating features. Second, we implement nine widely used machine learning classifiers (Logistic Regression, Support Vector Machine, Naive Bayes, K-Nearest Neighbor, Decision Trees, Random Forest, Extreme Tree classifier, AdaBoost, and Multilayer Perceptron), based on the three feature representation methods: TF-IDF, FastText, and Hybrid. To evaluate our methods, we use a publicly available Nepali-COVID-19 tweets dataset, NepCov19Tweets, which consists of Nepali tweets categorized into three classes (Positive, Negative, and Neutral).

Paper title with year: Acnn-lstm-based hybrid deep learning approach to detect sentiment polarities on monkeypox tweets. 2022 [6].

Summary: This study focuses on finding out what individuals think about monkeypox illnesses, which presents a hybrid technique based on CNN and LSTM. We have considered all three possible polarities of a user's tweet: positive, negative, and neutral. An architecture built on CNN and LSTM is utilized to determine how accurate the prediction models are. The recommended model's accuracy was 94% on the monkeypox tweet dataset. Other performance metrics such as accuracy, recall, and F1-score were utilized to test our models and results in the most time and resource-effective manner. The findings are then compared to more traditional approaches to machine learning. The findings of this research contribute to an increased awareness of the monkeypox infection in the general population.

Paper title with year: Using twitter and web news mining to predict the monkeypox outbreak 2022 [7].

Summary: The analysis and processing of social media data have revolutionized infodemiology, which helps researchers investigate human-related events accurately. Furthermore, these social networks report various statistical data such as the most comments, photos,videos, etc. about social-trend diseases like monkeypox. Therefore, this allows for predicting monkeypox morbidity rates in each area and brings awareness to health policymakers to implement educational and preventional programs in the higher-risk regions. Finally, this may help decrease the incidence of monkeypox cases and even mortality in communities.

Paper title with year: Once Bitten, Twice Shy: Our Attitude TowardsMonkeypox. 2022 [8].

Summary: It is mainly transmitted through close contact, but it can also be transmitted by droplets or aerosols. Since it is essentially a disease of wild animals, it is a zoonotic disease, but it can also be transmitted from person to person.3,5 The possibility of sexually transmitted diseases is being raised, but more verification is needed.

Paper title with year:COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis 2021 [9].

Summary: This study aims to inform policy that can be applied to social media platforms; for example, determining what degree of moderation is necessary to curtail misinformation on social media. This study also analyzes views concerning COVID-19 by focusing on people who interact and share social media on Twitter. As a platform for our experiments, we present a new large-scale sentiment data set COVIDSENTI, which

consists of 90 000 COVID-19-related tweets collected in the early stages of the pandemic, from February to March 2020.

Paper title with year: Twitter Sentiment Analysis, 2021 [10]

Summary: Provided results for sentiment analysis on Twitter. The developed unigram model was previously proposed as our baseline and we reported an overall gain for two rating tasks: binary, positive versus negative, and triple positive versus negative versus neutral. we provided a comprehensive set of experiments for each of these two tasks on manually annotated data that is a random sample of tweets. We looked at two types of models: tree kernel and feature-based models and showed that both models outperform Unigram's baseline.

Proposed System Architecture

The proposed system involves utilizing machine learning techniques for sentiment analysis on a comprehensive dataset related to the Monkeypox outbreak, sourced from Twitter.Access and gather tweets using Twitter's API based on specific keywords like "Monkeypox outbreak," "Monkeypox virus," etc. Clean the collected data by removing duplicates, irrelevant information, and noise. Label the tweets with sentiment categories (positive, negative, neutral) or a polarity score indicating the sentiment.Pre-labeled Datasets: Alternatively, use pre-labeled datasets to train the sentiment analysis model. Convert text data into a format suitable for machine learning models, such as TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (Word2Vec, GloVe). Train various machine learning models (e.g., Random Forest) using the labeled dataset to predict sentiment.Optimize models to improve accuracy and generalization.Evaluate the models using techniques like k-fold cross-validation to ensure robustness.Measure the model's performance using metrics like accuracy, precision, recall, and F1-score.

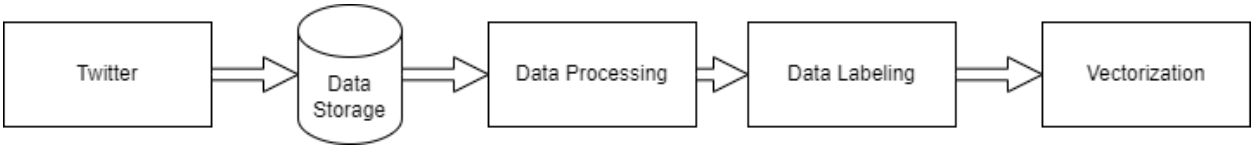

Figure : System Architecture

Details of algorithm
Random Forest

The algorithm used here is Random Forest. Random Forest is the most popular and powerful algorithm of machine learning.

Step 1: Assume N as number of training samples and M as number of variables within the classifier.

Step 2: The number m as input variables to decide the decision at each node of the tree; m should be much less than M.

Step 3: Consider training set by picking n times with replacement from all N available training samples. Use the remaining of the cases to estimate the error of the tree, by forecasting their classes.

Step 4: Randomly select m variables for each node on which to base the choice at that node. Evaluate the best split based on these m variables in the training set.

Step 5: Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier). For forecasting, a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is repeated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction. i.e. classifier having most votes.

Mathematical Model

Relevant mathematics associated with the Project : state transition diagram System Description as Let us consider S as a system.
S=A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweet

INPUT: Identify the inputs
F= f1, f2, f3 ....., FN— F as set of functions to execute commands.
I= i1, i2, i3—I sets of inputs to the function set
O= o1, o2, o3.—O Set of outputs from the function sets,
S= I, F, O I = Search Product
O = Output i.e. recommend product based on real reviews
F = Functions implemented to get the output
Space Complexity: The space complexity depends on Presentation and visualization of discovered patterns. More the storage of data more is the space complexity.
Time Complexity: Check No. of patterns available in the datasets= n If (n(1)) then retrieving of information can be time consuming.
So the time complexity of this algorithm is O (nˆn). = Failures and Success conditions.
Failures:
1. Huge database can lead to more time consumption to get the information.
2. Hardware failure.
3. Software failure.
Success:
1. Search the required information from available in Datasets.
2. User gets result very fast according to their needs.

RESULTS AND ANALYSIS

Experimental result of machine learning approach

After all basic steps like data collection & data pre- processing, feature extraction has been considered to be performed. For feature extraction, the BOW algorithm hasbeen used. The input data has been split into two parts training and testing. Training data size and testing data size are 75% and 25% respectively. Then different classification models are used and the accuracy of each model has recorded as shown below.

TABLE I. ACCURACY OF MACHINE LEARNING CLASSIFIERS

Sr no.	Classifier	Accuracy
1	SVM	92.31
2	Logistic regression	91.32
3	Random forest	92.39
4	NaÃ¯ve Bayes	88.38

CONCLUSIONS

The sentiment analysis revealed a varied distribution of sentiments regarding the Monkeypox Outbreak on Twitter. This distribution might include positive, negative, and neutral sentiments.Certain trends or patterns in public opinion might have emerged during different phases of the outbreak. These could be linked to news updates, government interventions, or other significant events related to the outbreak. Geotagging information associated with tweets could provide insights into how sentiments varied across different regions or countries affected by the outbreak.Identification of influential users or groups within the Twitter community who significantly shaped or influenced the sentiment around the Monkeypox Outbreak.Assessing the prevalence and impact of misinformation or rumors circulating on Twitter concerning the outbreak and how it affected public sentiment.

FUTURE WORK

Develop more sophisticated Machine Learning models for sentiment analysis that can account for nuanced sentiments, sarcasm, and context-specific meanings to improve accuracy.Conduct a time-series analysis to track sentiment changes over time during the outbreak, identifying spikes or drops in sentiment related to specific events or interventions.

References

[1] MiftahulQorib, Timothy Oladunni , Max Denis , Esther Ososanya,Paul Cotae,"Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset ",Elsevier,2023.

[2] P. Chinnasamy, V. Suresh, K. Ramprathap, B. Jency A. Jebamani, K. Srinivas Rao, M. Shiva Kranthi,"COVID-19 vaccine sentiment analysis using public opinions on Twitter"Elsevier,2022.

[3].F. M. JavedMehediShamrat, Sovon Chakraborty, M. M. Imran, JannatunNaeemMuna, Md. Masu,mBillah, Protiva Das, Md. Obaidur Rahman, "Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm",ijeecs,2021.

[4]. C. Sitaula , A. Basnet ,A. Mainali and T. B. Shahi,"Deep Learning-Based Methods for Sentiment Analysis on Nepali COVID-19-Related Tweet",Hindwi,2021

[5].T.B. Shahi ,C. Sitaula and N. Paudel,"A Hybrid Feature Extraction Method for Nepali COVID-19-Related Tweets Classification",2022