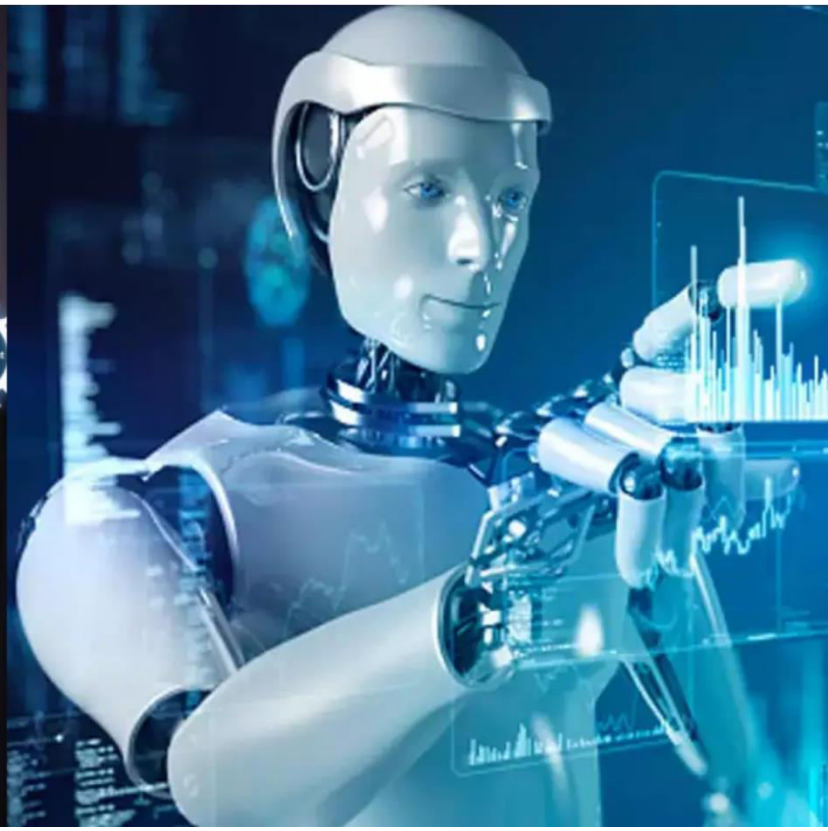
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 5, May 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505177|

Social Media Sentiment Analyzer using NLP

Sachi Bagade, Bharti Zankar, Rushikesh Shirsath, B.N.Babar, Bhagyashree Patil

U.G. Student, Department of AI&DS, SRTTC Engineering College, Kamshet, Pune, Maharashtra, India

U.G. Student, Department of AI&DS, SRTTC Engineering College, Kamshet, Pune, Maharashtra, India

U.G. Student, Department of AI&DS, SRTTC Engineering College, Kamshet, Pune, Maharashtra, India

Associate Professor, Department of AI&DS, SRTTC Engineering College, Kamshet, Pune, Maharashtra India

Associate Professor, Department of AI&DS, SRTTC Engineering College, Kamshet, Pune, Maharashtra India

ABSTRACT: People share a lot of opinions on social media through text. It takes a lot of time and efforts this paper describes a system that uses natural language processing (NLP) to understand and classify social media text. The system uses kaggle dataset and does text preprocessing like cleaning, tokenization, and removing stop words. TF-IDF is used to create features from the text. Different models are tested, and an Random Forest model is used for the final prediction. The model shows about 91% accuracy when tested on the data. The system is useful for understanding public opinion from social media posts comments and real-time user input.

KEYWORDS: Sentiment Analysis, NLP, LSTM, Social media, Text classification.

I. INTRODUCTION

People use social media everyday to share their thoughts feelings and experiences about services, products and events these short posts and comments create a large amount of text data that helps in understanding public opinion, but analyzing this data manually is slow and very difficult on a large datasets. Sentiment analysis is used to identify the emotional tone of a text as positive, Negative, Neutral by using NLP techniques, computers can analyze written text and find meaningful information from it. In this research, we develop a sentiment analysis model that is trained using example and identifies sentiment in social media posts. The system uses a dataset from Kaggle and perform basic text cleaning and feature extraction techniques before training the model. Different models are tested, and an Random Forest. Based model is selected to improve the handling of noisy and informed social media texts. This method allows the system to understand online user feedback and understand important information from it.

Furthermore, the proposed approach focuses on improving the overall accuracy and reliability of sentiment prediction by carefully preprocessing the textual form of data, including removing stop words, punctuations, and irrelevant symbols. Feature extraction techniques such as TFIDF helps in converting text information into numerical form, making it easier for machine learning algorithms to process the data effectively. The Random Forest model is particularly chosen because of its robustness and ability to handle complex and unstructured data which is easily found on social media content. Overall this research highlights the importance of sentiment analysis in today's digital world and shows how machine learning techniques cam simplify the process of analyzing large-scale textual data efficiently.

II. LITERATURE SURVEY

With the advancement techniques of machine learning, supervised classification techniques became one of the main approaches used for sentiment analysis. Pang et al. [1] (2002) applied method such as Naïve Bayes, Maximum Entropy, and Support Vector Machines (SVM) to analyse movie reviews. Their study showed that SVM obtained better accuracy as compared to other traditional classifiers. Later studies applied these process to social media text which is normally short, informal, and unstructured. To handle this type of data properly techniques like TF-IDF and n-grams became common methods for changing text into organized features compatible for machine learning models. Researcher also studied integrated methods such as Random Forest and Gradient Boosting to improve prediction accuracy and decrease overfitting. Chen at el. [2] (2017) showed that ensemble models can successfully manage high dimensional feature spaces and give more reliable sentiment predictions across different dataset. As technology developed further deep learning methods started playing important role in sentiment analysis. Recurrent Neural Network (RNNs) especially Long Short Term Memory (LSTM) network became

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505177|

commonly used for analysing text that appear in a sequence. These models can learn the relationship between words in a sentence which helps in better understandings the sentiment expressed in different parts of the text. Tang et al. [4] (2015) showed that LSTM based models perform well for twitter sentiment classification especially when they are used with pre-trained word representation methods such as Word2Vec or GloVe.

Comparison of Methods			
Study	Method used	Results	Remarks
SVM on Tweeter Data	Support Vector Machine	~80% accuracy on Structured data	Struggles with Short/noisy text
LSTM based Sentiment Analysis	Long Short Term Memory (LSTM)	~70% accuracy, in previous studies	Needs large Data and long training time
Comparison of ML Algorithms	Navie Bayes, KNN Random forest	Random Forest performed well across dataset	Less Effective with complex sentence structure
Our Approach	NLP with LSTM ,SVM, Random Forest	Random Forest Achieved 91% accuracy— Highest among all	Comparative evaluation of multiple models on social media text
BERT-Based Sentiment Analysis	Bidirectional Encoder Representation from Transformers (BERT)	75% accuracy on large datasets	Powerful for understanding the given data but requires high mathematical resources
Hybrid Deep learning Model	CNN + LSTM	83% accuracy on social media text	Combines Features extraction of CNN with Sequence learning of LSTM

III. METHODOLOGY

a. Data Collection The developed sentiment analysis system follows a organized method to analyse social media text and identify user opinions. The methods includes several steps such as data collection, text preprocessing, feature extraction, model training, and result analysis.

b. Text Preprocessing The dataset used in this research contains social media posts with different sentiment opinions. Due to social media text commonly contains informal language, short forms, emojis and other unwanted elements, text cleaning is needed to make the data uniform. Basic preprocessing steps such as deleting unwanted symbols, changing text to lowercase, and tokenization are used to improve data quality.

c. Feature Extraction After preprocessing the text is converted into numbers so that machine learning models can work with it. In this work, methods such as TF – IDF vectorization and word representation are used to represent the data.

d. Model Training Multiple models such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and XLM are used and compared.

e. Evaluation Metrices The performance of these models is tested using common testing such as accuracy, precision, recall, and F1 score. Confusion metrices are also created to better understand the classification performance for each sentiment category.

IV. EXPERIMENTAL RESULTS

Many machine learning and deep learning models were used to analyze the proposed sentiment analysis system. The efficiency of the model was determined using accuracy, precision, recall, and F1-score. The result display that Random Forest obtained the highest accuracy nearly 91% followed by the XLM model shows accuracy around 56% Logistic Regression and SVM showed average efficiency with accuracy values of about 85% and 73% respectively. The LSTM (Long Short Term Memory) model gives the lowest accuracy of about 26% in this study. These results shows that Random Forest performed best for this dataset, Whereas the other models showed in comparison lower efficiency.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505177|

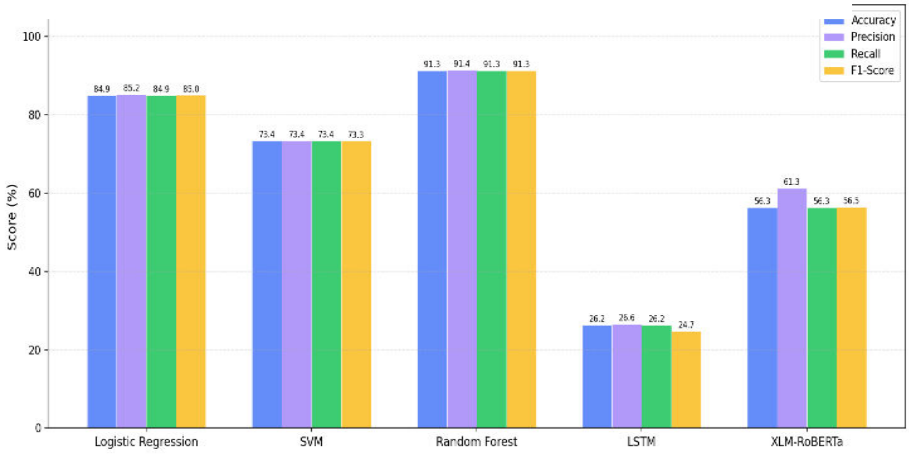


Fig.1. Accuracy comparison of models

Here's a graphical representation of the experimental results. The bar chart compares the accuracy of all models:

1) Logistic Regression: 85%
2) SVM: 73%
3) Random Forest: 91%
4) LSTM: 26%
5) XLM: 56%

This visualization clearly highlights Random Forest as the best-performing model, while the other models show lower performance on the dataset.

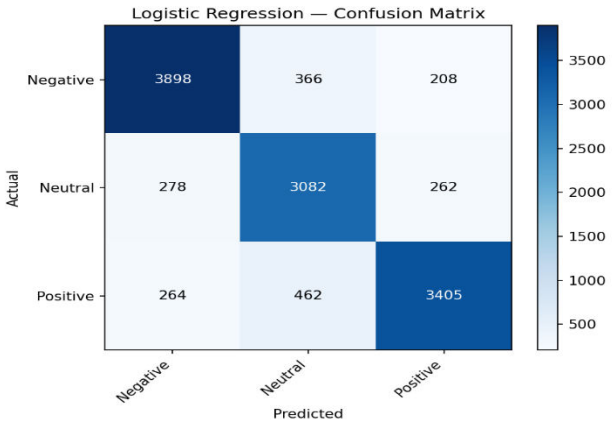


Fig.2. Logistic Regression - Confusion matrix

The confusion matrix for logistic Regression shows how well the model is able to classify sentiments into Negative, Neutral, and Positive categories. Most of the values are concentrated along the diagonal, which means the model is correctly predicting a large number of samples.

Overall, the confusion matrix suggests that logistic regression provides good performance, but there is still room for improvement in reducing wrong sentiment classification between sentiment classes.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505177|

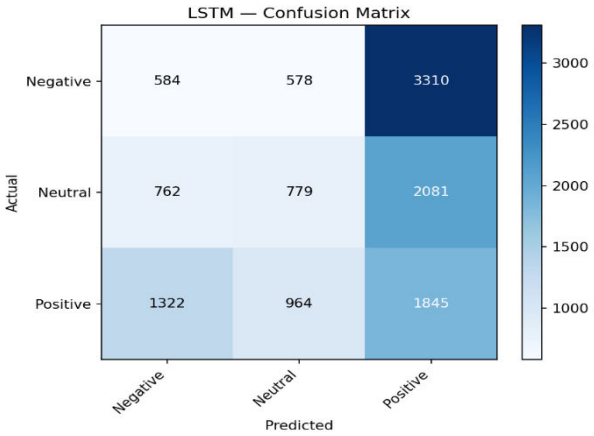


Fig.3. LSTM – Confusion matrix

The confusion matrix for LSTM model shows that its performance is not very strong compared to other models. From the matrix, it is clearly visible that the correct predictions are relatively low as compared to other models confusion matrix. A major issue with this model is that it misclassifies many samples into the positive category. Similarly, positive texts are also confused with negative and neutral classes.

This indicates that the LSTM model is struggling to properly understand and differentiate between the sentiments in the dataset. Overall the confusion matrix shows that it has misclassification and is less reliable in this case.

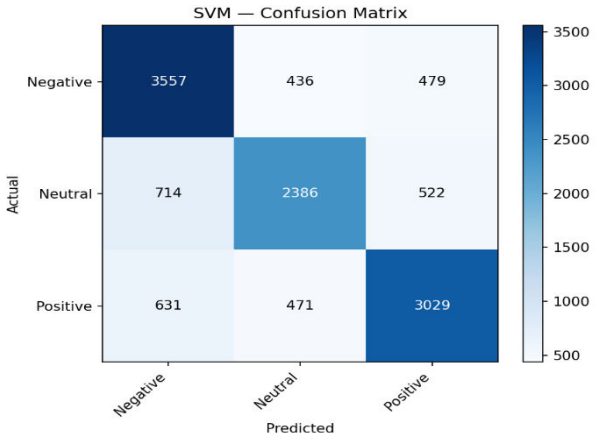


Fig. 4 SVM – Confusion matrix

The confusion matrix for the SVM model shows that it performs reasonably well in classifying sentiments into Negative, Neutral and Positive categories. A large number values are present on it, which means the model is correctly predicting many instances. At the same time, there are some misclassifications. Some negative texts are predicted as neutral and positive. At the same time, neutral and positive texts are sometimes confused with other categories. This shows that model faces some difficulty in clearly separating similar sentiments, especially between neutral and other two classes.

Overall, the SVM model gives decent performance but it is not perfect and still makes noticeable errors in classification.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505177|

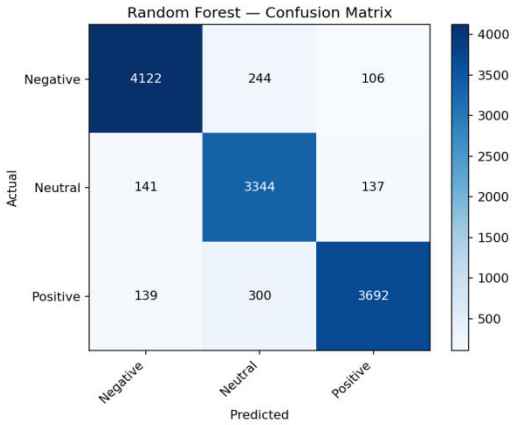


Fig.5. Random Forest – Confusion matrix

The confusion matrix for the random forest model shows very strong performance in classifying sentiments. Most of the values are clearly concentrated along the diagonal which means the model is correctly predicting a large number of samples in each category.

The number of misclassifications is relatively low. Only a small number of negative texts are predicted as neutral or positive. Similarly, neutral and positive classes also have fewer wrong predictions compared to other models. This indicates that the model is able to identify well between different sentiment classes.

Overall, the confusion matrix shows that random forest performs very accurately and consistently, making it the most reliable model for sentiment classification in this case.

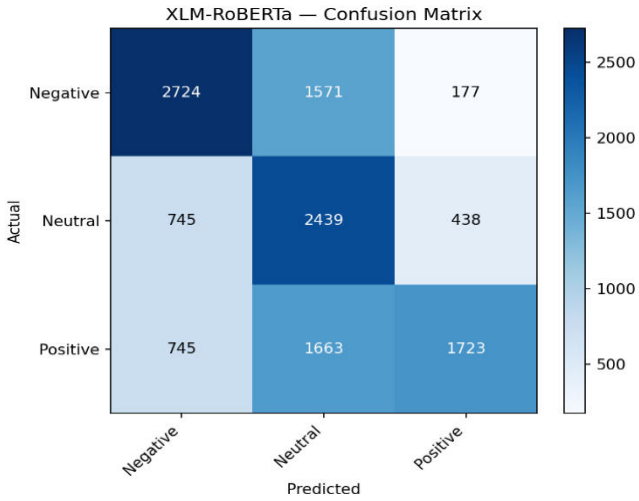


Fig.6. XLM- RoBERTa – Confusion matrix

The confusion matrix for the XLM- RoBERTa shows that the performance of the XLM- RoBERTa model across Negative, Neutral, and Positive sentiment classes. The model correctly classifies a large number of Negative samples, showing strong performance for this class. However, there is significant misclassification between Negative and Neutral categories. The Neutral class also shows confusion with both Negative and Positive sentiments. Similarly, many Positive samples are incorrectly predicted as Neutral. This indicates that the model finds it difficult to distinguish between closely related sentiments. Overall, the model performs well but requires improvement in handling Neutral and Positive classifications.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505177|

V. CONCLUSION

This study presented a comparative analysis of multiple machine learning and deep learning approaches for sentiment analysis of social media text. The project implemented traditional machine learning models, including Logistic Regression, Support Vector Machine (SVM), and Random Forest, along with a deep learning model based on Long Short-Term Memory (LSTM). In addition, a transformerbased pre-trained model, XLM- RoBERTa, was evaluated as a benchmark to compare classical approaches with modern natural language processing architectures.

The experimental results demonstrate that ensemble and transformer-based models generally provide better performance in sentiment classification tasks. Among the traditional machine learning models, Random Forest achieved the most balanced performance due to its ensemble learning capability and robustness against overfitting. However, the pre-trained transformer model, XLM RoBERTa, achieved the highest overall accuracy and F1score because of its contextual language understanding and knowledge learned from large-scale multilingual datasets. The project also integrated the trained models into a Flaskbased web application that allows users to input text and receive real-time sentiment predictions along with confidence scores and performance visualizations. This demonstrates how machine learning models can be effectively deployed in practical applications for opinion mining and social media analysis. Overall, the results highlight the importance of contextual language modeling in sentiment analysis and show that transformer-based models significantly improve prediction accuracy compared to traditional feature-based approaches. The system provides a useful framework for analyzing user opinions and can be extended further with larger datasets, improved model tuning, and scalable deployment.

REFERENCES

1. B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86, 2002.
2. T. Joachims, “Text categorization with Support Vector Machines: Learning with many relevant features,” Proc. 10th Eur. Conf. Mach. Learn. (ECML), pp. 137–142, 1998.
3. L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
4. Yassin M. Y. Hasan and Lina J. Karam, “Morphological Text Extraction from Images”, IEEE Transactions On Image Processing, vol. 9, No. 11, 2000
5. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997. [5] D. Tang, B. Qin, and T. Liu, “Document modeling with gated recurrent neural networks for sentiment classification,” Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 1422–1432, 2015.
6. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.
7. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pretraining of deep bidirectional transformers for language understanding,” Proc. Conf. North American Chapter of the Association for Computational Linguistics (NAACL), pp. 4171– 4186, 2019.
8. B. Liu, Sentiment Analysis and Opinion Mining. San Rafael, CA: Morgan & Claypool, 2012.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details