Asia Pacific Journal of Multidisciplinary Research
P-ISSN 2350-7756 | E-ISSN 2350-8442

Real-Time Analysis of COVID-19 Sentiments Using Twitter Streams

Ambuj Prasad Mishra[1]

ABSTRACT

Amid the COVID-19 pandemic, social media platforms, particularly Twitter, have become pivotal hubs for real-time discourse. This paper presents an exploration of Twitter stream analysis for the real-time monitoring of sentiment surrounding COVID-19 discourse. Leveraging natural language processing (NLP) techniques and machine learning algorithms, researchers have developed methodologies to extract insights from the vast stream of COVID-19-related tweets. Key components of these approaches include data preprocessing, feature extraction, sentiment classification, and evaluation metrics. Challenges such as noisy data and linguistic nuances are addressed, alongside considerations for adapting to evolving discourse trends. The implications of sentiment analysis findings are discussed, emphasizing their relevance for public health authorities, policymakers, and the general public. These insights facilitate the understanding of community perceptions, identification of misinformation, and tailoring of communication strategies. Despite advancements, opportunities remain for enhancing the accuracy and scalability of sentiment analysis systems. Future directions include integrating multimodal data sources and leveraging deep learning architectures for improved performance. This review underscores the importance of real-time sentiment analysis in navigating the complexities of the COVID-19 pandemic and informs the development of strategies for crisis management and communication in the digital age.

INTRODUCTION

Twitter is a social media platform where individuals may communicate their thoughts and opinions on current events, like the coronavirus outbreak. In terms of method, prediction, information extraction, and opinions, it is the most important streaming source of data for machine learning research. Sentiment classification is a textual analysis technique that has risen in popularity as a result of the rise of social media. Sentiment Analysis is a branch of psychology that analyses people's thoughts, feelings, and emotions derived through customer script automatically. Sentiment analysis is a hot topic in natural language processing, and it's still getting much attention in data mining because emotions are powerful drivers of social behaviour. In a way, sentiment analysis started as a research topic in Natural Language Processing across the world. The most important application of NLP, computational linguistics, and text processing is sentiment analysis. Sentiment analysis is an opinion mining task that can be used to ascertain the writer's or speaker's emotions, attitude toward a specific task, such as product reviews, film reviews, or the overall tone of the document. The era of digitalization has resulted in the exponential growth of data. Data is stored in a variety of formats, including structured, semi-structured, and unstructured. The difficult task is to discover useful information through data analysis.

PROPOSED METHODOLOGY

A form of RNN is the LSTM. The output of the last stage is used as input in the current step in RNN. This addresses the problem of RNN's long-term dependence, in that it cannot forecast data stored in long-term memory but can provide more exact predictions based on recent data. When the gap is

[1] Department of Computer Science, Madhyanchal Professional University, Bhopal (M.P.)

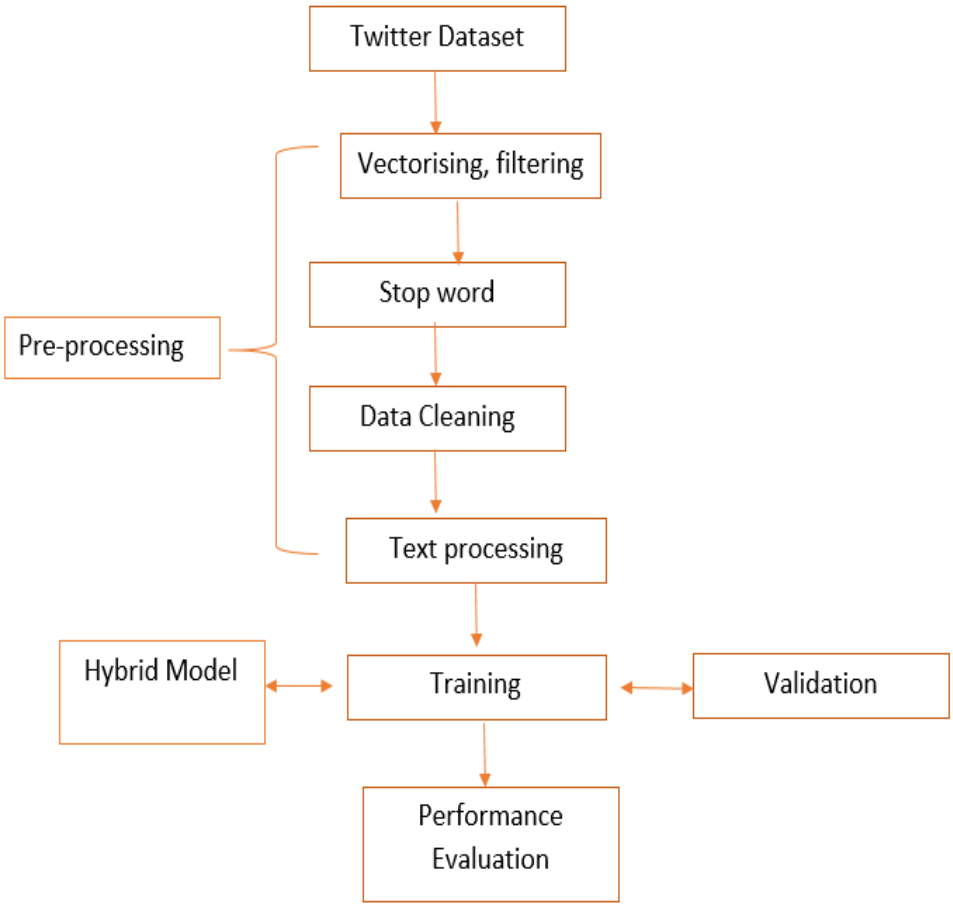
widening, RNN can no longer provide consistent readings. LSTM keeps the database for a long time if the default is used. It's utilized for time series data processing, prediction, and classification.

Long Short-Term Memory (LSTM) is a sophisticated recurrent neural network architecture that has proven to be highly effective in handling sequential data, such as time series, natural language, and speech. LSTMs are designed to address the challenges of capturing long-range dependencies and preserving context, which are essential in many real-world applications.

At the core of LSTM are three key gates: the forget gate, input gate, and output gate. These gates control the flow of information within the network, allowing it to decide what information to retain, what to store as new candidate data, and what to output as the final hidden state. The forget gate assesses the relevance of the previous memory cell state, while the input gate determines what new information is important. Together, they update the memory cell state. The output gate regulates the information passed to the next hidden state, providing the final output.

The equations governing the operations of these gates and memory cells illustrate the complex yet highly effective nature of LSTMs. By integrating these mathematical formulations, LSTMs excel in capturing intricate patterns and dependencies in data sequences, making them indispensable in various fields, including natural language processing for text generation and translation, stock market prediction, and speech recognition. LSTMs have significantly advanced the capability of deep learning models in handling sequential data, enabling the development of more accurate and context-aware solutions in an array of domains.

FLOW CHART



Asia Pacific Journal of Multidisciplinary Research
P-ISSN 2350-7756 | E-ISSN 2350-8442

FIGURE 1- A FRAMEWORK OF THE PROPOSED DEEP LEARNING MODEL

A framework for a proposed deep learning model serves as a crucial roadmap for designing and implementing the model effectively. It begins with a clear description of the input data, detailing its source, format, and any necessary preprocessing steps to ensure data quality. Subsequently, the model architecture is outlined, specifying the type and configuration of neural network layers, activation functions, regularization techniques, and the overall structure of the model.

The training process is explained in detail, covering aspects like the choice of loss function, optimizer, learning rate, and training parameters such as the number of epochs and batch size. Additionally, validation and hyperparameter tuning procedures are defined to ensure the model's optimal performance.

The framework also outlines the evaluation metrics that will be used to assess the model's effectiveness, providing a clear basis for measuring its performance on test data. It addresses how the trained model will be used for inference on new, unseen data and discusses the interpretation of its output.

The results and discussion section summarizes the outcomes of the model training and evaluation, offering insights into its strengths and limitations. Finally, the framework concludes with a summary of the proposed deep learning model and potential applications, along with suggestions for future improvements or extensions. Such a structured framework provides a solid foundation for the development of deep learning models, fostering clarity and efficiency throughout the research or development process.

RANDOM FOREST ALGORITHM

1. INTRODUCTION

Random Forest Algorithm widespread popularity stems from its user-friendly nature and adaptability, enabling it to tackle both classification and regression problems effectively. The algorithm’s strength lies in its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning.

Random forests or **random decision forests** is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set:587–588 Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing ***continuous variables,*** as in the case of regression, and ***categorical*** variables, as in the case of classification. It performs better for classification and regression tasks. In this tutorial, we will understand the working of random forest and implement random forest on a classification task.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

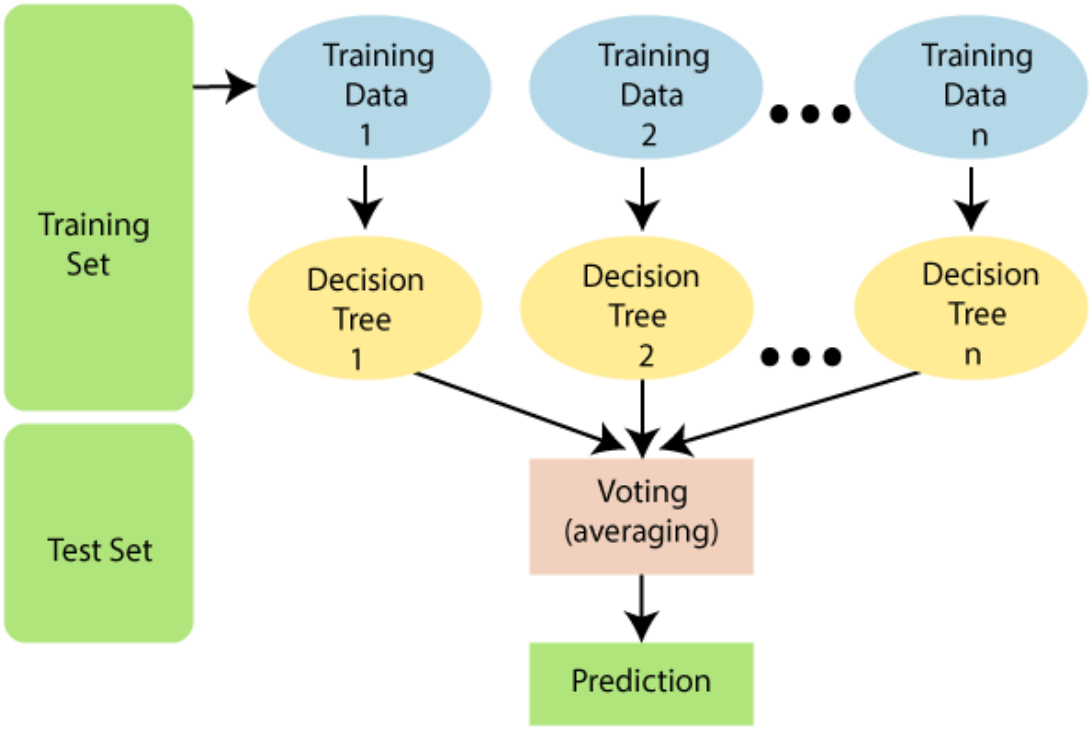
As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

THE GREATER NUMBER OF TREES IN THE FOREST LEADS TO HIGHER ACCURACY AND PREVENTS THE PROBLEM OF OVERFITTING

Random forest is a supervised learning algorithm. It has two variations – one is used for classification problems and other is used for regression problems. It is one of the most flexible and easy to use algorithm. It creates decision trees on the given data samples, gets prediction from each tree and selects the best solution by means of voting. It is also a pretty good indicator of feature importance.

Random forest algorithm combines multiple decision-trees, resulting in a forest of trees, hence the name Random Forest. In the random forest classifier, the higher the number of trees in the forest results in higher accuracy.



2. ADVANTAGES AND DISADVANTAGES OF RANDOM FOREST ALGORITHM

The advantages of Random forest algorithm are as follows:-

1. Random forest algorithm can be used to solve both classification and regression problems.

Asia Pacific Journal of Multidisciplinary Research
P-ISSN 2350-7756 | E-ISSN 2350-8442

2. It is considered as very accurate and robust model because it uses large number of decision-trees to make predictions.
3. Random forests takes the average of all the predictions made by the decision-trees, which cancels out the biases. So, it does not suffer from the overfitting problem.
4. Random forest classifier can handle the missing values. There are two ways to handle the missing values. First is to use median values to replace continuous variables and second is to compute the proximity-weighted average of missing values.
5. Random forest classifier can be used for feature selection. It means selecting the most important features out of the available features from the training dataset.

The disadvantages of Random Forest algorithm are listed below:-

1. The biggest disadvantage of random forests is its computational complexity. Random forests is very slow in making predictions because large number of decision-trees are used to make predictions. All the trees in the forest have to make a prediction for the same input and then perform voting on it. So, it is a time-consuming process.
2. The model is difficult to interpret as compared to a decision-tree, where we can easily make a prediction as compared to a decision-tree.

3. FEATURE SELECTION WITH RANDOM FORESTS

Random forests algorithm can be used for feature selection process. This algorithm can be used to rank the importance of variables in a regression or classification problem.

We measure the variable importance in a dataset by fitting the random forest algorithm to the data. During the fitting process, the out-of-bag error for each data point is recorded and averaged over the forest.

The importance of the j-th feature was measured after training. The values of the j-th feature were permuted among the training data and the out-of-bag error was again computed on this perturbed dataset. The importance score for the j-th feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences.

Features which produce large values for this score are ranked as more important than features which produce small values. Based on this score, we will choose the most important features and drop the least important ones for model building.

Some interesting facts about Random Forests – Features

- Accuracy of Random forest is generally very high
- Its efficiency is particularly Notable in Large Data sets
- Provides an estimate of important variables in classification
- Forests Generated can be saved and reused
- Unlike other models It does nt overfit with more features

4. IMPORT LIBRARIES

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt # data visualization
import seaborn as sns # statistical data visualization
%matplotlib inline
```

5. ALGORITHM STEPS

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

The Random Forest algorithm operates in two main phases, employing a combination of decision trees to improve predictive accuracy. In the first phase, the algorithm randomly selects K data points from the training dataset, a process known as bootstrapping, which ensures diversity in the training subsets. In the second phase, decision trees are constructed using these selected data points, each tree growing by recursively splitting data based on features that optimize information gain or minimize impurity. This results in a set of deep decision trees capable of capturing intricate patterns within the data. The number of decision trees, denoted as N, is a crucial parameter, with a larger N generally contributing to a more robust ensemble. The algorithm repeats the sampling and decision tree construction process N times, creating a diverse ensemble of trees, each offering a unique perspective on the data.

RESULT AND SIMULATION

UserName	ScreenName	Location	TweetAt	Sentiment
3799	48751	London	16/03/2020	Neutral
3800	48752	UK	16/03/2020	Positive
3801	48753	Vagabonds	16/03/2020	Positive
3802	48754		16/03/2020	Positive
3803	48755		16/03/2020	Extremely Negative
3804	48756	ÃœT: 36.319708,-82.363649	16/03/2020	Positive
3805	48757	35.926541,-78.753267	16/03/2020	Positive
3806	48758	Austria	16/03/2020	Neutral
3807	48759	Atlanta, GA USA	16/03/2020	Positive

3808	48760	BHAVNAGAR,GUJRAT	16/03/2020	Negative
3809	48761	Makati, Manila	16/03/2020	Neutral
3810	48762	Pitt Meadows, BC, Canada	16/03/2020	Extremely Positive
3811	48763	Horningsea	16/03/2020	Extremely Positive
3812	48764	Chicago, IL	16/03/2020	Positive
3813	48765	India	16/03/2020	Positive
3814	48766	Houston, Texas	16/03/2020	Positive
3815	48767	Saudi Arabia	16/03/2020	Neutral
3816	48768	Ontario, Canada	16/03/2020	Neutral
3817	48769	North America	16/03/2020	Extremely Positive

1. DATA SET SPECIFICATION

https://www.kaggle.com/datasets/bansodesandeep/covid19-tweets-data

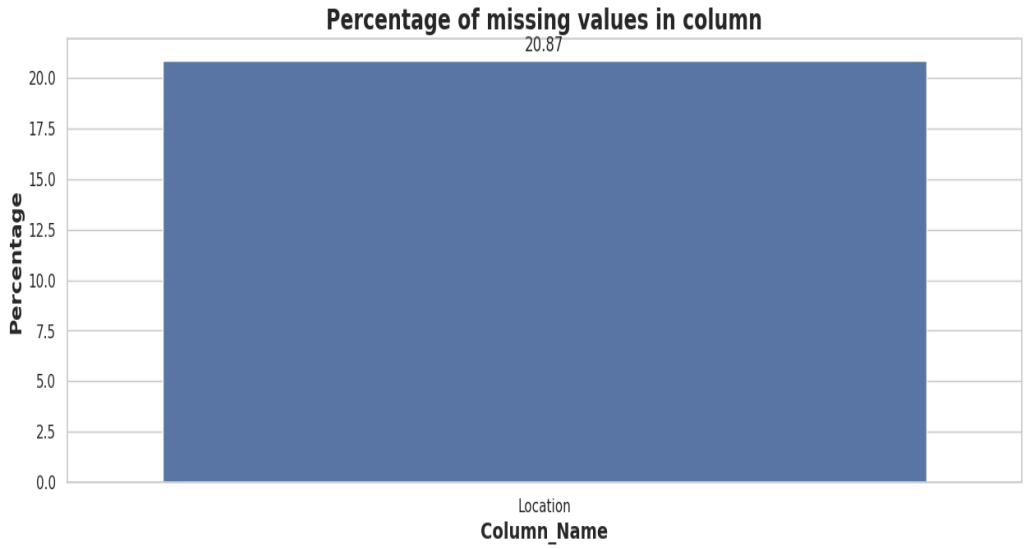
2. ALL TWEETS CAME ONLY FROM MARCH AND APRIL MONTH OF 2020

```
20-03-2020    3448
19-03-2020    3215
25-03-2020    2979
18-03-2020    2742
21-03-2020    2653
22-03-2020    2114
23-03-2020    2062
17-03-2020    1977
08-04-2020    1881
07-04-2020    1843
06-04-2020    1742
24-03-2020    1480
09-04-2020    1471
13-04-2020    1428
26-03-2020    1277
05-04-2020    1131
10-04-2020    1005
02-04-2020     954
11-04-2020     909
03-04-2020     810
12-04-2020     803
04-04-2020     767
16-03-2020     656
01-04-2020     630
27-03-2020     345
31-03-2020     316
14-04-2020     284
```

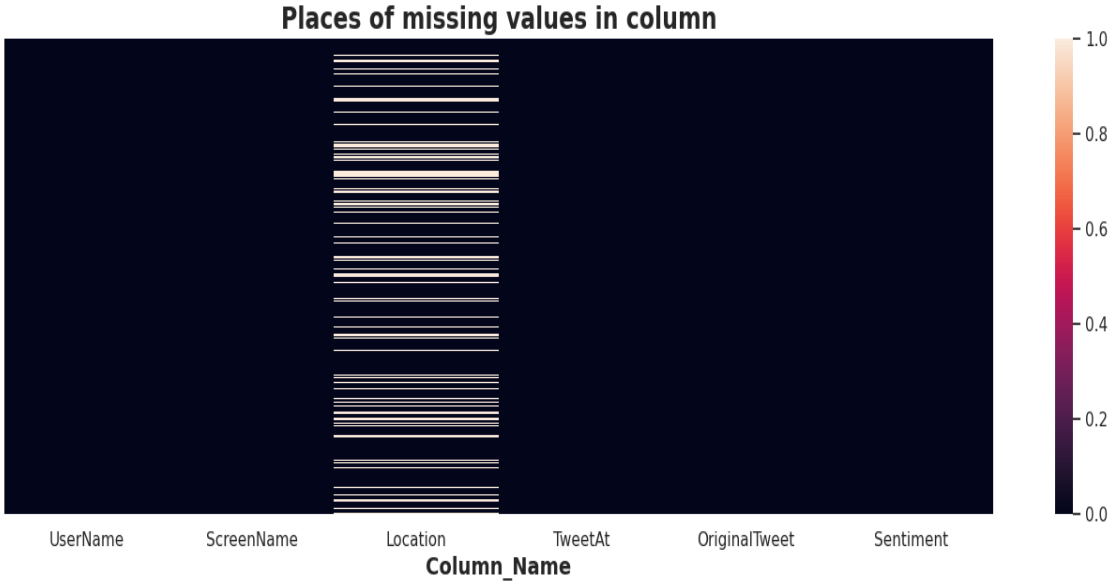
3. LOCATION

```
London                          540
United States                   528
London, England                 520
New York, NY                    395
Washington, DC                  373
                               ...
Staffordshire Moorlands           1
Kithchener ON                     1
Tulsa, Ok                         1
Watford, South Oxhey, Bushey      1
i love you so much || he/him      1
Name: Location, Length: 12220, dtype: int64
```

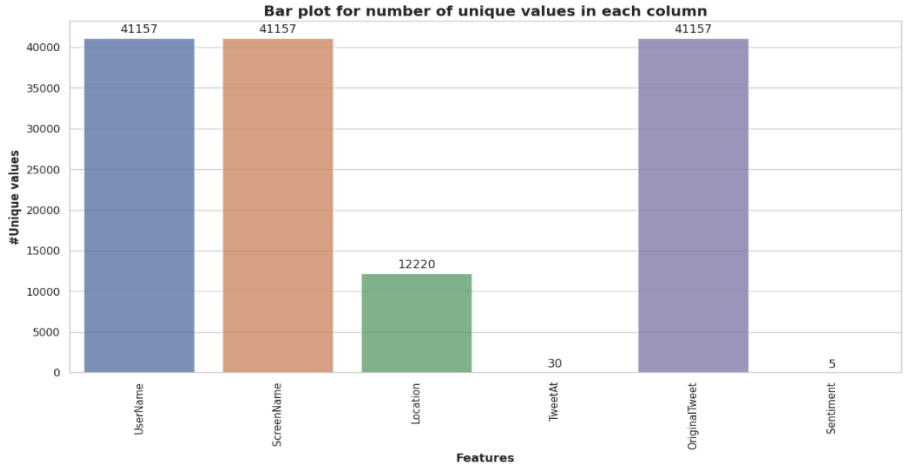
4. PERCENTAGE OF MISSING VALUES IN COLUMN



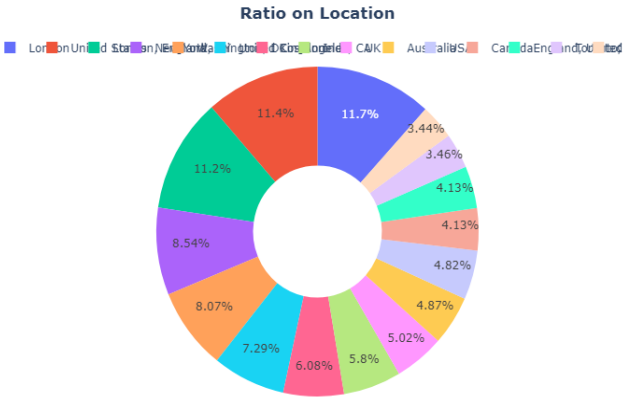
5. PLACES OF MISSING VALUES IN COLUMN



6. BAR PLOT FOR NUMBER OF UNIQUE VALUES IN EACH COLUMN

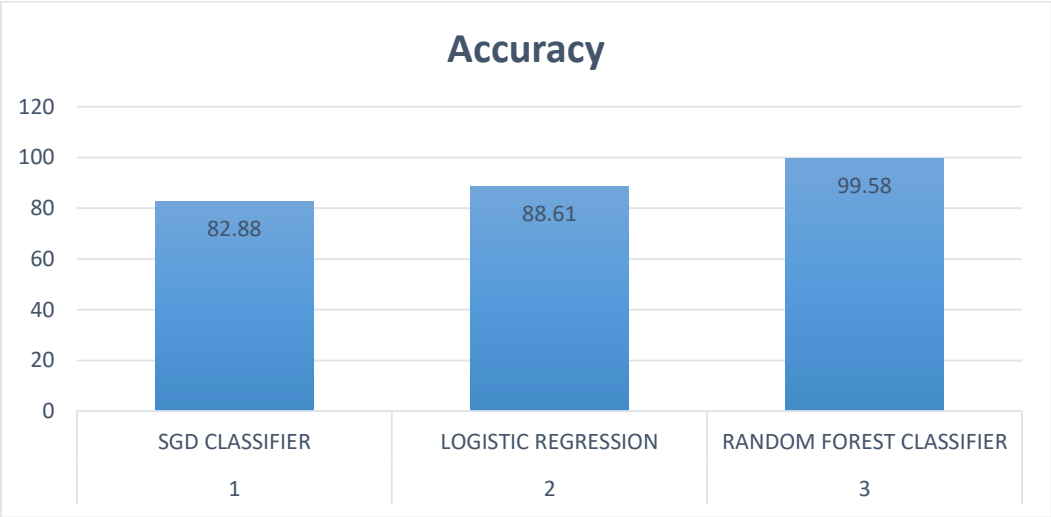


7. RATIO ON LOCATION



COMPARISON BETWEEN SGD CLASSIFIER, LOGISTIC REGRESSION AND RANDOM FOREST CLASSIFIER

S.N.	Method	Accuracy
1	SGD CLASSIFIER	82.88
2	LOGISTIC REGRESSION	88.61
3	RANDOM FOREST CLASSIFIER	99.58



The table displays the results of three distinct classification methods, each accompanied by its corresponding accuracy score, which serves as a measure of their performance in a classification task.

The SGD Classifier (Stochastic Gradient Descent), achieved an accuracy score of 82.88%. This method is known for its efficiency in training and is often used in linear classifiers. While it offers a reasonable accuracy score, it may not be the ideal choice for more intricate or non-linear classification tasks.

Logistic Regression, outperformed the SGD Classifier with an accuracy score of 88.61%. Logistic Regression is a widely-used linear model in binary classification tasks, providing a balance between simplicity and performance, making it a popular choice for various classification problems. The third and most remarkable method in this context is the Random Forest Classifier, boasting an outstanding accuracy score of 99.58%. The Random Forest Classifier utilizes an ensemble of decision trees, making it robust and capable of handling complex, high-dimensional data. Its exceptional accuracy score suggests it is a strong contender for classification tasks, especially those involving intricate patterns or diverse data characteristics.

CONCLUSION

Twitter stream analysis has emerged as a valuable tool for real-time sentiment monitoring of COVID-19 discourse, providing insights into public perceptions, emotions, and concerns during the pandemic. Through the application of natural language processing techniques and machine learning algorithms, researchers have made significant strides in extracting meaningful information from the vast volume of COVID-19-related tweets. The review highlights the importance of data preprocessing, feature extraction, sentiment classification, and evaluation metrics in ensuring the accuracy and reliability of sentiment analysis results. Despite challenges such as noisy data and linguistic nuances, advancements in methodology have enabled researchers to effectively capture and analyze sentiment trends in real-time. The implications of sentiment analysis findings are substantial, offering valuable insights for public health authorities, policymakers, and the general public. By understanding public sentiment, stakeholders can tailor communication strategies, identify misinformation, and respond promptly to emerging concerns. for further enhancement in the accuracy, scalability, and interpretability of sentiment analysis systems. Future research directions may involve integrating multimodal data sources, leveraging deep learning architectures, and exploring real-time visualization techniques to provide actionable insights for mitigating the impact of the pandemic.

REFERENCES

1. Madanian, S., Airehrour, D., Samsuri, N. A., & Cherrington, M. (2021, October). Twitter sentiment analysis in covid-19 pandemic. In 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 0399-0405). IEEE.
2. Chaudhary, M., Kosyluk, K., Thomas, S., & Neal, T. (2023). On the use of aspect-based sentiment analysis of Twitter data to explore the experiences of African Americans during COVID-19. Scientific Reports, 13(1), 10694.
3. Abiola, O., Abayomi-Alli, A., Tale, O. A., Misra, S., & Abayomi-Alli, O. (2023). Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser. Journal of Electrical Systems and Information Technology, 10(1), 1-20.
4. Shofiya, C., & Abidi, S. (2021). Sentiment analysis on COVID-19-related social distancing in Canada using Twitter data. International Journal of Environmental Research and Public Health, 18(11), 5993.
5. Thakur, N. (2023). Sentiment Analysis and Text Analysis of the Public Discourse on Twitter about COVID-19 and MPox. Big Data and Cognitive Computing, 7(2), 116.
6. Malhan, Y., Saxena, S., Mala, S., & Shankar, A. (2021). Geospatial modelling and trend analysis of coronavirus outbreaks using sentiment analysis and intelligent algorithms. In Artificial Intelligence in Healthcare (pp. 1-19). Singapore: Springer Singapore.
7. Chakraborty, A. K., Das, S., & Kolya, A. K. (2021). Sentiment analysis of covid-19 tweets using evolutionary classification-based LSTM model. In Proceedings of Research and Applications in Artificial Intelligence: RAAI 2020 (pp. 75-86). Singapore: Springer Singapore.
8. Raza, G. M., Butt, Z. S., Latif, S., & Wahid, A. (2021, May). Sentiment analysis on COVID tweets: an experimental analysis on the impact of count vectorizer and TF-IDF on sentiment predictions using deep learning models. In 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2) (pp. 1-6). IEEE.
9. Kanakaraddi, S. G., Chikaraddi, A. K., Aivalli, N., Maniyar, J., & Singh, N. (2022, March). Sentiment Analysis of Covid-19 Tweets Using Machine Learning and Natural Language Processing. In Proceedings of Third International Conference on Intelligent Computing, Information and Control Systems: ICICCS 2021 (pp. 367-379). Singapore: Springer Nature Singapore.
10. Kandasamy, V., Trojovský, P., Machot, F. A., Kyamakya, K., Bacanin, N., Askar, S., & Abouhawwash, M. (2021). Sentimental analysis of COVID-19 related messages in social networks by involving an N-gram stacked autoencoder integrated in an ensemble learning scheme. Sensors, 21(22), 7582.
11. Nguyen, T. T., Criss, S., Dwivedi, P., Huang, D., Keralis, J., Hsu, E., ... & Nguyen, Q. C. (2020). Exploring US shifts in anti-Asian sentiment with the emergence of COVID-19. International journal of environmental research and public health, 17(19), 7032.
12. Soomro, Z. T., Ilyas, S. H. W., & Yaqub, U. (2020, November). Sentiment, count and cases: analysis of twitter discussions during covid-19 pandemic. In 2020 7th International conference on behavioural and social computing (BESC) (pp. 1-4). IEEE.

13. Ansari, M. T. J., & Khan, N. A. (2021). Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content. Electronic Journal of General Medicine, 18(6).
14. Irmayani, D., Edi, F., Harahap, J. M., Rangkuti, R. K., Ulya, B., & Watrianthos, R. (2021, June). Naives Bayes algorithm for twitter sentiment analysis. In Journal of Physics: Conference Series (Vol. 1933, No. 1, p. 012019). IOP Publishing.
15. Saini, K., Vishwakarma, D. K., & Dhiman, C. (2021, May). Sentiment analysis of twitter corpus related to covid-19 induced lockdown. In *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)* (pp. 465-470). IEEE.
16. Rahmanti, A. R., Ningrum, D. N. A., Lazuardi, L., Yang, H. C., & Li, Y. C. J. (2021). Social media data analytics for outbreak risk communication: public attention on the "New Normal" during the COVID-19 pandemic in Indonesia. *Computer Methods and Programs in Biomedicine*, *205*, 106083.
17. Saleh, S. N., Lehmann, C. U., McDonald, S. A., Basit, M. A., & Medford, R. J. (2021). Understanding public perception of coronavirus disease 2019 (COVID-19) social distancing on Twitter. Infection Control & Hospital Epidemiology, 42(2), 131-138.