iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80486

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Sentivox: An End-to-End Scalable NLP-Based Twitter Sentiment Analysis System

Dr. Sanjay Balamwar[1], Prof. Atul Akotkar[2], Avish Walde[3]
[1]*Associate Scientist Autonomous Body Of Planning Department, Government Of Maharashtra, Maharashtra Remote Sensing Application Centre, Maharashtra, India*
[2]*Professor Department Of Computer Science And Engineering, Nagarjuna Institute Of Engineering Technology And Management, Maharashtra, India*
[3]*UG Student, Department Of Computer Science and Engineering, Nagarjuna Institute Of Engineering Technology And Management, Maharashtra, India*

Abstract: ***This paper presents a scalable Twitter sentiment analysis system using Natural Language Processing (NLP) and Logistic Regression. A dataset of approximately [1]6 million labeled tweets is utilized for binary sentiment classification. Data preprocessing includes tokenization, stopword removal, and text normalization, followed by TF-IDF vectorization and vector normalization to achieve effective feature representation. Logistic Regression is selected due to its computational efficiency and strong performance on high-dimensional data. The system is deployed using Streamlit to enable real-time sentiment prediction. The results show that, although deep learning models may achieve higher accuracy, but they are prone to overfitting; in contrast, the proposed approach ensures balanced performance with lower computational cost and improved generalization.***
Keywords: Sentiment Analysis, Natural Language Processing, Logistic Regression, Python, Streamlit, TF-IDF, Text Classification, Deep Learning, Supervised Machine Learning

I. INTRODUCTION

Social media platforms like Twitter are rich sources of public opinion and sentiment data. Sentiment analysis, a subfield of Natural Language Processing (NLP), focuses on identifying and categorizing opinions expressed in text. This research addresses the challenge of analyzing large volumes of Twitter data to extract meaningful insights. The study proposes a scalable system using Python and Logistic Regression on a dataset of [1]6 million tweets. It incorporates preprocessing techniques such as emoji conversion and TF-IDF feature extraction, along with a Streamlit-based interface for real-time sentiment prediction. Logistic Regression, a supervised learning algorithm, provides a balance between interpretability and performance.

II. PROBLEM STATEMENT

Analyzing Social media platforms such as Twitter generate vast amounts of unstructured textual data, making sentiment analysis challenging. Informal language, slang, abbreviations, emojis, and noise further complicate accurate classification. Additionally, the large volume of approximately [1]6 million tweets creates scalability and processing challenges. Traditional deep learning approaches require high computational resources. Therefore, the problem is to develop an accurate, interpretable, and efficient sentiment classification model using NLP and Logistic Regression with real-time prediction capability.

III. OBJECTIVES

The primary objectives of this project are:

1) To develop a scalable Twitter sentiment analysis system using NLP techniques in Python.
2) To preprocess and clean a large-scale twitter dataset for sentiment classification,
3) To implement TF-IDF vectorization for effective feature extraction from textual data.
4) To evaluate a Logistic Regression model for accurate and interpretable sentiment classification.
5) To develop a web application using Streamlit for real-time sentiment prediction

IV. LITERATURE REVIEW

Sentiment analysis on social media data has been widely studied using lexicon-based, machine learning, and deep learning approaches. Traditional models such as Logistic Regression, Naïve Bayes, and SVM are preferred for their simplicity and

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

scalability. Large datasets like Sentiment140 emphasize preprocessing techniques and TF-IDF feature extraction. Deep learning models like BERT achieve higher accuracy but are prone to overfitting and require high computational resources. Streamlit supports efficient deployment of sentiment analysis systems. This project builds upon these approaches to develop an efficient and scalable Twitter Sentiment Analysis System.

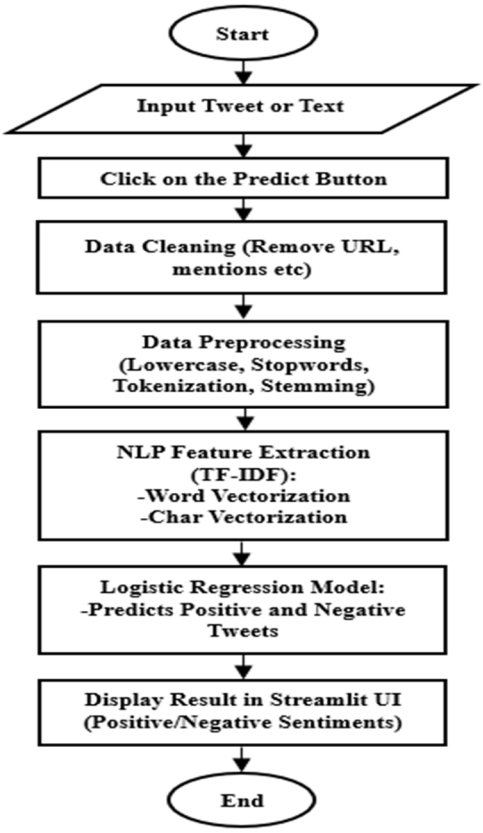
V. METHODOLOGY

The system follows a modular architecture typical of machine learning-based NLP applications, where stages such as preprocessing, feature extraction, model training, and deployment are logically organized. Python is used as the primary development language. A large-scale dataset of approximately [1]6 million labeled tweets from Kaggle (Sentiment140) is used for training and evaluation. The implementation consists of a Jupyter Notebook for model development and a Streamlit-based application for real-time prediction.

A. Core Modules

1) Data Preprocessing Module: Cleans tweets using tokenization, stopword removal, normalization, URL and mention removal, and emoji-to-text conversion.
2) Feature Extraction Module: Applies TF-IDF using word- and character-level features with weighted combination and normalization.
3) Model Training Module: Uses Logistic Regression for classification.
4) Evaluation Module: Uses accuracy, precision, recall, and F1-score.
5) Prediction Module: Performs threshold-based sentiment prediction.
6) Deployment Module: Provides a Streamlit-based user interface with result visualization and confidence score
.

VI. SYSTEM DESIGN



Workflow:

1) User Input: Tweet entered via Streamlit for sentiment analysis.
2) Text Preprocessing: input text cleaned using NLP lowercasing URL removal stemming.
3) Feature Extraction: word-level and character-level TF-IDF weighted features extracted.
4) Vector Normalization: feature vector normalized for consistent scaling.
5) Model Prediction: Logistic Regression predicts sentiment with probability threshold.
6) Result Visualization: Streamlit displays sentiment label, processed text, probability distribution graph, and confidence score to user interface output view.

VII. IMPLEMENTATION DETAILS

A. *Technologies Used*

1) Web Application Framework: Streamlit
2) Language: Python
3) Machine Learning Model: Logistic Regression
4) NLP Library: NLTK / spaCy for preprocessing
5) Feature Extraction: TF-IDF Vectorizer (word-level and character-level)
6) Data Handling: Pandas, NumPy
7) Dataset: Kaggle dataset containing ~1.6 million labelled tweets
8) Environment: Python, Jupyter Notebook, Streamlit runtime

B. *Logic Overview*

The system (Twitter_Sentiment_Analysis.ipynb) analyzes a large-scale tweet dataset using NLP techniques such as lowercasing, URL and mention removal, emoji conversion, normalization, and stemming.

The cleaned text is converted into numerical features using TF-IDF vectorization at word and character levels. These features are combined and used to train a Logistic Regression model for sentiment classification. The trained model (trained_model.sav and vectorizer.sav). is deployed using Streamlit for real-time prediction.

C. *User Interface*

The Streamlit interface (app.py) enables tweet input and displays predicted sentiment, confidence score, probability values, and visualization outputs.

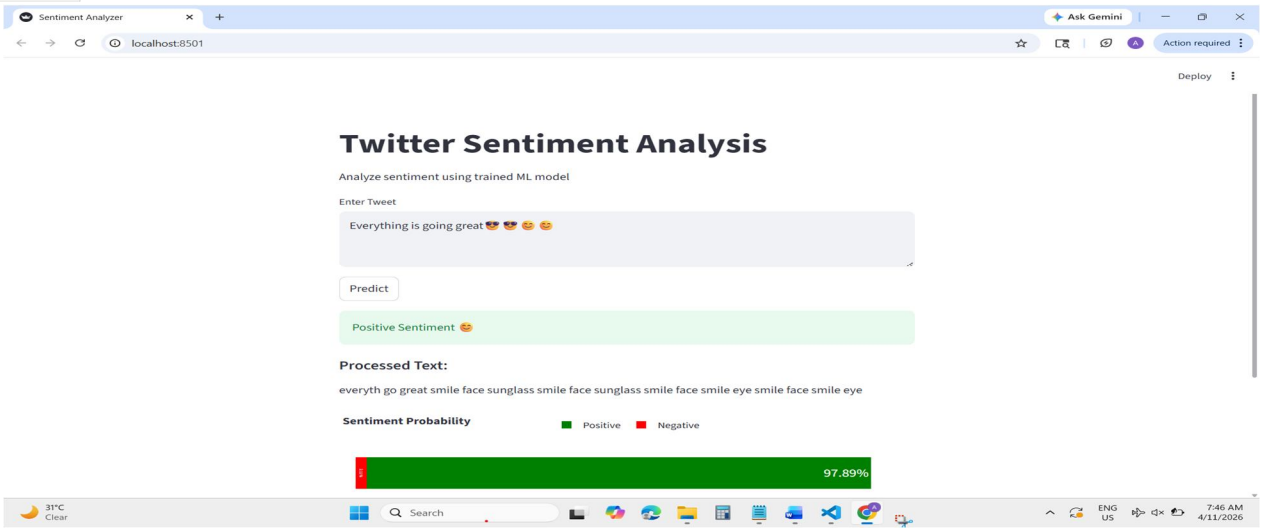
VIII. RESULT & DISCUSSION

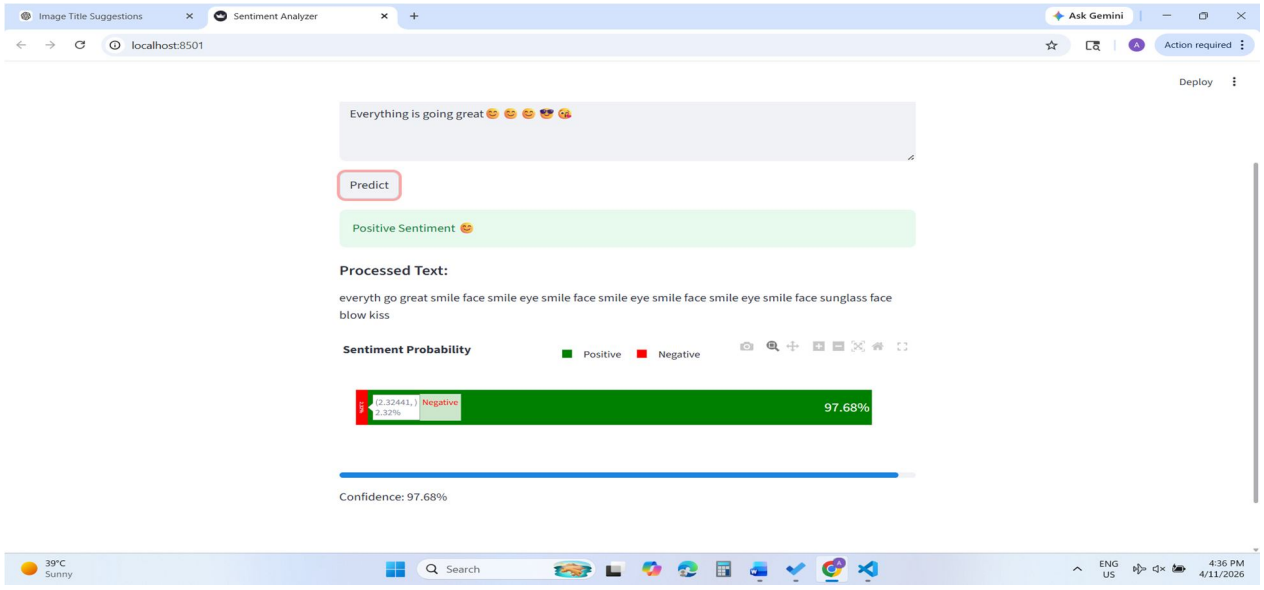
The developed Twitter Sentiment Analysis system effectively classifies tweets into positive and negative sentiment categories using a Logistic Regression model trained on a large-scale dataset of approximately [1]6 million labelled tweets from Kaggle.

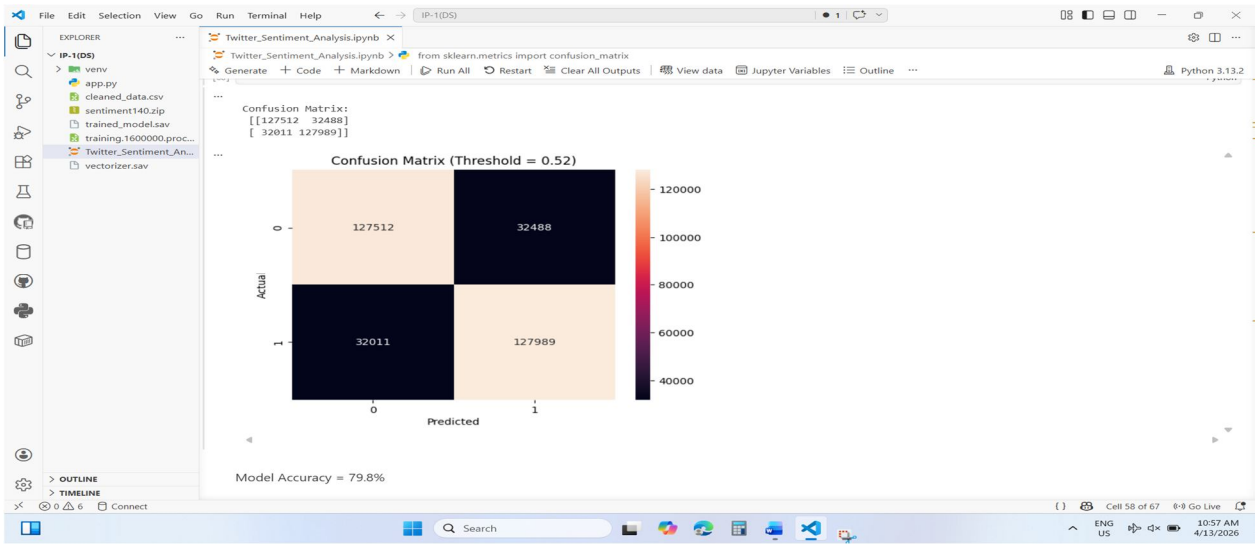
1) Sentiment Classification: The model successfully distinguishes between positive and negative tweets with reliable accuracy based on TF-IDF extracted features.
2) Real-Time Prediction: The Streamlit web application enables users to input tweets and instantly receive sentiment predictions along with probability scores.
3) Performance: The use of word-level and character-level TF-IDF vectorization improves feature representation, resulting in stable and consistent classification performance.
4) Model Efficiency: Logistic Regression provides fast training and prediction, making the system suitable for large-scale text data processing.
5) Visualization**:** The interface displays confidence scores and probability distribution graphs, improving interpretability of results.

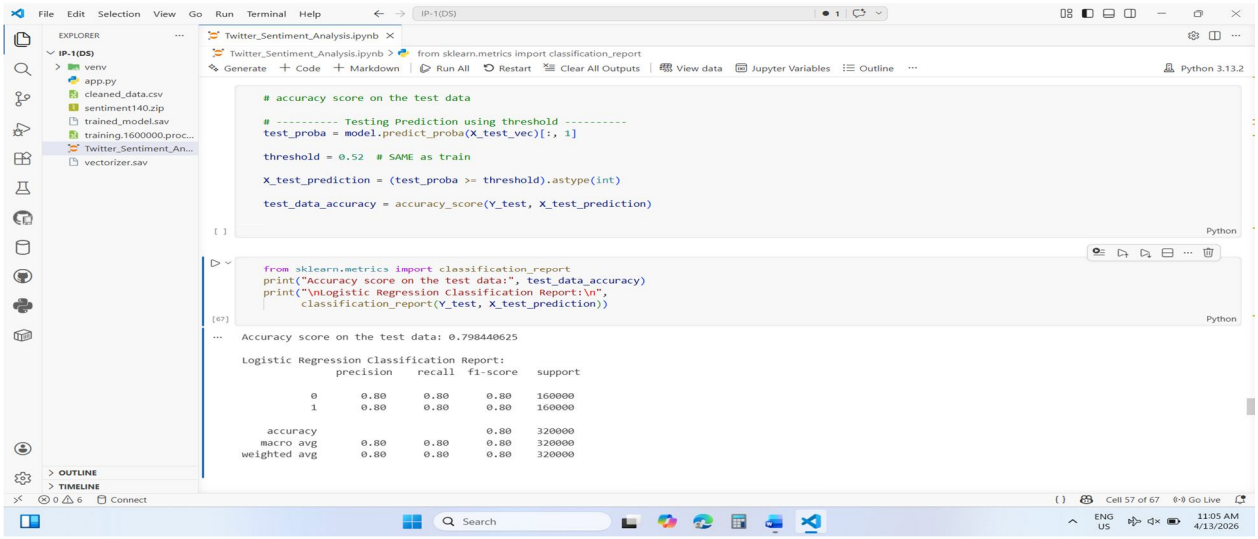
User testing demonstrated smooth execution, quick response time, and accurate sentiment prediction, confirming the system's effectiveness for real-time sentiment analysis applications.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com









IX. CONCLUSION

This project presents a Twitter Sentiment Analysis system using NLP, Logistic Regression, and Streamlit. Trained on a Kaggle dataset of 1.6 million labelled tweets, it enables large-scale text classification using TF-IDF features. The system supports real-time sentiment prediction and demonstrates that traditional machine learning with Streamlit integration is effective for deploying practical and scalable NLP applications.

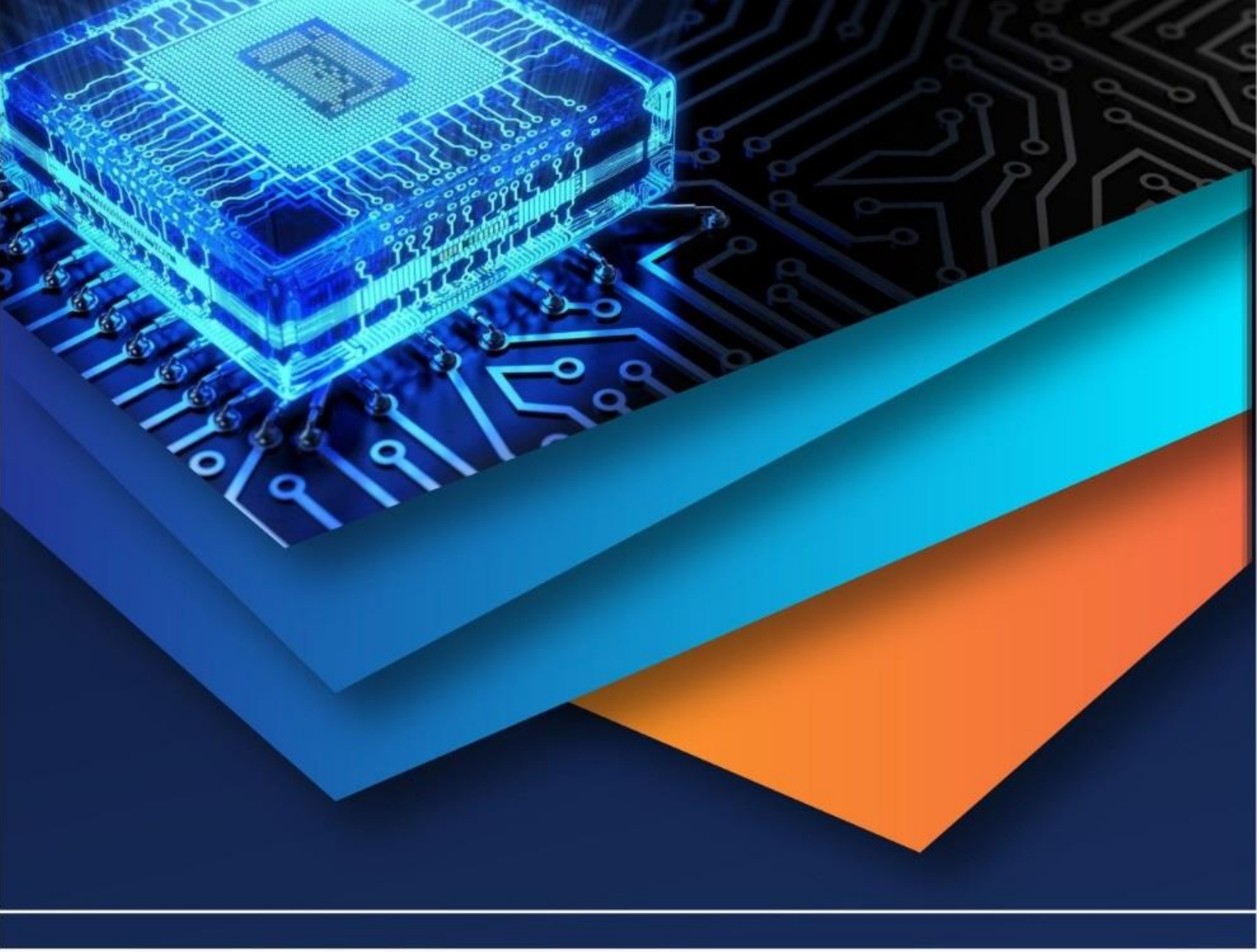
X. FUTURE SCOPE

Future research may focus on:

1) Incorporating deep learning models to capture nuances.
2) Enhancing preprocessing to better handle sarcasm, emojis and multilingual tweets.
3) Expanding the system to support multi-class sentiment analysis. (positive, negative, neutral)
4) Integrating real-time data streaming and updating models dynamically.
5) Leveraging domain-specific sentiment lexicons and hybrid machine learning approaches for improved accuracy.

REFERENCES

[1] Qutab, U. Fatima and I. Ahmed, International journal of innovative science and research technology, “Analyzing COVID-19 Sentiments on Twitter: An Effective Machine Learning Approach,” pp.841–850, Aug.2024 doi 10.38124/ijisrt/ijisrt24aug640.
[2] D. Kavitha, S. Venkatraman, K. CR, and N. S. Nair, Advances in systems analysis, software engineering, and high-performance computing book series, “Machine Learning-Based Sentiment Analysis of Twitter Using Logistic Regression,” pp. 308–319, June 2024, doi: 10.4018/979-8-3693-3502-4.ch020.
[3] E. Vladić, B. Mehanović, M. Novalić, D. Kečo, and D. Mehanović, Proceedings of the 3rd International Conference on NLP and Machine Learning Trends (NLMLT 2024), “Sentiment Classification of Tweets using Machine Learning and NLP Techniques,” Oct. 2024, pp. 35–45, doi: 10.5121/csit.2024.142004
[4] P. Dhanalakshmi, G. Ashish Kumar, B. Sai Satwik, K. Sreeranga, A. Tharun Sai, and G. Jashwanth, International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), "Sentiment Analysis Using VADER and Logistic Regression Techniques," Coimbatore, India, 2023, pp. 139-144, doi: 10.1109/ICISCoIS5654[1]2023.10100565.
[5] S. Raheja and A. Asthana, International journal of software innovation, “Sentiment Analysis of Tweets During the COVID-19 Pandemic Using Multinomial Logistic Regression,” vol. 11, no. 1, pp. 1–16, Jan. 2023, doi: 10.4018/ijsi.315740
[6] E. Cerrahoğlu and P. Cihan, International Conference on Pioneer and Innovative Studies (ICPIS), "Sentiment Analysis and Emojification of Tweets," vol. 1, pp. 481–486, June 2023. DOI: 10.59287/icpis.876
[7] A. Muslim, A. Benny, R. Refianti, C. Maisyarah, and G. Setiawan, International Journal of Advanced Computer Science and Applications, “Comparison of Accuracy between Long Short-Term Memory-Deep Learning and Multinomial Logistic Regression-Machine Learning in Sentiment Analysis on Twitter,” vol. 11, no. 2, Jan. 2020, doi: 10.14569/IJACSA.2020.0110294.
[8] R. Lakshmi, S. R. B. Divya, and R. Valarmathi, International journal of engineering and technology, “Analysis of sentiment in twitter using logistic regression,” vol. 7, p. 619, June 2018, doi: 10.14419/IJET.V7I2.33.14849.
[9] A. Gangawane, International Journal of Computer Applications, “Opinion Mining and Sentiment Analysis on Twitter,” vol. 182, no. 10, pp. 32–35, Aug. 2018, doi: 10.5120/IJCA2018917718.
[10] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” Stanford University, Stanford, CA, USA, Tech. Rep., 2009. [Online]. Available: https://www.kaggle.com/datasets/kazanova/sentiment140



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY