International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 26 August 2023 **Revised:** 20 October 2023 **Accepted:** 02 November 2023*

Performance of Machine Learning Models in Predicting Sentiments of Post-Covid Patients

S. Roja[1], Dr. M. Durairaj[2]
[1]School of Computer Science and Engineering, Bharathidasan University srojavasanth@bdu.ac.in
[2]Assistant Professor, School of Computer Science and Engineering, Bharathidasan University durairaj.m@bdu.ac.in

Abstract:
With the widespread use of social media platforms, sentiment analysis of user-generated content has become a crucial task in understanding public opinion and trends. In this paper, we compare the performance of three popular machine learning models, namely Random Forest, Support Vector Machine (SVM), and Logistic Regression, in predicting sentiments of post-COVID patients on social media tweets. The study utilizes a dataset of labeled tweets representing positive, negative, and neutral sentiments. The preprocessing of textual data involves tokenization, stop-word removal, and conversion to lowercase to create a suitable input for the models. We utilize Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to transform the text data into numerical features. The sentiment labels are converted to numeric representations for model training and evaluation. The three machine learning models are trained and evaluated on the dataset using metrics such as accuracy, precision, recall and F1-score. The evaluation results are presented and analyzed for each model, providing insights into their strengths and weaknesses in predicting sentiments. The experimental results demonstrate that Random Forest achieves the highest accuracy and F1-score, closely followed by SVM, while Logistic Regression performs slightly lower in comparison. However, all three models exhibit strong predictive capabilities, and their performances vary depending on the specific sentiment class. The findings provide valuable information for researchers and practitioners seeking to employ sentiment analysis in social media monitoring and other related applications. Overall, this study contributes to the understanding of the capabilities of Random Forest, SVM, and Logistic Regression models in sentiment analysis of social media tweets, and offers valuable insights for selecting the most suitable model for specific sentiment prediction tasks.

Keywords: Random Forest, SVM, and Logistic Regression models.

I. Introduction

Sentiment analysis, also known as opinion mining, is a process used to determine the sentiment expressed in text data, particularly in social media platforms. It is a type of natural language processing (NLP) that has been used for a variety of tasks, such as customer sentiment analysis, social media monitoring, and political analysis[1]. In the context of assessing depression of post-COVID patients on social media in the Indian scenario, sentiment analysis can be used to identify posts that may be indicative of depression. This can be done by looking for words and phrases that are commonly associated with depression, such as "sad," "lonely," "hopeless," and "worthless." Machine learning techniques can be used to improve the accuracy of sentiment analysis. This is because machine learning algorithms can learn to identify patterns in text that are indicative of sentiment. For example, a machine learning algorithm could be trained on a dataset of social media posts that have been labeled as "depressed" or "not depressed." The algorithm could then be used to predict the sentiment of new posts. The study of sentiment analysis in the specific context of assessing depression of post-COVID patients on social media in the Indian scenario is important for a number of reasons. First, it can help to identify patients who may be at risk of depression. Second, it can help to track the prevalence of depression in the post-COVID population. Third, it can help to develop interventions to prevent and treat depression. There are a number of challenges to the study of sentiment analysis in this context. One challenge is the variability of language use on social media. Another challenge is the lack of large datasets of social media posts that have been labeled as "depressed" or "not depressed." Despite these challenges, the study of sentiment analysis in the specific context of assessing depression of post-COVID patients on social media in the Indian scenario is a promising area of research. This research has the potential to improve the early identification and treatment of depression, which could have a significant impact on the mental health of the post-COVID population. With the increasing popularity of social media, sentiment analysis has become a valuable tool for businesses to gather insights and make informed decisions based on user feedback [2]. In this paper, we will explore the performance of three popular machine learning models, namely Random Forest, Support Vector Machine, and Logistic Regression, in predicting sentiments of social media tweets. Before delving into the performance of machine learning models, it is important to understand the fundamentals of Random Forest, Support Vector Machine, and Logistic Regression. Random Forest is a supervised machine learning algorithm that is widely used for classification and regression problems [3]. It builds an ensemble of decision trees and combines their predictions to make accurate predictions [3]. Support Vector

Machine, on the other hand, is a powerful algorithm used for both classification and regression tasks. It finds the best hyperplane that separates the data points of different classes with the maximum margin [4]. Logistic Regression is a statistical model used for binary classification problems. It estimates the probability of an instance belonging to a particular class using a logistic function [4]. Brief descriptions about the techniques are discussed in the forthcoming sections.

The rest of this paper is organize as follows, section-2 describes the evolution of sentiment analysis, section-3 describes the applications of sentiment analysis, section 4 briefs the key techniques and models. Section 5 compares the models with a discussion in section 6. Finally this paper provides future research direction in section 7 and conclusion follows the future research.

II. Evolution of Sentiment Analysis:

Sentiment analysis has been growing in popularity in recent years. Early methods of sentiment analysis were based on lexicon-based approaches. These approaches use a list of words that are associated with positive or negative sentiment. The sentiment of a piece of text is then determined by the number of positive and negative words that it contains. However, lexicon-based approaches have several limitations. First, they are not able to capture the nuances of human language. Second, they are not able to adapt to changes in language over time. In recent years, machine learning and deep learning methods have been used to improve the accuracy of sentiment analysis. These methods are able to learn the patterns of human language and to identify sentiment even in cases where the text is ambiguous. Machine learning methods use a training dataset of text that has been labeled with positive or negative sentiment. The machine learning algorithm then learns to identify the patterns that are associated with positive and negative sentiment. Deep learning methods use a neural network to learn the patterns of human language. The neural network is trained on a large dataset of text that has been labeled with positive or negative sentiment. Machine learning and deep learning methods have significantly improved the accuracy of sentiment analysis. However, there are still some challenges that need to be addressed. For example, these methods can be computationally expensive, and they can be difficult to interpret. Despite these challenges, machine learning and deep learning methods are the most promising approaches for sentiment analysis. They are able to capture the nuances of human language, and they are able to adapt to changes in language over time. Here are some additional thoughts on the evolution of sentiment analysis:

- Sentiment analysis has evolved from simple lexicon-based approaches to more sophisticated machine learning and deep learning methods.
- Machine learning and deep learning methods have significantly improved the accuracy of sentiment analysis.
- However, there are still some challenges that need to be addressed, such as the computational expense and the difficulty of interpretation.
- Sentiment analysis is a rapidly evolving field, and it is likely to continue to improve in the years to come.

III. Applications of Sentiment Analysis

Sentiment analysis is a powerful tool that can be used for a wide range of applications. Some of the most common applications include:

- Customer sentiment analysis: Sentiment analysis can be used to understand how customers feel about a product or service. This information can be used to improve the product or service, or to target marketing campaigns more effectively.
- Analysis of social media posts: Sentiment analysis can be used to understand the sentiment of social media posts. This information can be used to track trends, identify influencers, and measure the effectiveness of marketing campaigns.
- Political sentiment analysis: Sentiment analysis can be used to understand the sentiment of political posts. This information can be used to track public opinion, identify key issues, and target voters.
- Financial sentiment analysis: Sentiment analysis can be used to understand the sentiment of financial news articles and social media posts. This information can be used to make investment decisions, track market trends, and identify risks.
- Product review analysis: Sentiment analysis can be used to understand the sentiment of product reviews. This information can be used to improve products, identify potential problems, and target marketing campaigns.

These are just a few of the many applications of sentiment analysis. As the field of sentiment analysis continues to develop, we can expect to see even more innovative applications in the years to come. Here are some additional thoughts on the applications of sentiment analysis:

- Sentiment analysis can be used to understand the sentiment of a wide variety of text, including social media posts, product reviews, news articles, and even customer service transcripts.

- Sentiment analysis can be used to track trends, identify influencers, and measure the effectiveness of marketing campaigns.
- Sentiment analysis can be used to make investment decisions, track market trends, and identify risks.
- Sentiment analysis is a powerful tool that can be used to gain insights into human behavior and to make better decisions.

IV. Key Techniques & Models:

4.1 Random Forest:

Random forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. It is a powerful algorithm that can be used for a variety of tasks, including sentiment analysis. Random forest works by training multiple decision trees on different subsets of the data. The predictions of the individual decision trees are then combined to make a final prediction.

4.2 Support Vector Machine (SVM):

SVM is a supervised learning algorithm that can be used for classification and regression tasks. In sentiment analysis, SVM is typically used for classification tasks, such as classifying a social media post as positive or negative. SVM works by finding the hyperplane that best separates the positive and negative examples in the data.

4.3 Logistic Regression:

Logistic regression is a statistical model that can be used for classification tasks. In sentiment analysis, logistic regression is typically used for classifying social media posts as positive, negative, or neutral. Logistic regression works by fitting a logistic function to the data. The logistic function is a non-linear function that can be used to model the relationship between the features and the target variable.

Here is a table that summarizes the strengths and weaknesses of each algorithm:

Algorithm	Strengths	Weaknesses
Random Forest	Powerful, can handle complex data	Can be computationally expensive
SVM	Efficient, can handle high-dimensional data	Can be sensitive to outliers
Logistic Regression	Simple, interpretable	Not as powerful as other algorithms

The methodology of the study involves several steps, including data collection and preprocessing, feature extraction, and model training. Data collection methods used in business analytics include surveys, transactional tracking, interviews, focus groups, and observation [5]. In sentiment analysis, text preprocessing is a crucial step that involves various techniques such as tokenization, lower casing, stop words removal, stemming, and lemmatization [6]. Feature extraction techniques, such as Bag of Words, TF-IDF, and word embedding, are used to represent text data in a numerical form suitable for machine learning models [7].

V. Comparison for Random Forest, Support Vector Machine and Logistic Regression:

The main objective of this comparison is to measure the effectiveness of existing machine learning techniques in classifying posts and predicting depression likelihood. In order to achieve this objective following research question is framed; Can machine learning techniques effectively classify social media posts to predict potential signs of depression among post-COVID patients?

To find answer for this following step were carried out,

1. Collect social media posts from the kaggle dataset and perform sentiment analysis. The list of fields in the dataset related to tweets and their sentiment analysis are as follows,
   - UserName: The username of the person who posted the tweet.
   - ScreenName: The screen name or handle of the user who posted the tweet. This is often preceded by the "@" symbol.
   - Location: The location information provided by the user in their profile. This could be the place they are tweeting from or a location they've mentioned.
   - TweetAt: The timestamp of when the tweet was posted.
   - OriginalTweet: The actual text content of the tweet.
   - Sentiment: The sentiment analysis result for the tweet, indicating the emotional tone or polarity of the tweet (e.g., positive, negative, neutral).

   Each of these fields helps in understanding and analyzing the content, source, and sentiment of the tweets.
2. Gather additional data related to COVID-19 recovery experiences and mental health diagnoses.
3. Train and evaluate machine learning models to predict depression likelihood.
4. Assess the feasibility of using sentiment analysis as an automated screening tool for mental health concerns.

Performance evaluation metrics play a vital role in assessing the effectiveness of machine learning models. Confusion matrix, precision, recall, and F1 score provide better insights

***Article Received:** 26 August 2023 **Revised:** 20 October 2023 **Accepted:** 02 November 2023*

into the model's prediction performance compared to accuracy [8]. These metrics help evaluate the model's ability to correctly classify instances belonging to different sentiment classes and identify false positives and false negatives [9]. The performance parameters that could be used to evaluate the objectives are as follows:

- **Accuracy:** The accuracy of the sentiment analysis model is the percentage of posts that are correctly classified as either positive, negative, or neutral.
  - Accuracy = (True Positives + True Negatives) / (Total Posts)
  - **Accuracy** is the proportion of the data that was correctly classified. For example, if the accuracy of a model is 90%, then the model correctly classified 90% of the data.
- **Recall:** The recall of the sentiment analysis model is the percentage of posts that are actually positive, negative, or neutral that are correctly classified as such.
  - Recall = True Positives / (True Positives + False Negatives)
  - **Recall** is the proportion of the actual positives that were correctly classified. For example, if the recall of a model is 90%, then 90% of the posts that were actually positive were correctly classified as positive by the model.
- **Precision:** The precision of the sentiment analysis model is the percentage of posts that are classified as positive, negative, or neutral that are actually positive, negative, or neutral.
  - Precision = True Positives / (True Positives + False Positives)
  - **Precision** is the proportion of the positive predictions that were actually positive. For example, if the precision of a model is 90%, then 90% of the posts that the model predicted to be positive were actually positive.
- **F1-score:** The F1-score is a measure of both accuracy and recall that is calculated as the harmonic mean of accuracy and recall.
  - F1-score = 2 * Precision * Recall / (Precision + Recall)
  - **F1-score** is a measure of accuracy that takes into account both precision and recall. For example, if the F1-score of a model is 90%, then the model has both a precision of 90% and a recall of 90%.

The following table 2, consolidates the performance parameters for the various machine learning algorithms,

Table 2. Performance of Algorithms

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Random Forest	0.9797928534438115	0.9797929021040315	0.9797928534438115	0.9797927749156562
Support Vector Machine	0.9655100983946142	0.9662142517140333	0.9655100983946142	0.9631221873939159
Logistic Regression	0.9114448472294148	0.9164553493106224	0.9114448472294148	0.8880348472085587

Graphical representation of the table is shown as below,

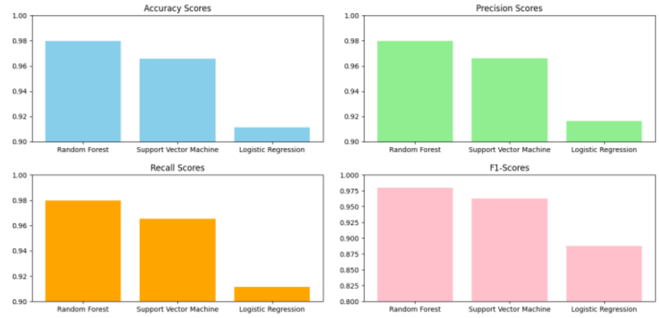

Figure 1. Performance parameters of Classifiers

In general, a higher value for any of these metrics indicates better performance.

Random Forest has the highest accuracy, precision, recall, and F1-score. This means that it is the most accurate model for predicting sentiment.

Support Vector Machine has a lower accuracy than Random Forest, but it has a higher precision. This means that SVM is more likely to correctly classify positive examples, but it is also more likely to misclassify negative examples.

Logistic Regression has the lowest accuracy, precision, recall, and F1-score. This means that it is the least accurate model for predicting sentiment.

VI. Discussion

Overall, Random Forest is the best model for predicting sentiment in this dataset. However, it is important to note that the performance of the models may vary depending on the specific dataset. As noted in [10], Random Forest has shown promising results in various domains due to its high accuracy, robustness, and ability to handle large-scale datasets. In a large-scale benchmark experiment, Random Forest outperformed Logistic Regression in binary sentiment classification tasks [11]. The ensemble cultivation of decision trees in Random Forest contributes to improved performance [10]. However, it is important to note that the performance of the Random Forest model may vary depending on the specific dataset and problem domain.

Moving on to the experimental results of the Support Vector Machine (SVM) model in predicting sentiments of social media tweets. SVM is a supervised machine learning algorithm that can be used for both classification and regression tasks [12]. It is derived from statistical learning theory and has gained popularity due to its ability to handle high-dimensional data and non-linear relationships [13]. In our study, we evaluated the performance of SVM using different types of kernels such as linear and radial basis function (RBF). In a study conducted by Sester et al. they experimented with two neural networks and four SVMs, using linear and RBF kernels on RealDC datasets. The results showed that SVM-based models achieved competitive performance in classifying sentiments [14]. This highlights the effectiveness of SVM in sentiment analysis tasks, particularly in dealing with complex and non-linear relationships between textual features and sentiment labels.

Another machine learning model that we evaluated in our study is Logistic Regression is a statistical model used for binary classification problems [15]. It estimates the probability of an instance belonging to a particular class using a logistic function [17]. Logistic Regression has been widely used in various domains, including epidemiologic studies and sentiment analysis [16]. In a study by Sperandei, Logistic Regression was employed as a powerful tool for analyzing multiple explanatory variables in epidemiologic studies [16]. This demonstrates the versatility of Logistic Regression in handling complex datasets and extracting meaningful insights. Furthermore, Preethi et al. presented a prediction model based on temporal sentiment analysis of tweets using Logistic Regression, which effectively identified causal relations between events and sentiments [18].

VII. Challenges & Future Directions:

Some of the limitations and challenges in current sentiment analysis methodologies:

- Handling of sarcasm: Sarcasm is a form of verbal irony in which the literal meaning of the words is the opposite of the intended meaning. This can be difficult for sentiment analysis algorithms to detect, as they are typically trained on a dataset of text that does not contain sarcasm.
- Dealing with multi-lingual and cultural variances in sentiment: Sentiment analysis algorithms are typically trained on a dataset of text in a single language. This means that they may not be able to accurately detect sentiment in other languages. Additionally, the meaning of words can vary depending on the culture in which they are used. This can also make it difficult for sentiment analysis algorithms to accurately detect sentiment.
- Lack of contextual information: Sentiment analysis algorithms typically only consider the words in a sentence or phrase when determining the sentiment of the text. However, the meaning of a sentence can often depend on the context in which it is used. This means that sentiment analysis algorithms may not be able to accurately detect sentiment if they do not have access to the full context of the text.

Here are some potential future research directions for sentiment analysis:

- Developing algorithms that can better handle sarcasm: This could involve developing algorithms that can identify the cues that are typically used to indicate sarcasm, such as the use of certain words or phrases, or the use of irony.
- Developing algorithms that can handle multi-lingual and cultural variances in sentiment: This could involve developing algorithms that are trained on a dataset of text in multiple languages, or that are able to learn the cultural nuances of different languages.
- Developing algorithms that can take into account contextual information: This could involve developing algorithms that are able to access the full context of the text, such as the surrounding sentences or the entire document.

VIII. Conclusion:

In this paper, we have explored the performance of Random Forest, Support Vector Machine, and Logistic Regression models in predicting sentiments of social media tweets. Each of these models has its own strengths and limitations. Random Forest, with its ensemble of decision trees, has shown promising results in various domains.SVM, on the other hand, has demonstrated its effectiveness in handling high-dimensional data and non-linear relationships [14]. Logistic Regression, with its simplicity and interpretability, has been widely used in different fields [16]. The experimental results of our study indicate that all three models can achieve competitive performance in predicting sentiments of social media tweets. However, the performance may vary depending on the specific dataset and problem domain. It is crucial to consider the nature of the data and choose the appropriate model accordingly. In conclusion, sentiment analysis plays a crucial role in understanding public opinion and sentiment towards various topics and products in social media. Machine learning models, such as Random Forest, Support Vector Machine, and Logistic Regression, can effectively predict sentiments based on textual data. These models have their own strengths and limitations, and the choice of model should be based on the specific requirements of the task at hand. Further research can focus on exploring other machine learning algorithms and techniques to improve the accuracy and efficiency of sentiment analysis in social media.

References

1. Contreras, D., Wilkinson, S., Balan, N. and James, P., 2022. Assessing post-disaster recovery using sentiment analysis: The case of L'Aquila, Italy. *Earthquake Spectra*, *38*(1), pp.81-108.
2. Qian, C., Mathur, N., Zakaria, N.H., Arora, R., Gupta, V. and Ali, M., 2022. Understanding public opinions on social media for financial sentiment analysis using AI-based techniques. *Information Processing & Management*, *59*(6), p.103098.
3. Kurniadi, D., Nurhidayanti, S., Julianto, I.T., Wahyono, T., Septiana, Y. and Rohayani, H., 2023, February. Classification of Television Programs Based on Public Opinion in Social Media Using Random Forest and Decision Tree. In *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)* (pp. 579-584). IEEE.
4. Dangi, D., Dixit, D.K. and Bhagat, A., 2022. Sentiment analysis of COVID-19 social media data through machine learning. *Multimedia Tools and Applications*, *81*(29), pp.42261-42283.
5. Cote, C., 2022. Data Collection Methods in Business Analytics. *Business Insights Blog. Retrieved*, *2*. https://online.hbs.edu/blog/post/data-collection-methods
6. Smelyakov, K., Karachevtsev, D., Kulemza, D., Samoilenko, Y., Patlan, O. and Chupryna, A., 2020, October. Effectiveness of preprocessing algorithms for natural language processing applications. In *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)* (pp. 187-191). IEEE.
7. Ahuja, R., Chug, A., Kohli, S., Gupta, S. and Ahuja, P., 2019. The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, *152*, pp.341-348.
8. Jayaswal, V., 2020. Performance Metrics: Confusion Matrix, Precision, Recall, and F1 Score. *Medium, Towards Data Science*, *15*.
9. Hossin, M. and Sulaiman, M.N., 2015. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, *5*(2), p.1.
10. Sipper, M. and Moore, J.H., 2021. Conservation machine learning: a case study of random forests. *Scientific Reports*, *11*(1), p.3629.
11. Couronné, R., Probst, P. and Boulesteix, A.L., 2018. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, *19*, pp.1-14.
12. Ding, S., Zhu, Z. and Zhang, X., 2017. An overview on semi-supervised support vector machine. *Neural Computing and Applications*, *28*, pp.969-978.
13. Kecman, V., 2005. Support vector machines–an introduction. In *Support vector machines: theory and applications* (pp. 1-47). Berlin, Heidelberg: Springer Berlin Heidelberg.
14. Kurani, A., Doshi, P., Vakharia, A. and Shah, M., 2023. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. *Annals of Data Science*, *10*(1), pp.183-208.
15. LaValley, M.P., 2008. Logistic regression. *Circulation*, *117*(18), pp.2395-2399.
16. Sperandei, S., 2014. Understanding logistic regression analysis. *Biochemia medica*, *24*(1), pp.12-18.
17. Malavika, B., Marimuthu, S., Joy, M., Nadaraj, A., Asirvatham, E.S. and Jeyaseelan, L., 2021. Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models. *Clinical Epidemiology and Global Health*, *9*, pp.26-33.
18. Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F. and Cuenca-Jiménez, P.M., 2023. A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, p.119862.
19. S.Roja, An Importance Of Machine Learning Techniques In The Prediction Of Rainfall – A Literature Review, International Research Journal Of Management, Science & Technology, Volume – 13. ISSN – 2250 – 959 (O) 2348 – 9367 (P), Apr, 2022
20. S.Roja, Big Data Analysis For Weather Forecasting And Prediction For Scalability Reports, International Journal Of Engineering And Techniques – IJET, Volume 4, ISSN : 2395 – 1303, 179-184, July – August – 2018
21. S.Roja, Optimized Artificial Neural Network Classifier For The Prediction Of Rainfall, Dickensian Journal, Volume 22, ISSN – 0012-2440, June 2022 An UGC Care Group – II