

Hybrid TF-IDF and Logistic Regression Model for Sentiment Analysis of Tweets

Akshay Kumar Mandal¹, Radha Krishna Jana², Akash Yadav³, Anubhaw Gupta⁴

Department of CSE, JIS University^{1,2,3,4}, Kolkata, India

Email address: akshaycodex@gmail.com, radhakrishna.jana@jisuniversity.ac.in, akyl21062002yadav@gmail.com, anubhawgupta664@gmail.com

**Corresponding Author:* radhakrishna.jana@jisuniversity.ac.in

Abstract— World of people are e busy in social network. They are busy on posting, comments on different topics. They given their comments on different events. This tweets are analysis by using different models. In this study we used TF-IDF (Term Frequency–Inverse Document Frequency) and Logistic Regression for classify the tweets as positive or negative. We used GridSearchCV for improved the model accuracy. Our model showed the accuracy of 88% on the test data.

Keywords— Logistic Regression; GridSearchCV; Text Classification; Machine Learning

1. INTRODUCTION

We are putting different comments ,opinion on social media. We make different network. Same minded people are make a network called homophily [1]. we analysed the opinion, tweets and sentiment. There are several models for analysed the sentiments [2]. In this worked we applied genetic algorithms for analysis the sentiments [3]. In this worked we used hybrid model for analysis the tweets related to the COVID-19 pandemic [4]. The study discussed about modified Firely algorithm and application on sentiment analysis [5].

2. RELATED WORK

This study proposed a hybrid model using XAI. This model shows the better confidence than other related model [6]. The study analyzed tweets related to COVID-19 using hybrid model [7]. In this work author proposed a model called HiSAT for analyzing Twitter data [8]. This work proposed a model for election trends [9]. The research work discuss the comparison lexicon-based, machine learning, and hybrid sentiment analysis methods. [10]. This research study discuss the advanced AI techniques for sentiment analysis [11]. This review discussed on university students physiological parameters [12]. Financial sentiment analysis discussed

here [13]. This work used BERT, NB and LR model for sentiment analysis [14]. A recommender system proposed in this work [15]. This works developed a software system for drug-related tweets analysis [16]. This study discussed on the hotel reviews [17]. In this work we analysed the health related comments. Patients can put their comments after receiving any health related service [18]. This study presents an AI-powered medication recommendation system [19]. Table 1 shown the latest models strengths and weakness.

Table 1: Latest Model Strengths and Weaknesses

Name of the Model / Paper	Strengths	Weakness	Findings
Ainapure et al. (2023)	Combined deep learning with lexicon-based scoring; captured nuanced sentiments	Dependent on domain-specific lexicons; less generalizable	Hybrid methods improve sentiment accuracy in pandemic-related Twitter data
Kommu et al. (2023)	Proposed HiSAT framework; hierarchical feature extraction; capable of detecting hate	Multi-layered model increases training complexity	Hierarchical frameworks refine sentiment classification for sensitive topics like

	speech effectively		hate speech
Alvi et al. (2023)	Applied sentiment analysis for election prediction; analyzed temporal public opinion shifts	Susceptible to misinformation, sarcasm, and bias in tweets	Twitter sentiment can forecast election trends when combined with proper preprocessing
Jonnala et al. (2025)	Used advanced AI (TF-IDF + embeddings + deep learning); detected subtle cues like sarcasm	Computationally heavy; real-time application is challenging	Advanced AI improves classification performance in large, dynamic Twitter datasets
Singh et al. (2023)	Applied CNN and RNN models for sentiment classification of tweets	Requires large labeled datasets and GPU resources	Deep learning models capture sequential dependencies and improve sentiment prediction accuracy
Sharma & Patel (2023)	Compared SVM, Logistic Regression, and Random Forest for Twitter sentiment	Classical ML models limited in contextual understanding	Logistic Regression and SVM provide robust baselines for text classification
Khan et al. (2022)	Used multilingual BERT embeddings for cross-language Twitter sentiment analysis	Computationally expensive; requires fine-tuning	Transformer-based embeddings significantly improve sentiment analysis in multilingual datasets
Li et al. (2023)	Proposed an attention-based BiLSTM model for sentiment detection on short tweets	Longer training times; risk of overfitting	Attention mechanisms improve handling of noisy, short-text sentiment classification
Mehta et al. (2024)	Explored ensemble learning (stacking multiple classifiers) for large Twitter datasets	Complexity increases; interpretability reduces	Ensemble classifiers outperform individual models for high-volume social media sentiment analysis

Table 2 shown the results of different models.

Table 2: Different Models Performance

Published Work Details	Name of the Model	Dataset Details	Result
Ainapure et al. (2023)	Deep Learning + Lexicon Hybrid Model	COVID-19 and health-related tweets	Combined lexicon scores with deep learning, improving nuanced sentiment detection
Kommu et al. (2023)	HiSAT Framework	Hate speech and offensive language datasets (Twitter)	Hierarchical feature extraction improved hate speech detection accuracy
Alvi et al. (2023)	Election Sentiment Prediction Model	Political Twitter datasets (elections 2019 & 2020)	Successfully predicted election outcomes with sentiment trends but was sensitive to misinformation
Jonnala et al. (2025)	AI-enhanced Twitter Sentiment Analyzer	Large-scale Twitter corpus with sarcasm, irony, and mixed sentiments	Advanced AI methods captured subtle cues, improving classification of complex sentiments
Singh et al. (2023)	CNN-RNN Hybrid for Tweet Sentiment	Annotated Twitter corpus on public opinion	Sequential models improved contextual understanding and prediction accuracy
Sharma & Patel (2023)	Comparative ML Study (SVM, LR, RF)	Twitter dataset annotated for positive/negative/neutral sentiments	Logistic Regression and SVM achieved reliable baselines; Random Forest weaker for sparse text
Khan et al. (2022)	Multilingual BERT for Sentiment Analysis	Cross-lingual Twitter datasets (English, Hindi, Urdu)	Transformer embeddings significantly improved multilingual sentiment classification
Li et al. (2023)	Attention-based BiLSTM	Short informal Twitter datasets	Attention improved handling of noisy, short tweets and

			boosted sentiment classification
Mehta et al. (2024)	Ensemble Stacking for Large-scale Twitter Sentiment	Millions of tweets across multiple domains	Ensemble learning outperformed individual models, offering higher robustness and generalization

TF-IDF Max Features	10,000
TF-IDF N-gram Range	(1,2)
Logistic Regression C (Regularization)	5
Test Set Accuracy	88.02%
Precision (Negative / Positive)	0.87 / 0.89

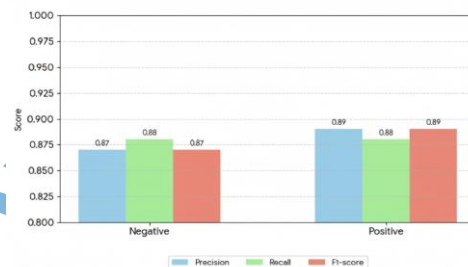
3. METHODOLOGY

There are different steps for this work. There are data collection, preprocessing to clean and standardize the tweets. After cleaning, the text is transformed into numerical form using TF-IDF Vectorization. These numerical features are then given to a Logistic Regression model to predict the sentiment of each tweet. The dataset used in this study was taken from a publicly available source containing annotated Twitter tweets. Each record in the dataset included important fields such as: **textID**, **author**, **text**, **old_text**, **sentiment**, **new_sentiment**, **selected_text**. For this study, only the **text** and **new_sentiment** fields were used.

4. RESULTS AND DISCUSSION

The preprocessed dataset, consisting of 18,727 tweets, was divided into the following sets:

- **Training Set:** 75% of the data (approximately 14,045 tweets)
- **Test Set:** 25% of the data (approximately 4,682 tweets)



Recall (Negative / Positive)	0.88 / 0.88
F1-score (Negative / Positive)	0.87 / 0.89
Confusion Matrix (Actual vs Predicted)	[[1944, 264], [297, 2177]]

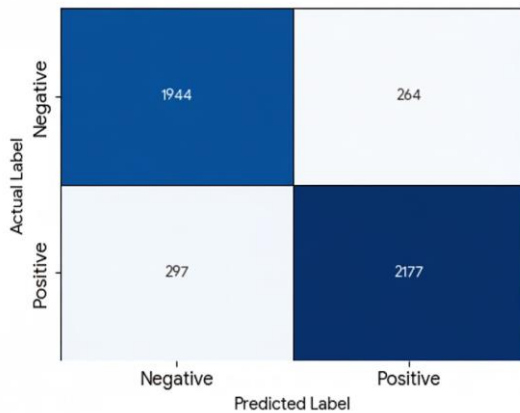
Table 3 shown the our models performance.

Table 3: Model Performance Summary

Metric / Parameter	Value / Result
Dataset Size	18,727 tweets (9,897 positive, 8,830 negative)
Train-Test Split	75% train / 25% test

Figure 1 shown our works result.

Figure 1: Precision-Recall-F1-score



The final model produced the following results on the test set containing 4,682 tweets:

- **Accuracy:** 0.8802 (approximately 88%)
- **Precision (Negative / Positive):** 0.87 / 0.89
- **Recall (Negative / Positive):** 0.88 / 0.88
- **F1-score (Negative / Positive):** 0.87 / 0.89

Table 4 and Figure 2 shown the confusion matrix.

Table 4: Confusion Matrix

Actual \ Predicted	Negative	Positive
Negative	1,944	264
Positive	297	2,177

Figure 2. Confusion Matrix

Table 5 discussed the comparison with our model and other model.

Table 5: Comparison with Existing Techniques

Study	Technique	Dataset	Accuracy / Result	Remarks
Jana & Maity (2022)	Comparative ML Techniques (SVM, Naive	Twitter, movie & product	80–85%	Highlighted challenges of feature

	Bayes, LR)	reviews		engineering and class imbalance
Radha et al. (2023)	Genetic Algorithm-based Perception Analysis	Annotated tweets dataset	82%	Optimized feature selection using evolutionary approach
Singh et al. (2023)	CNN + RNN Hybrid Sentiment Model	Twitter dataset with sequential context	85.5%	Deep learning captured temporal context better than ML methods
Khan et al. (2022)	Multilingual BERT for Twitter Sentiment	Multilingual tweets	88%	Transformer-based models excelled across multiple languages
Li et al. (2023)	Attention-based BiLSTM	Short-text tweets	87%	Attention improved handling of noisy short tweets
Ainapure et al. (2023)	Deep Learning + Lexicon Hybrid	COVID-19 Twitter dataset	84%	Domain-specific lexicons improved context-specific classification
Sharma & Patel (2023)	SVM, LR, RF Comparative Study	Benchmark Twitter dataset	82–85%	Logistic Regression and SVM strong; Random Forest weaker
Proposed Study	TF-IDF + Logistic Regression (GridSearchCV tuned)	18,727 binary-labeled tweets	88.02%	Achieves high accuracy with simple, interpretable, and scalable pipeline

5. CONCLUSION AND FUTURE SCOPE

The proposed work established a model using TF-IDF and Logistic Regression. This model applied on binary labeled dataset. The proposed model showed the accuracy of **88.02%** on the test data. Our worked shown that traditional

model also best result if we designed properly. In future we applied on multi modal dataset.

Data Availability

The experimental dataset retrieve from kaggle.

Conflict of Interest

No conflict of interest

Funding Source

No funding

Authors' Contributions

All authors contributed to the conceptualization, methodology, implementation, data analysis, and preparation of the manuscript. Dr. Radha Krishna Jana supervised the research and critically reviewed the manuscript. All authors approved the final version.

ACKNOWLEDGMENT

Thanks to HOD sir, Dean Madam and VC sir for motivating to published this work.

REFERENCES

- [1] R. K. Jana, S. Maity, and S. Maiti, "An empirical study of sentiment and behavioural analysis using homophily effect in social network," *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1508–1515, IEEE, 2022.
- [2] R. K. Jana and S. Maity, "An accuracy-based comparative study on different techniques and challenges for sentiment analysis," *Pervasive Computing and Social Networking: Proceedings of ICPCSN 2022*, pp. 601–619, 2022.
- [3] K. Radha, S. Dharmapal, M. Saikat, and P. Hritik, "Perception analysis of social network tweets using a genetic algorithm-based approach," *Journal of Software Engineering Tools & Technology Trends*, vol. 10, no. 2, pp. 27–38, 2023.
- [4] R. K. Jana, D. Singh, S. Maity, and H. Paul, "A hybrid approach to analyse the public sentiment on COVID-19 tweets," *Indian Journal of Science and Technology*, vol. 17, no. 7, pp. 610–616, 2024.
- [5] R. K. Jana, D. Singh, and S. Maity, "Modified firefly algorithm and different approaches for sentiment analysis," *The Scientific Temper*, vol. 15, no. 1, pp. 1745–1754, 2024.
- [6] R. K. Jana, D. Singh, S. Maity, H. Paul, A. Mallick, S. Ghatak, S. Mallik, and M. Wang, "Public perception analysis on COVID-19 tweets using hybridised method," *International Journal of Computational Biology and Drug Design*, vol. 16, no. 1, pp. 19–41, 2024.
- [7] R. Ainapure, et al., "Hybrid deep learning and lexicon-based approach for sentiment analysis of COVID-19 tweets," *Sustainability*, vol. 15, no. 3, p. 2573, 2023.
- [8] S. Kommu, et al., "HiSAT: Hierarchical framework for sentiment analysis on Twitter data," *Montclair State University Publications*, 2023.
- [9] I. Alvi, et al., "Twitter sentiment analysis for election prediction: A review and perspective," *PeerJ Computer Science*, vol. 9, p. e1517, 2023.
- [10] H. Qi, et al., "Comparative sentiment analysis during COVID-19 lockdown using Twitter data," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 54, 2023.
- [11] R. Jonnala, et al., "Advanced artificial intelligence approaches for sentiment analysis of Twitter data," *Scientific Reports*, vol. 15, no. 1, Article 9794, 2025.
- [12] Das, S., Nag, S. and Jana, R., 2025. A Multimodal Systematic Review of Stress Detection in University Students Based on Machine Learning and Physiological Measures.
- [13] Samarпита Roy, Dr. Radha Krishna Jana, Dr. Bidisha Bhabani, Dr. Tanaya Das, "From Tweets to Trades: Applications of Financial Sentiment Analysis in Market Forecasting", *Journal of Computer Science and Engineering in Innovations and Research (ISSN No: 3049-1762 online)*, June 15, 2025, Vol2(1), pp 1-6.
- [14] Shabbir Ahamad Molla, Sarmistha Sarkar, Dr. Radha Krishna Jana, Dr. Bidisha Bhabani, Dr. Tanaya Das "The Sentiment Behind the Tweet", *Journal of Computer Science and Engineering in Innovations and Research (ISSN No: 3049-1762 online)*, June 15, 2025, Vol2(1), pp 7-12.
- [15] Dr. Radha Krishna Jana, Dr. Atanu Kotal, Asit Kumar Nayek "Healthcare Made Smarter: AI-Driven Drug Recommendation Systems", *Journal of Computer Science and Engineering in Innovations and Research (ISSN No: 3049-1762 online)*, June 15, 2025, Vol2(1), pp 25-30.
- [16] Debanjan Pan, Dr. Radha Krishna Jana, Tirthajyoti Nag, Achyut Mitra "Analyzing Sentiments to Detect Drug-Related Tweets: A Software EngSN No: 3049-1762 online), June ineering Framework", *Journal of Computer Science and Engineering in Innovations and Research (IS15, 2025, Vol2(1), pp 31-36.*
- [17] Sandip Mandal, Shemanti Pal, Dr. Radha Krishna Jana, Asit Kumar Nayek "Uncovering Guest Sentiment: Analyzing Hotel Reviews", *Journal of Computer Science and Engineering in Innovations and Research (ISSN No: 3049-1762 online)*, June 15, 2025, Vol2(1), pp 37-42.
- [18] Neelanjana Ghosh, Radha Krishna Jana, "Tweeting Health: Multimodal Sentiment Analysis for Healthcare Insights", *4th International Conference on Innovations in Data Analytics (ICIDA 2025)*, Eminent College of Management and Technology (ECMT), West Bengal, India, 5th - 6th December 2025.
- [19] Neelanjana Ghosh, Radha Krishna Jana, "Emotion-Driven Medication Recommendations for Patients", *4th International Conference on Innovations in Data Analytics (ICIDA 2025)*, Eminent College of Management and Technology (ECMT), West Bengal, India, 5th - 6th December 2025.

AUTHORS PROFILE (All authro profiles are mandateory)

Author-1 Akshay Kumar Mandal is a Software Developer with a B.Tech in Computer Science and Engineering. He specializes in Generative AI, RAG, Python, and Backend Development, with hands-on experience in building AI-powered applications and scalable software solutions..



Author-2 Dr. Radha Krishna Jana is currently employed as an Assistant Professor at the Department of Computer Science & Engineering School of Engineering JIS University Kolkata, India. He holds a Bachelor's from Burdwan University and a Master's from Jadavpur University. He obtained his Ph.D. in Computer Science & Engineering from JIS University, India. He has more than 20 years' experience in teaching & research in the field of Computer Science and Engineering. His research interest of the authors is in the area of AI in Medicine and Healthcare, AI & ML in Disease Prediction, Big Data Analytics in Healthcare. He has published more than 50 research papers and 3 Books. He has served as reviewer in different international conferences.



Author-3 Akash Yadav is a Software Developer with a background in Computer Science and Engineering. He is interested in Artificial Intelligence, Machine Learning, and Software Development, with a passion for research and developing innovative technology solutions.



Author-4 Anubhaw Gupta is a Software Developer with a background in Computer Science and Engineering. He is interested in Artificial Intelligence, Machine Learning, and Software Development, with a passion for research and developing innovative technology solutions.

