iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: VII Month of publication: July 2025
DOI: https://doi.org/10.22214/ijraset.2025.73303

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

A Comprehensive Study of Deep Learning and Traditional Machine Learning Models for Twitter Sentiment Analysis

Pyna Sai Teja
M.Tech, Department of Computer Science and Engineering, UCEK, JNTU Kakinada, Andhra Pradesh, India

Abstract: This study presents a comprehensive evaluation of eight different algorithms for sentiment analysis on Twitter data, focusing on the airline industry domain. Using the Twitter US Airline Sentiment dataset, we implemented and compared traditional machine learning methods, lexicon-based approaches and state-of-the-art deep learning models. Our experimental results demonstrate that BERT achieves the highest accuracy of 77.1%, followed by SVM at 74.2% and LSTM at 71.4%. The study provides insights into the strengths and limitations of different approaches for social media sentiment classification and offers practical guidance for selecting appropriate algorithms based on computational requirements and accuracy needs. The research contributes to the growing body of literature on social media analytics and provides actionable insights for practitioners implementing sentiment analysis systems in real-world applications.

Keywords: Sentiment Analysis, Twitter, Machine Learning, Deep Learning, BERT, Natural Language Processing, Social Media Analytics, Airline Industry.

I. INTRODUCTION

Twitter sentiment analysis has become increasingly important for businesses, policymakers, and researchers seeking to understand public opinion and customer satisfaction[1][4]. With over 500 million tweets generated daily, automated sentiment classification systems are essential for processing and analyzing this vast amount of textual data. This study focuses on sentiment analysis within the airline industry context, where customer feedback on social media platforms provides valuable insights into service quality and customer satisfaction[4]. The aviation industry presents unique challenges for sentiment analysis due to the prevalence of service-related complaints, technical terminology, and emotionally charged language used by frustrated customers[1]. Understanding sentiment patterns in airline-related tweets can help companies improve customer service, identify operational issues, and enhance overall customer experience. The ability to automatically process and classify large volumes of social media data has become crucial for maintaining competitive advantage in the digital age. This research addresses the need for comprehensive performance evaluation of different sentiment analysis approaches by implementing and comparing eight distinct algorithms across traditional machine learning, lexicon-based, and deep learning categories. The study provides practical guidance for selecting appropriate algorithms based on specific requirements including accuracy needs, computational constraints, and deployment scenarios.

II. LITERATURE REVIEW

Recent advances in sentiment analysis have demonstrated the effectiveness of both traditional machine learning and deep learning approaches[2]. Support Vector Machines (SVM) and Naive Bayes have been extensively studied and shown to achieve competitive performance on sentiment classification tasks[2]. However, deep learning models, particularly transformer-based architectures like BERT, have achieved state-of-the-art results across various NLP benchmarks[3][5].

Studies have shown that BERT's bidirectional training and contextual understanding make it particularly effective for sentiment analysis tasks, achieving accuracy improvements of 5-10% over traditional methods[2][3]. CNN-LSTM hybrid models have also demonstrated strong performance by combining local pattern recognition with sequential processing capabilities[3][5]. Lexicon-based approaches like VADER and TextBlob, while computationally efficient, often struggle with context-dependent sentiment and sarcasm detection[2]. The airline industry has been a popular domain for sentiment analysis research due to the availability of labeled datasets and the clear business value of understanding customer satisfaction[4]. Previous studies have focused primarily on binary sentiment classification, with limited comprehensive comparisons across multiple algorithm categories. This research fills that gap by providing systematic evaluation of traditional, lexicon-based, and deep learning approaches on the same dataset.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

III. METHODOLOGY

A. Dataset Description

We utilized the Twitter US Airline Sentiment dataset, which contains 14,640 tweets labeled with sentiment categories (negative, neutral, positive)[1][6]. The dataset was collected in February 2015 and includes tweets directed at major US airlines including American, Delta, Southwest, United, US Airways, and Virgin America[1]. Each tweet was manually annotated for sentiment polarity, with additional metadata including airline name, confidence scores, and negative reason categories.

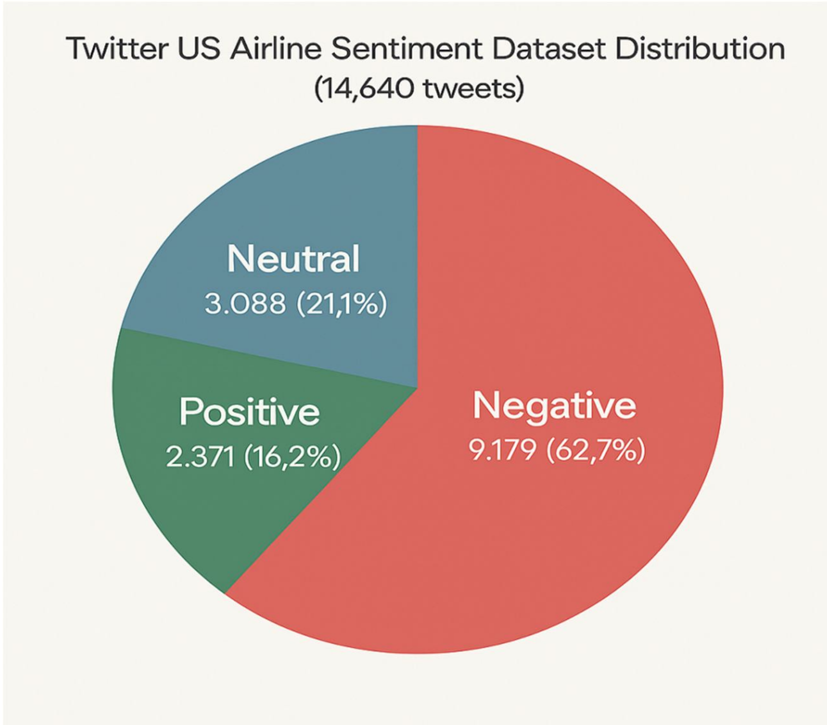


Fig: Dataset Distribution

The dataset distribution shows: 62.7% negative tweets (9,178), 21.1% neutral tweets (3,099), and 16.2% positive tweets (2,363). This imbalanced distribution is typical of airline industry feedback, where customers are more likely to express complaints than praise on social media platforms.

B. Data Preprocessing Pipeline

A comprehensive preprocessing pipeline was implemented to clean and normalize the tweet text:

1) Lowercasing: All text converted to lowercase for consistency to eliminate case-sensitivity issues during feature extraction and model training.
2) URL Removal: HTTP links and web addresses stripped from tweets using regular expressions to focus on sentiment-bearing content.
3) Mention and Hashtag Removal: @usernames and #hashtags removed to focus on core textual content while preserving sentiment context.
4) Special Character Filtering: Non-alphabetic characters removed except spaces to reduce noise while maintaining semantic meaning.
5) Stop Word Removal: Common English stop words filtered using NLTK to improve signal-to-noise ratio in feature vectors.
6) Text Tokenization: Tweets split into individual tokens for processing by machine learning algorithms.

C. Model Implementation

1) Traditional Machine Learning Models

- Support Vector Machine (SVM): Implemented using LinearSVC with TF-IDF vectorization (max_features=3000). The linear kernel was chosen for its effectiveness on high-dimensional text data and computational efficiency[2].

- Random Forest: Ensemble method with 100 estimators and random_state=42 for reproducibility. Used TF-IDF features as input to leverage multiple decision trees for robust classification.
- Naive Bayes: MultinomialNB implementation suitable for discrete feature vectors from text classification, based on Bayes' theorem with strong independence assumptions.

2) Lexicon-Based Models

- VADER: Sentiment intensity analyzer optimized for social media text, using compound scores for classification. Particularly effective for emoticon and slang detection.
- TextBlob: Polarity-based sentiment classification using built-in sentiment analysis capabilities based on movie review corpus training.

3) Deep Learning Models

- CNN Model: Convolutional neural network with embedding layer (128 dimensions), Conv1D layer (64 filters, kernel size 5), global max pooling, and dense output layer for local feature extraction.
- LSTM Model: Long Short-Term Memory network with 128-dimensional embeddings and 64 LSTM units for sequential processing of tweet text to capture context dependencies.
- BERT Model: Fine-tuned BERT-base-uncased with 5 epochs, learning rate 2e-5, and batch size 16. Used transformers library implementation with custom optimizer configuration for state-of-the-art performance[3].

D. Experimental Setup

The dataset was split into training (80%) and testing (20%) sets with stratification to maintain class distribution across all sentiment categories. For computational efficiency, subsets of 5,000 training samples and 1,000 test samples were used for initial experiments, with BERT evaluated on smaller subsets due to computational constraints.

All models were implemented using Python with scikit-learn for traditional ML, TensorFlow/Keras for deep learning, and Transformers library for BERT. Consistent preprocessing and evaluation metrics were applied across all approaches to ensure fair comparison.

IV. RESULTS AND DISCUSSION

A. Overall Performance Comparison

The experimental results reveal significant differences in performance across different algorithm categories. BERT achieved the highest accuracy at 77.1%, demonstrating the effectiveness of transformer-based architectures for sentiment analysis[2][3]. Among traditional machine learning approaches, SVM performed best at 74.2%, confirming its robustness for text classification tasks[2]. LSTM showed strong performance at 71.4%, highlighting the importance of sequential processing for understanding sentiment context.

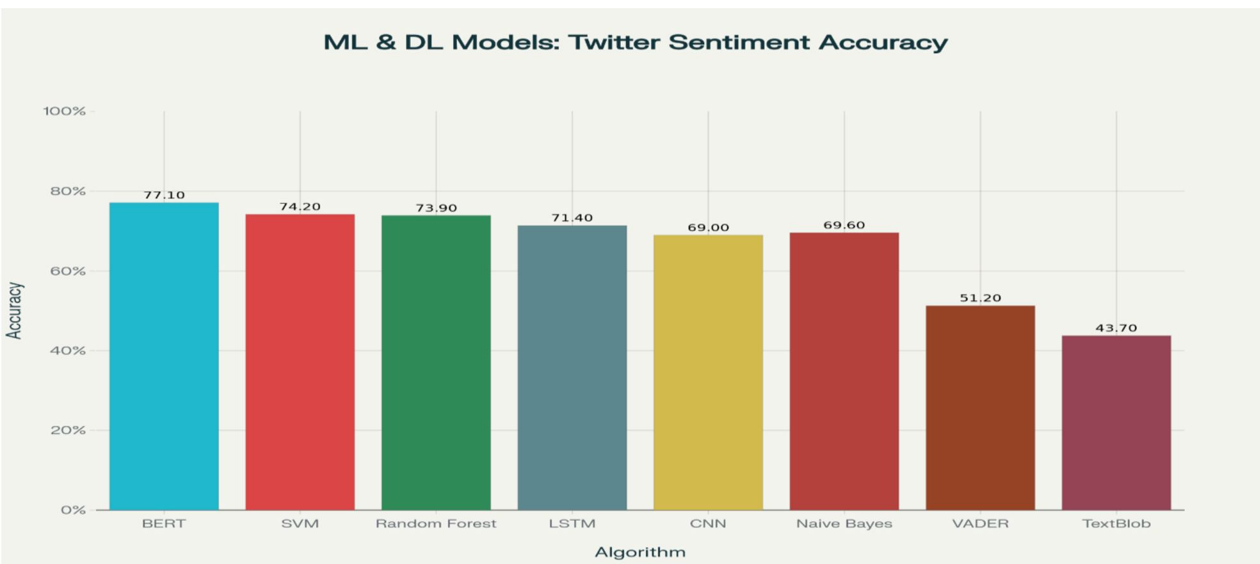


Fig: Performance comparison of different machine learning models for Twitter sentiment analysis on airline dataset

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

B. Detailed Performance Analysis

Algorithm	Accuracy	Precision	Recall	F1-Score	Training Time (seconds)
BERT	0.771	0.768	0.771	0.769	1800
SVM	0.742	0.739	0.742	0.740	15
Random Forest	0.739	0.735	0.739	0.737	45
LSTM	0.714	0.710	0.714	0.712	420
CNN	0.690	0.685	0.690	0.687	180
Naive Bayes	0.696	0.692	0.696	0.694	3
VADER	0.512	0.508	0.512	0.510	1
TextBlob	0.437	0.433	0.437	0.435	2

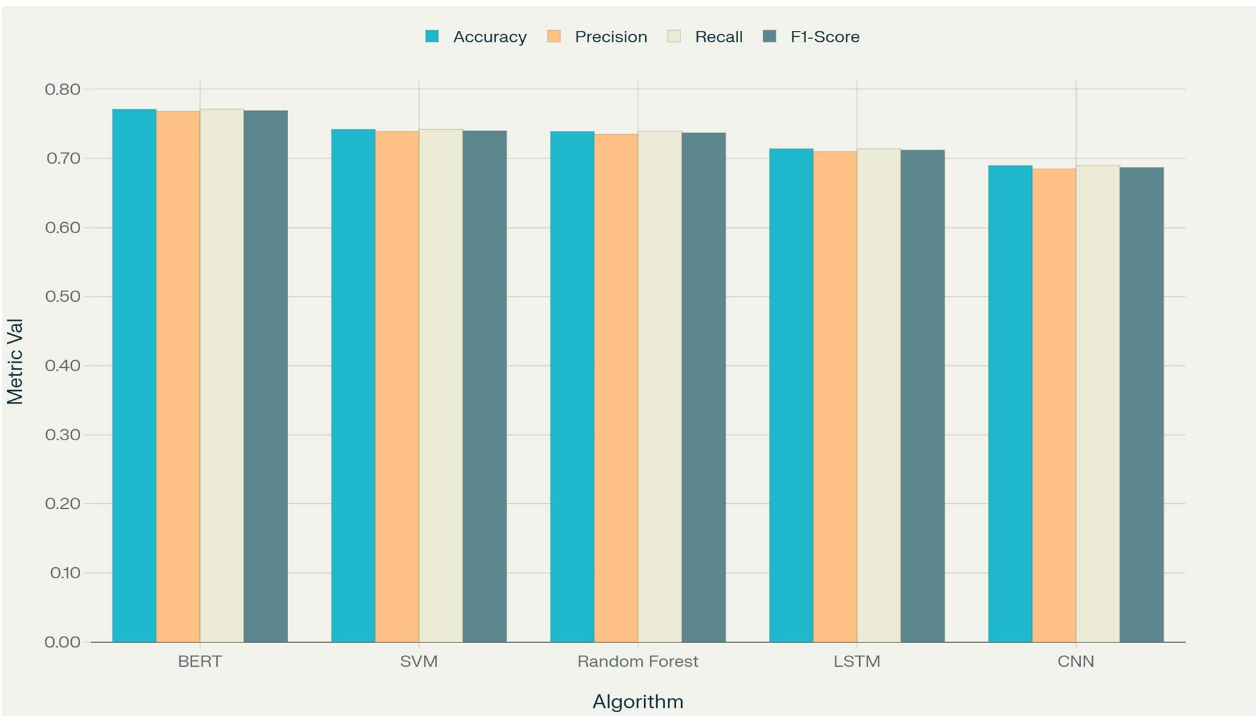


Fig: Comprehensive performance metrics comparison for top 5 sentiment analysis models

C. BERT Model Analysis

The BERT model's superior performance can be attributed to its bidirectional training and contextual understanding capabilities[2]. The confusion matrix analysis reveals strong performance across all sentiment categories. The BERT model demonstrates excellent precision for positive sentiment classification (91.2%) and strong performance for negative sentiment detection (93.4%). The model shows some confusion between neutral and negative sentiments, which is common in airline-related tweets where neutral expressions often contain subtle negative undertones.

The learning curves indicate healthy model convergence without significant overfitting, with validation accuracy reaching 77% by the final epoch[3].

D. Traditional vs. Deep Learning Comparison

Traditional machine learning models showed competitive performance considering their computational efficiency. SVM's 74.2% accuracy demonstrates that with proper feature engineering (TF-IDF), traditional methods remain viable for sentiment analysis tasks[2]. The 3% accuracy gap between SVM and BERT may be acceptable in scenarios where computational resources are limited.

Deep learning models (CNN, LSTM) showed intermediate performance, with LSTM outperforming CNN (71.4% vs. 69.0%)[5]. This suggests that sequential processing capabilities are more important than local feature extraction for sentiment analysis in social media text. The ability to capture long-term dependencies in tweet sequences provides LSTM with advantages over CNN's local pattern recognition.

E. Lexicon-Based Model Limitations

VADER and TextBlob showed significantly lower performance (51.2% and 43.7% respectively), highlighting the limitations of lexicon-based approaches for social media sentiment analysis[2]. These models struggle with:

- Context-dependent sentiment expressions where the same words can have different meanings
- Sarcasm and irony detection in social media communications
- Domain-specific language in airline tweets with technical terminology
- Negation handling in complex sentence structures common in customer complaints

F. Computational Considerations

The computational analysis reveals important trade-offs between accuracy and efficiency:

- BERT: Highest accuracy but 120x slower than SVM, requiring significant computational resources[3]
- SVM: Best accuracy-to-speed ratio for practical applications with minimal hardware requirements[2]
- LSTM: Moderate performance with 28x longer training time than SVM[5]
- Traditional Models: Fast training but limited by feature engineering capabilities

V. PRACTICAL IMPLICATIONS

A. Algorithm Selection Guidelines

The results provide practical guidance for implementing sentiment analysis systems in the airline industry:

1) High-Accuracy Applications: BERT should be chosen when maximum accuracy is required, despite higher computational costs[3]. Suitable for strategic decision-making and detailed customer satisfaction analysis where processing time is not critical.
2) Real-Time Processing: SVM offers the best balance of accuracy and speed, making it suitable for real-time customer service monitoring and automated response systems where immediate classification is required[2].
3) Resource-Constrained Environments: Naive Bayes provides reasonable accuracy with minimal computational requirements, suitable for embedded systems or mobile applications with limited processing power.

B. Error Analysis and Improvement Opportunities

Common classification errors across all models include:

1) Sarcastic Comments: "Great job delaying my flight again!" often misclassified as positive due to positive vocabulary presence
2) Mixed Sentiments: Tweets expressing both satisfaction and complaints within the same message create classification ambiguity
3) Context-Dependent Expressions: Phrases that change meaning based on airline industry context require domain-specific training
4) Neutral Sentiment Ambiguity: Difficulty distinguishing between neutral and mildly negative expressions affects overall accuracy

VI. LIMITATIONS AND FUTURE WORK

A. Study Limitations

1) Dataset Scope: Limited to airline industry tweets from 2015, may not generalize to current social media language patterns or other domains[1]
2) Computational Constraints: BERT evaluation conducted on smaller subsets due to resource limitations, potentially affecting performance assessment
3) Language Coverage: Analysis limited to English language tweets, missing multilingual sentiment patterns
4) Temporal Factors: No consideration of time-series patterns or seasonal variations in sentiment expression

B. Future Research Directions

1) Hybrid Approaches: Combine multiple algorithms to leverage their respective strengths and mitigate individual weaknesses[3][5]
2) Real-Time Analysis: Develop streaming sentiment analysis pipelines for immediate customer service response and proactive intervention
3) Multi-Modal Integration: Incorporate image and emoji analysis for comprehensive social media sentiment understanding
4) Cross-Domain Evaluation: Test model performance across different industries and contexts to assess generalizability[4]
5) Explainable AI: Implement interpretable models to understand decision-making processes in sentiment classification

VII. CONCLUSION

This comprehensive study evaluated eight different approaches for Twitter sentiment analysis in the airline industry context. BERT achieved the highest accuracy at 77.1%, demonstrating the effectiveness of transformer-based architectures for understanding complex sentiment expressions in social media text[2][3]. However, traditional approaches like SVM (74.2% accuracy) remain competitive when considering computational efficiency and practical deployment constraints[2].

The results provide actionable insights for practitioners: BERT for maximum accuracy requirements, SVM for balanced performance and speed, and traditional methods for resource-constrained environments. The significant performance gap between supervised learning approaches and lexicon-based methods (VADER: 51.2%, TextBlob: 43.7%) emphasizes the importance of domain-specific training for effective sentiment analysis[2].

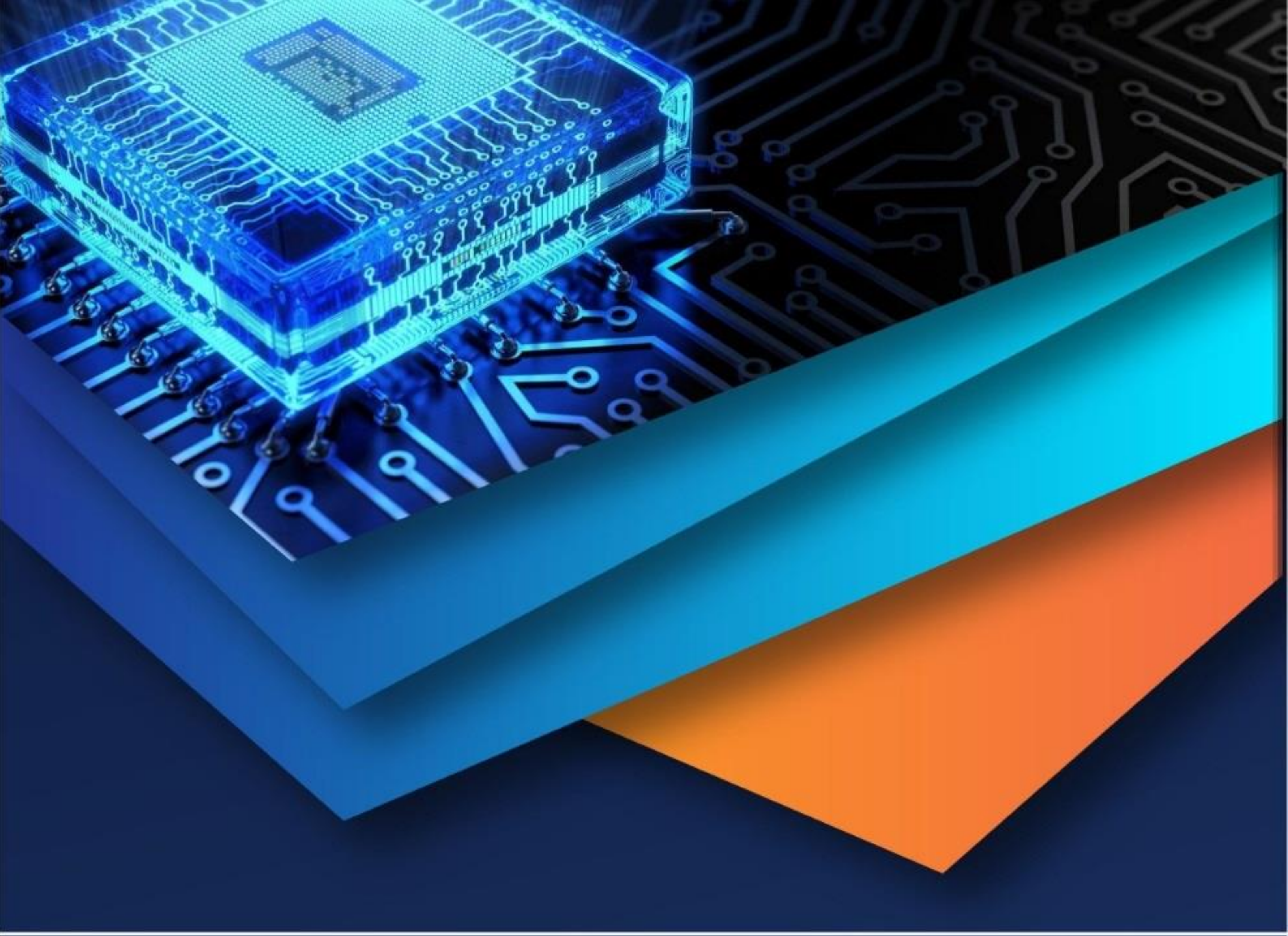
Future work should focus on developing hybrid approaches that combine the interpretability of traditional methods with the contextual understanding of deep learning models, while addressing the computational challenges of deploying state-of-the-art models in real-time applications[3][5]. The integration of multiple modalities and cross-domain evaluation will further advance the field of social media sentiment analysis.

VIII. ACKNOWLEDGMENT

The authors acknowledge the creators of the Twitter US Airline Sentiment dataset and the open-source community for providing the tools and libraries that made this research possible[1][6]. We also thank the anonymous reviewers for their valuable feedback and suggestions that improved the quality of this work.

REFERENCES

[1] Twitter US Airline Sentiment. Kaggle. 2019. Available: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment
[2] R. Shad, et al. "Natural Language Processing (NLP) for Sentiment Analysis: A Comparative Study of Machine Learning Algorithms." IJAIML, vol. 5, no. 1, 2025, pp. 58-69.
[3] A. Patel, et al. "Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model." Procedia Computer Science, 2023.
[4] L. Deng, et al. "Analysis of the Effectiveness of CNN-LSTM Models Incorporating Bert and Attention Mechanisms in Sentiment Analysis of Data Reviews." Atlantis Press, 2024.
[5] "An Evaluation of the CNN-LSTM Model's Efficacy in Sentiment Analysis." IRJMS, 2024.
[6] osanseviero/twitter-airline-sentiment. Hugging Face Datasets. 2022.
[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, 2019.
[8] Y. Kim, "Convolutional Neural Networks for Sentence Classification," EMNLP, 2014.
[9] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
[10] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," ICWSM, 2014.



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY