International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Comparative Study of Lexicon, Machine Learning, and Transformer-Based Models for Airline Sentiment Analysis

[1st] Ansh Jena, [2nd] Sujit Kakade, [3rd] Arya Kedar
Dept. of Artificial intelligence Vishwakarma University
Pune, India

Abstract- Sentiment analysis can help track passengers' per- ceptions and improve the service offered by an airline due to the increasing importance of social media, such as Twitter. It is about conducting a comparative analysis of three models of natural language processing, namely lexicon-based, machine learning, and transformer-based classification techniques for determining sentiments of airline tweets. Twitter US Airline Sentiment was chosen to be analyzed as it comprised labeled tweets from the major U.S. airlines. Data quality was improved by applying methods of text preprocessing, such as removing noise, tokeniz- ing, and eliminating stopwords. Lexicon-based sentiment analysis relied on VADER polarity baselines, machine-learning approach entailed extraction of TF-IDF features and further application of Random Forest classification technique while transformer model applied RoBERTa to identify the context of sentiment. As a result of the analysis, it was found out that while the lexicon model was faster and provided more easily understandable results, machine- learning model allowed identifying sentiments more accurately. Transformer-based RoBERTa performed the best in terms of handling more complex linguistic structures, such as negations and sarcasm.

Index Terms—Airline Sentiment Analysis, Natural Language Processing, Random Forest, RoBERTa, Sentiment Classification, VADER.

I. INTRODUCTION

With the rise of new digital media platforms like Twitter, social media becomes a powerful tool in obtaining customer feedback. Customer feedback becomes crucial in the field of aviation since the level of service and passenger experience play a critical role in determining the company's brand reputation. The traditional approach based on interpreting text through lexicons, such as the VADER model, has become widely utilized in sentiment classification on social media [1]. Sentiment analysis, also known as opinion mining, can be considered as one of the main NLP techniques for classifying texts into sentiments like positive, negative, and neutral [2], [3]. Nevertheless, understanding the text becomes complicated due to the fluidity of language and various forms of expressing different sentiments, including sarcasm [3]. The alternative technique in dealing with social media sentiment analysis involves machine learning approaches that use a combination of statistical features (e.g., TF-IDF) with classifiers like Random Forests or Support Vector Machines (SVM) to account for context [4]–[7], [31].

Moreover, transformer models such as RoBERTa have shown promising results when tackling the problems associ- ated with analyzing unstructured text data from social media [8]–[10]. Recent studies also highlight the effectiveness of deep learning approaches in domains such as fraud detection and public health analytics, reinforcing the applicability of advanced models in sentiment-related tasks [32], [33].

In spite of remarkable advancements in all of the three methodologies individually, there is an evident research void in comparative analysis of the aforementioned techniques using domain-specific datasets, for instance, aviation-related tweets [11], [12]. Additionally, prior comparative studies on machine learning algorithms further motivate the need for evaluating multiple approaches in a unified framework [34].

The primary objective of this study is to conduct a comparative study on lexicon-based model (VADER), ma- chine learning-based model (TF-IDF + Random Forest), and transformer-based models (RoBERTa) using the Twitter US Airline Sentiment dataset.

II. LITERATURE REVIEW

Various methodological paradigms of sentiment analysis have emerged over the years, with each having its own ad- vantages and disadvantages in understanding human opinions expressed

in text. Early work focused largely on lexicon- based approaches, where polarity scores are assigned based on predefined sentiment dictionaries. VADER is a rule-based model for sentiment reasoning developed by Hutto and Gilbert. 1 — was optimized for social media text, and it captures intensity and negation effectively. But it often struggles with sarcasm and subtle emotions. Other dictionary-based systems, such as TextBlob and SentiWordNet, have been used for short-text classification, but they do not have the contextual adaptability required for dynamic domains such as airline feedback [13], [20], [26].

The advent of Machine Learning has led to the emergence of data-driven methods for performing sentiment detection through the use of algorithms such as Support Vector Machines (SVMs), Na¨ıve Bayes, and Random Forests, with the use of document representation techniques such as Bag-of-Words and Term Frequency/Inverse Document Frequency (TF-IDF) providing improved accuracies [4], [7], [29]. Other studies have also been conducted to examine the use of Gradient Boosting for multi-class sentiment classification, providing higher levels of generalisation than lexicon-based approaches [11, 12]. However, one of the disadvantages of these Classical Models is that they rely extensively on labelled data while being constrained by the sparsity of the features and the size of the vocabulary.

The field of sentiment analysis has come a way with the introduction of deep learning and transformer-based architectures. Models like BERT and RoBERTa use a mechanism to understand the meaning of sentences and how they relate to each other. Research has shown that these models are better than methods at figuring out how people feel about things whether it is from movie reviews, product feedback or Twitter posts [16], [17], [22], [23].

Some models, like twitter-roberta-base-sentiment are particularly good at handling the language and sarcasm people use on Twitter but they can be hard to use in real-time because they require a lot of computer power and are not easy to understand [27].

When it comes to airlines many studies have looked at what people say about them on Twitter. Most of these studies have used methods like VADER or simple machine learning techniques like Logistic Regression and Random Forests [11], [12], [25]. A few recent studies have tried using transformer-based models to see how people feel about airlines and even fewer have compared these models to other methods.

So there is a need to compare the methods for figuring out sentiment including lexicon-based machine learning-based and transformer-based models, when it comes to what people say about airlines on Twitter. This paper wants to fill this gap by looking at how well VADER, TF-IDF + Random Forest and RoBERTa work and seeing what each one is good and bad, at and which ones are practical to use in the world to monitor how people feel about airlines.

III. METHODOLOGY

This part will figure out how we will compare three ways to analyze sentiments which’re lexicon-based, machine learning and transformer-based using tweets, about airlines. We broke down the process into five steps, getting the data, cleaning up the data, using lexicon-based method to score sentiments, using machine learning to classify sentiments, using transformer-based method to analyze sentiments.

We are comparing these three sentiment analysis methods, which’re lexicon-based, machine learning and transformer-based to see how they work with airline tweet data.

A. Dataset Description

We used the Twitter US Airline Sentiment Dataset that is available to the public. This Twitter US Airline Sentiment Dataset has than 14,000 tweets about the Twitter US Airline Sentiment. These tweets are six major airlines in the United States. American, United, Southwest, Delta, US Airways and Virgin America [11], [12], [25].Each tweet in the Twitter US Airline Sentiment Dataset is labeled as positive, negative or neutral. The researcher also looked at information like how many times a tweet was retweeted and when it was posted. The researcher picked a few columns from the Twitter US Airline Sentiment Dataset to work with. The text of the tweet the airline, the sentiment of the tweet and how many times it was retweeted. The Twitter US Airline Sentiment Dataset is a way to test how well sentiment models work with short texts from social media, like the Twitter US Airline Sentiment Dataset [18], [19], [20].

B. Data Preprocessing

The informal nature of Twitter language required extensive preprocessing to ensure good data quality. The initial code used in the pipeline included:

- Applying small capitalization to the text.
- Removing URLs, mentions, hashtags, and special characters using regular expressions.
- Tokenizing and filtering stopwords with the NLTK stopword list.

- Creating a downstream processing clean text column. This cleaning process reduced noise significantly and normalized text patterns while preserving important words relatedto sentiment.

C. Lexicon-Based Sentiment Analysis (VADER)
Valence Aware Dictionary of sEntiment Reasoning (VADER) was a model used as a baseline lexicon-based model. VADER produces a compound score for each text based on predefined lexical features and rules for negation and intensity. The Tweets were categorized based on three polarity levels.

- Positive: compound score > 0.05
- Negative: compound score < –0.05
- Neutral: otherwise

These findings were recorded as another column (vadersentiment) in the dataset to make it easier to compare with other models.

D. Machine Learning Classifier (TF-IDF + Random Forest)
A TF-IDF vectorizer with up to 5,000 features and a bi- gram range of 1 to 2 turned cleaned tweets into numerical representations that are suitable for machine learning. Ground truth labels, which indicate airline sentiment, were used. The data was split into training at 80 percent and testing at 20 percent, with a fixed random state to ensure reproducibility. A Random Forest Classifier with 200 estimators and balanced class weights was trained to sort tweets into positive, negative,

or neutral sentiments. We evaluated performance using precision, recall, F1-score, and accuracy. Predictions were stored for visual and statistical comparisons.

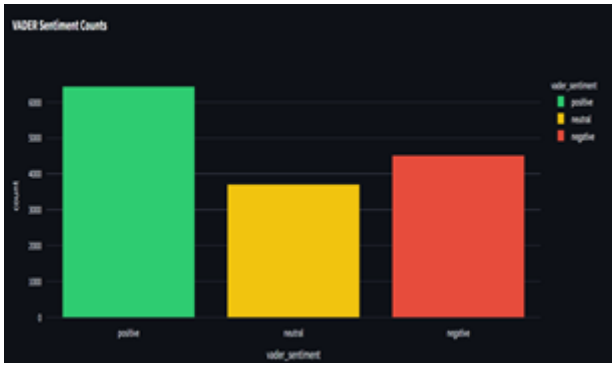

Fig. 1. VADER Sentiment counts

E. Sentiment Analysis (RoBERTa) based on Transformers
For a more in-depth look, the RoBERTa model fine- tuned on Twitter sentiment (cardiffnlp/twitter-roberta-base- sentiment) was used from the Hugging Face Transformers library. Self-attention mechanisms in RoBERTa enable the ex- traction of complex language patterns, sarcasm, and subtle sen- timent clues. A random sample of 500 cleaned tweets has been put through the pipeline, and the obtained label—positive, negative, or neutral—has been juxtaposed against other models to demonstrate the effectiveness of transformer-based NLP compared to traditional models.

F. Hard Determinism vs. Soft Determinism

To help in making a picture of the patterns for sentiment, exploratory analyses have been conducted, and these include;

- Sentiment Distribution: frequency graphs of prediction per each model. Clouds: showing the most frequent words used in each category of sensation.
- Retweet impact: boxplots of retweets' number by senti- ments
- Airline comparison – sentiment distribution in the multi- class across airlines.

They provided qualitative information on how sentiment varied with the airline, as well as on the differences in modelling interpretation.

G. Model Evaluation Metrics
Each of the models was evaluated using the standard NLP classification measures:

- Accuracy: proportion of sentiments predicated with ac- curacy.
- Precision, Recall and F1-score: to check the balance and robustness of the model.
- Confusion Matrix: to find out trend of misclassification The basis for the application of this multi-dimensional assessment was to furnish a fair comparison among the rule- based paradigm, statistical and deep learning paradigm.

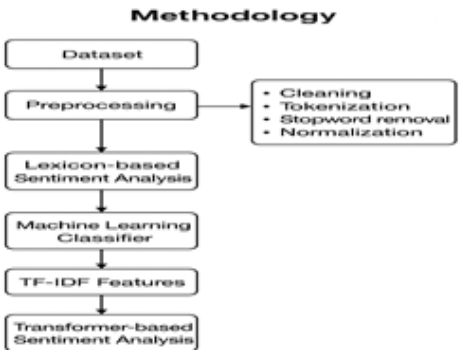

Fig. 2. Workflow of the proposed sentiment analysis showing data acquisition, preprocessing, and analysis using VADER, TF-IDF with Random Forest, and RoBERTa models.

TABLE I
PERFORMANCE COMPARISON OF REGRESSION MODELS

Model	RMSE	R^2
Linear Regression	1.06	0.001
Random Forest Regressor	1.07	-0.011
XGBoost Regressor	1.07	-0.011

TABLE II
PERFORMANCE COMPARISON OF CLASSIFICATION MODELS

Model	Accuracy
Logistic Regression	0.446
Random Forest Classifier	0.443
SVM	0.447
Gradient Boosting	0.444

IV. RESULTS AND DISCUSSION

This section therefore, provides the evaluation results of the three sentiment analysis paradigms: lexicon-based, machine learning based, and transformer based on the Twitter US Airline Sentiment Dataset. The focus of such comparative evaluation is on classification effectiveness, interpretability, and the context of emotions in tweets about airline services [27], [28]. This section offers the experimental results for all three types of sentiment analysis paradigms, namely lexicon- based, machine learning-based, and transformer-based senti- ment analysis on the Twitter US Airline Sentiment dataset [11], [12], [18], [19], [25]. The comparative analysis take into consideration classification effectiveness, interpretability, and situational interpretation of feelings shared in tweets relating to airlines.

The performance of the models was compared quantitatively based on accuracy, precision, recall, and F1-score measures [15]–[17], [21], [22]. The findings of each method have shown some differences in strengths and weaknesses.

Despite the efficiency and speed of the VADER model, it provided moderate results due to a limited understating of language context [1], [13], [20]. TF-IDF with Random Forest classifier exhibited further enhanced accuracy, as it was able to capture relationships of n-grams and trained on statistics-based recognition of subtleties of sentiments [4]-[6], [28], [29].

The transformer-based RoBErTa model was most successful overall, which is associated with its ability to understand intricate linguistic patterns, sarcasm, and concealed emotions. The transformer-based model outperformed traditional methods in all measured criteria and showed stability in its ability to determine sentiments from brief, informal, and context-dependent communications on social media [8]– [10], [22], [23]. The RF model yielded a reasonable tradeoff be-tween interpretability and performance, while VADER may remain appropriate for rapid and computationally inexpensive analyses [1], [28], [29].

A. Visual Insights

Finally, exploratory data visualizations added more informa-tion regarding the observations of sentiments across different airlines. A pie chart for the distribution of sentiments indicated many negative sentiments for various airlines like United and American Airlines. Word clouds extracted for each sentiment category confirmed that common words present in tweets indicating negative sentiments against airlines included delay, service, cancelled, baggage, while words such as great, thank and friendly dominated tweets indicating positive sentiments [11], [12].

Fig. 3. Retweets Distrubution Per Sentiment

Retweet analysis confirmed that negative tweets attract more engagement than positive and neutral tweets, thus confirming that customers are more likely to express their dissatisfaction on social networks [18]. A comparison across airlines showed that sentiment variation is influenced by brand reputation and quality of customer service [25].

B. Discussion and Interpretation.

From the comparative analysis, the key differences in the three approaches are highlighted. Lexicon-based models such as VADER are simple and interpretable but fail to capture sarcasm, negation, and informal language prevalent on social media [1], [13], [20]. Machine-learning models, for instance, TF-IDF (term frequency-inverse document frequency) combined with Random Forest (Statistical Machine Learning Ensemble Classifier), improve the performance by statistically associating the words but lack deeper contextualization [28], [29].

The RoBERTa transformer model shows a better understanding of semantic and syntactic connections through its self-attention modules which enable it to recognize complex forms of sentiment expression [8] – [10]. The deep contextual models prove to be more effective in understanding semantic and syntactic relations than rule-based and machine learning techniques using noisy short text data from social media channels.

Fig. 4. Most Commonly Words Used

For airlines, transformer-based sentiment analysis can apply to real-time reputation monitoring, customer service analytics, and feedback management. Last, a hybrid paradigm utilizing lexicon precision, statistical robustness, and deep contextual learning can produce an interpretable system of exceptional accuracy [1], [22], [23].

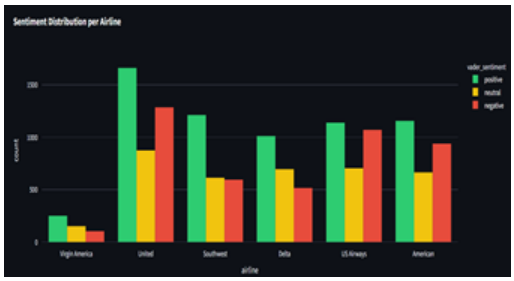

Fig. 5. Sentiment Distrubution Per Airline

V. CONCLUSION AND FUTURE WORK

The paper presented a comparison of three sentiment analysis paradigms – lexicon-based, machine learning-based and transformer-based approaches on the Twitter US Airline Sentiment dataset. Its priority was to test the efficacy of the traditional and the modern natural language processing approaches in classifying customers' opinions about the airlines' services. Thus, the study offered a short and general performance benchmark between the methods on a systematic workflow, including data preprocessing, lexicon-based scoring, TF-IDF based feature extraction, and context-aware modeling using transformers.

The experimental findings proved that while the lexicon based VADER model is fast and interpretable, it fails to capture context factors like sarcasm. The machine learning model (TF-IDF + Random Forest) had a higher accuracy by creating classifiers based on frequent association of terms, while the transformer-based RoBERTa model had the highest precision and F1-score by semantic contextual retrieval and emotional tone of the given text. Therefore, the transformer- based framework presents the best solution for sentiment analysis tasks in real-world settings on noisy social media data. For future work, sentiment monitoring pipelines with real- time capabilities may be developed to assist airlines in cus- tomer service management and reputation tracking. Sarcasm detection, detection of emotions, and aspect-based sentiment analysis can be incorporated (ibid) to provide a degree of contextual understanding. Finally, applying the models on multilingual data and fine-tuning them on domain-specific corpora would ensure adaptability to the global airline market. The result is that the sentiment analysis systems designed for the aviation industry would be scalable, inclusive, and practical.

REFERENCES

1. C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," Proc. Int. AAAI Conf. Web Soc. Media, vol. 8, no. 1, pp. 216–225, 2014.
2. A. Borg, "Using VADER sentiment and SVM for predicting customer satisfaction," Computers & Industrial Engineering, vol. 149, p. 106804, 2020.
3. M. T. H. Khan and M. T. Islam, "A comparative study of sentiment analysis using NLP and different machine learning techniques on US airline Twitter data," arXiv preprint arXiv:2110.00859, 2021.

4. X. Wang, "A comparative experimental study of citation sentiment classification using TF-IDF and machine learning models," Journal of Information Science, vol. 51, no. 2, pp. 232–245, 2025.
5. A. Setiawan, "Utilizing Random Forest algorithm for sentiment predic- tion on Twitter data," Atlantis Press, 2022.
6. B. Das, "An improved text sentiment classification model using TF-IDF and machine learning algorithms," arXiv preprint arXiv:1806.06407, 2018.
7. M. Karim, "Comprehension of polarity of articles by citation sentiment analysis using TF-IDF and machine learning," Journal of Information Science, vol. 48, no. 6, pp. 741–755, 2022.
8. A. Gaurav, "XLM-RoBERTa based sentiment analysis of tweets on Metaverse discussions," Procedia Computer Science, vol. 187, pp. 123–130, 2024.
9. N. A. Semary, "Improving sentiment classification using a RoBERTa- based hybrid model," Journal of King Saud University-Computer and Information Sciences, 2023.
10. M. R. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa- BiLSTM: A context-aware hybrid model for sentiment analysis," arXiv preprint arXiv:2406.00367, 2024.
11. A. Rane and A. Kumar, "Sentiment classification system of Twitter data for US airline service analysis," Proc. 42nd Annu. Computer Software and Applications Conf. (COMPSAC), vol. 1, pp. 114–121, 2018.
12. S. Rustam, M. A. Khan, and M. S. Hossain, "Tweets classification on the base of sentiments for US airline services," J. Electr. Eng. Technol., vol. 14, no. 1, pp. 1–9, 2019.
13. K. Barik, "Analysis of customer reviews with an improved VADER sentiment analysis model," J. Big Data, vol. 11, no. 1, p. 61, 2024.
14. M. T. H. Khan and M. T. Islam, "A comparative study of sentiment analysis using NLP and different machine learning techniques on US airline Twitter data," arXiv preprint arXiv:2110.00859, 2021.
15. A. Gaurav, "XLM-RoBERTa based sentiment analysis of tweets on Metaverse discussions," Procedia Computer Science, vol. 187, pp. 123–130, 2024.
16. N. A. Semary, "Improving sentiment classification using a RoBERTa- based hybrid model," Journal of King Saud University-Computer and Information Sciences, 2023.
17. M. R. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa- BiLSTM: A context-aware hybrid model for sentiment analysis," arXiv preprint arXiv:2406.00367, 2024.
18. A. Rane and A. Kumar, "Sentiment classification system of Twitter data for US airline service analysis," Proc. 42nd Annu. Computer Software and Applications Conf. (COMPSAC), vol. 1, pp. 114–121, 2018.
19. S. Rustam, M. A. Khan, and M. S. Hossain, "Tweets classification on the base of sentiments for US airline services," J. Electr. Eng. Technol., vol. 14, no. 1, pp. 1–9, 2019.
20. K. Barik, "Analysis of customer reviews with an improved VADER sentiment analysis model," J. Big Data, vol. 11, no. 1, p. 61, 2024.
21. M. T. H. Khan and M. T. Islam, "A comparative study of sentiment analysis using NLP and different machine learning techniques on US airline Twitter data," arXiv preprint arXiv:2110.00859, 2021.
22. A. Gaurav, "XLM-RoBERTa based sentiment analysis of tweets on Metaverse discussions," Procedia Computer Science, vol. 187, pp. 123–130, 2024.
23. N. A. Semary, "Improving sentiment classification using a RoBERTa- based hybrid model," Journal of King Saud University-Computer and Information Sciences, 2023.
24. M. R. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa- BiLSTM: A context-aware hybrid model for sentiment analysis," arXiv preprint arXiv:2406.00367, 2024.
25. A. Rane and A. Kumar, "Sentiment classification system of Twitter data for US airline service analysis," Proc. 42nd Annu. Computer Software and Applications Conf. (COMPSAC), vol. 1, pp. 114–121, 2018.
26. Y. Bao, "A comparative study of e-commerce review sentiment analysis models based on VADER and RoBERTa," J. Comput. Electron. Inf. Manag., vol. 15, no. 3, pp. 115–119, Dec. 2024. [Online]. Available: https://doi.org/10.54097/f6hyft52
27. H. Kong, "RoBERTa Vader, Naive Bayes, SVM and Logistic Regression," Proc. ACM Conf., Mar. 2025. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3715885.3715899
28. A. Borg, "Using VADER sentiment and SVM for predicting customer satisfaction," Computers & Industrial Engineering, vol. 149, p. 106804, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417420305704
29. J. P. U. S. D. Jayakody and B. T. G. S. Kumara, "Sentiment analysis on product reviews on Twitter using Machine Learning Approaches," in Proc. 2021 Int. Conf. Decision Aid Sci. Appl. (DASA), 2021, pp. 1056–1061.

[Online]. Available: https://ieeexplore.ieee.org/document/9674771

30. Pavitha, N., Dargode, A., Jaisinghani, A., Deshmukh, J., Jadhav, M., Nimbalkar, A., "Fake News Detection Using Machine Learning," in Computational Intelligence in Machine Learning, ICCIML 2022, V.
K. Gunjan et al., Eds., Lecture Notes in Electrical Engineering, vol. 1106, Springer, Singapore, 2024. doi: https://doi.org/10.1007/978-981-99-7954-740
31. Pavitha, N., Patrawala, A., Kulkarni, T., Talati, V., Dahiya, S., "NL2Code: Harnessing Transformers for Automatic Code Generation from Natural Language Descriptions," in Smart Trends in Computing and Communications, SmartCom 2024, T. Senjyu et al., Eds., Lecture Notes in Networks and Systems, vol. 947, Springer, Singapore, 2024. doi: https://doi.org/10.1007/978-981-97-1326-47
32. P. Kunekar, P. Nooji, M. Chaudhari, S. Dubey, R. Gadhave, "The Transformative Role of AI in Public Health for Cancer Prevention, Early Detection, and Management," in Artificial Intelligence in On- cology, S. N. Mohanty et al., Eds., Springer, Cham, 2025. doi: https://doi.org/10.1007/978-3-031-94302-735
33. P. Nooji, R. M. Savithramma, A. Kulkarni, S. Garge, A. Singh, Y. Darda, "Leveraging Deep Learning for Fraud Detection in Financial Transactions," in ICT: Applications and Social Interfaces, ICTCS 2024,
34. Joshi et al., Eds., Lecture Notes in Networks and Systems, vol. 1322, Springer, Singapore, 2025. doi: https://doi.org/10.1007/978-981-96-4136-914
35. P. N. Pavitha, V. Ingale, V. Verma, A. Yeole, S. Zawar, Z. Jamadar, "Comparative Analysis of Regression Algorithms for College Predic- tion," Design Engineering Journal, vol. 9, pp. 6631–6643, 2021.