Available online at
https://mijuitm.com.my

e-ISSN: 2735-0703

Mathematical Sciences and Informatics Journal 7(1) 2026, 93 – 101.

Mathematical Sciences and Informatics Journal

Comparison of Machine Learning Approaches to Sentiment Classification of Malaysian Airline Reviews

Muhammad Irham Abdul Razab[1], Haslizatul Fairuz Mohamed Hanum[1*]

[1]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor*

ARTICLE INFO

Article history:
Received 15 December 2025
Revised 3 March 2026
Accepted 18 March 2026
Online first
Published 30 April 2026

Keywords:
Sentiment Analysis
Machine Learning
Feature Extraction
Airline Reviews
Supervised Learning
Performance

DOI:
10.24191/mij.v7i1.11578

ABSTRACT

This study explores the use of machine learning (ML) methods to analyse customer reviews of Malaysia Airlines. The core problem is the need to correctly identify sentiment in unstructured online reviews, especially given language nuances, such as sarcasm, and the limited adaptability of prior models to Malaysia's local, multilingual context. The main aim is to identify the most effective among four supervised ML models: Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, and Random Forest (RF) to classify sentiment. Core aims include developing and training classifiers using TF-IDF and LDA-based feature extraction, and assessing performance using accuracy, recall, precision, and F1-score. The expectation is to find an optimal model for optimised sentiment analysis that can provide structured insights for airline operators. The study is limited to English-only text, excludes multimedia data, and uses moderately sized datasets.

1. INTRODUCTION

Customer reviews of airlines are widely disseminated online and have an impact on prospective passengers while reflecting public opinions about service quality. However, manual analysis is challenging because these reviews are frequently unstructured, diverse, and large in volume. Messy text can be transformed into useful input for machine learning models through efficient pre-processing and feature extraction techniques, such as Latent Dirichlet Allocation (LDA), to identify hidden topics. Feedback from sites such as Skytrax offers valuable information about passenger experiences in Malaysia's increasingly competitive aviation sector. This study focuses on Malaysia Airlines reviews and recognises that model performance depends significantly on the quality and structure of the text data used.

[1*] Corresponding author. E-mail address: haslizatul@uitm.edu.my
https://doi.org/10.24191/mij.v7i1.11578

The main problem is to be able to accurately identify the sentiment polarity positive, negative or neutral in reviews that may include sarcasm, blended emotions, or use of culturally specific expressions. Models of traditional ML such as Naive Bayes, SVM, and Logistic Regression can have a hard time picking up on these nuances since they focus too much on individual words rather than context. More sophisticated techniques, such as n-grams, topic modelling (LDA), and word embeddings (Word2Vec) are required to gain a better understanding of the relationships between words and to improve the accuracy of classification. Although there are many algorithms in ML that have been used in sentiment analysis, there is no clear evidence on which one is best suited for Malaysian airline reviews.

Existing studies often use general or international data sets that do not reflect local language patterns or cultural behavior. As a result, the global findings might not be reliable when applied in Malaysia. This gap indicates the need to assess various ML models with Malaysia airlines-specific data to identify the most accurate and interpretable sentiment classification algorithm to support customer satisfaction strategies of Malaysia Airlines. The main objective of this study is to determine the most effective method of sentiment analysis for Malaysia Airlines reviews by specifically evaluating the performance of different machine learning algorithms on classifying sentiment.

2. LITERATURE

Datasets play a crucial role in sentiment analysis because dataset's quality, size, and representativeness influence the accuracy of machine learning models. Previous studies on airline reviews have used various publicly available datasets such as the Twitter US Airline Sentiment dataset (Teja, 2025) and Kaggle repository (Patel et al., 2023; Lakshmanarao et al., 2022) to analyse passenger sentiment and evaluate airline service performance.

2.1 Datasets

Available literature tends to utilize universal or cross-national datasets which lack local linguistic and cultural tendencies. Consequently, the global findings may not be valid in Malaysia. The existence of this gap highlights the necessity to assess the various ML models on Malaysian airline data to understand what algorithm creates the best and confidentially most sensible sentiment classification, which aids Malaysia Airlines in its customer satisfaction policies. In this project, approximately 3,200 reviews of Malaysia Airlines, retrieved from Skytrax, a reputable aviation review site, will form the first dataset. Similar Skytrax datasets have been used in the past by Liu and Zhang (2025) to measure customer satisfaction, demonstrating that the platform provides genuine and comprehensive user-created information. It has been chosen as this type of dataset supplements more varied opinions and unlabeled text, which are effective for unsupervised feature extraction and improve the accuracy of model generalization.

The second dataset used in this study is the Airline Reviews Dataset, a publicly available dataset published on Kaggle (Bhojani, 2023) which was originally sourced from Airlinequality.com which is a reputable platform for verified passenger feedback. The dataset contains a comprehensive set of attributes, which include airline name, overall rating, review title, review date, reviewer verification status, textual review content, aircraft type, travelers' category, seat / cabin class, route and date flown. It also contains several service quality sub ratings such as seat comfort, cabin staff service, food and beverages, ground service, inflight entertainment, wifi connectivity, value for money and a binary recommendation indicator. This dataset is well suited for sentiment analysis and supervised learning studies due to the richness of structured variables and unstructured textual reviews. Prior research has stressed the importance of domain-specific review sites such as Airlinequality.com because they provide authentic and comprehensive user experience that enhances the accuracy of sentiment classification and assessment of service quality. Therefore, by helping to extract sentiment polarity, aspects related to the service and establishing the

https://doi.org/10.24191/mij.v7i1.11578

models of machine learning based on the real experiences of the passengers, this dataset facilitates the objectives of the current study.

Both datasets include textual reviews; however, they are different in the format and sentiment labelling. In the Kaggle dataset, the available sentiment labels can be used in the supervised learning protocol but the Skytrax dataset has raw reviews that can be used in text cleaning experiments which require scraped off a web site. The combination of structured and unstructured sources of reviews also strengthens the sentiment analysis with the help of both datasets. Together, these datasets provide comprehensive view of customer perceptions of Malaysia Airlines and presents the project with the ability to carry out both monitored analyses to improve the sentiment classification reliability.

2.2 Classification

Determining the sentiment polarity (neutral, negative, or positive) in reviews is challenging due to the complexity of natural language, which includes blended emotions, sarcasm, and culturally specific language prevalent in a multicultural environment like Malaysia. Traditional ML models (like Naïve Bayes, SVM, Logistic Regression) often fail to capture contextual meaning, relying heavily on individual keywords, which reduces precision. Advanced techniques such as n-gram modelling, topic modelling (LDA), and word embeddings (Word2Vec) are needed to better orient models to word relations and context for improved classification.

Despite the widespread use of various ML algorithms for sentiment classification (Patel et al., 2023; Lakshmanarao et al., 2022), there's no clear consensus on which model performs best for Malaysian airline sentiments. Existing studies often rely on generic or international datasets that may not fully account for regional language patterns, cultural context, and customer behaviour in Malaysia (Kang et al., 2021). This inconsistency necessitates an assessment of different machine learning models specifically targeted at the Malaysian context to determine the algorithm that provides the best and most readable values in predicting airline review polarity, which can then assist Malaysia Airlines in its data-driven customer happiness strategy. This study aims to find the most effective sentiment analysis approach for insights from Malaysia Airlines' seat, lounge, and service reviews.

In the present work, four machine learning algorithms were chosen to be trained on: Support Vector Machine (SVM), Logistic Regression, Naive Bayes, and Random Forest, and each of them was selected due to certain reasons. SVM is said to be robust in the high dimensional space as well as being useful in avoiding over-fitting particularly when the size of the features is greater than the size of the samples. The reason behind the selection of Logistic Regression is that it is easy to interpret; it provides clear understanding of the relationship between the predictors and the response variable and therefore the coefficient is easy to interpret. Naive Bayes is appreciated by its ease and speed, it works correctly with huge amounts of data, and it represents a good baseline that can be used in comparison. Finally, Random Forest is known to be highly accurate, and it could process many features without the necessity of large-scale feature selection, as well as allow one to see the importance of the features. These selections of algorithms are in line with the objectives of the study and use the advantages of both approaches to benefit the research problem.

https://doi.org/10.24191/mij.v7i1.11578

3. METHODOLOGY

This study employs experimental research design with both qualitative and quantitative results. The quantitative results will include the accuracy, precision, recall, and F1-score of the methods. The experimental process is structured using the CRISP-DM (Cross-Industry Standard Process for Data Mining) method.

3.1 Datasets

The dataset consists of Malaysia Airlines passenger reviews taken from 2 different sources the Skytrax website for the year 2025 and Kaggle. For first dataset (known later as Dataset A) was collected using Python's *BeautifulSoup* module to extract text and rating information from websites. The second dataset (labelled as Dataset-B) was extracted from the Kaggle website. The dataset comprises Malaysia Airlines passenger reviews collected from the Skytrax website over the period of 2025. Reviews are collected across three service categories (service, lounge, and seat) with an equal number of reviews from each section selected at random. Web scraping is executed using the Python *BeautifulSoup* module to parse HTML content and extract text and rating information

To enable supervised learning, numerical rating scores (1-10) are converted into categorical sentiment labels (positive, neutral, or negative) using defined thresholds. Sentiment polarity can be inferred from rating values by defining thresholds, where lower ratings represent negative sentiment, mid-range ratings represent neutral sentiment, and higher ratings represent positive sentiment (Ashbaugh et al., 2024). For example, ratings between 1 and 3 are typically considered negative, 4 to 6 as neutral, and 7 to 10 as positive as shown in Table 1.

Table 1. Categorization of the sentiment rating

Rating range	Suggested Polarity
1 through 3	Negative
4 to 6	Neutral
7 to 10	Positive

In order to guarantee a dependable and objective model performance, the distribution of sentiment classes within a dataset is crucial. Machine learning models may become biased toward forecasting the dominant class when the percentages of negative, neutral, and positive assessments are significantly out of balance. This may result in poor generalization, deceptive precision, and subpar performance in minority courses. The polarity distribution for both datasets is displayed in Table 2. There are approximately equal numbers of negative and positive evaluations in Dataset A, with fewer neutral remarks. Dataset B, on the other hand, has a significant bias in favor of the Bad class, which comprises most of the samples. The learning process may be impacted by this imbalance, leading the model to concentrate more on negative sentiment patterns.

Table 2. Polarity distribution for Dataset A and Dataset B

Dataset	Negative Polarity	Neutral Polarity	Positive Polarity
A	491	210	482
B	8016	962	1950

Before training the models, the text reviews undergo several pre-processing steps to improve data quality. First, sentiment segmentation splits each review into individual sentences so mixed opinions within long reviews can be analyzed more accurately. Next, tokenization using regular expressions breaks sentences into word tokens, which is a key step in NLP and works well for informal online language. Then, punctuation and noise removal is performed to clean elements like usernames, URLs, numbers, and

https://doi.org/10.24191/mij.v7i1.11578

unnecessary symbols, while keeping hashtag words for their potential meaning. After that, the text is converted to lowercase and stop words are removed using a predefined list. Finally, lemmatization reduces words to their base form to standardize vocabulary, reduce dimensionality, and improve the model's overall performance

3.2 Features

Latent Dirichlet Allocation (LDA) and Term Frequency–Inverse Document Frequency (TF-IDF) are the best feature extraction techniques used in this study. By highlighting discriminative phrases and minimizing the impact of frequently occurring but less informative words, TF-IDF transforms textual reviews into weighted numerical vectors. By modeling each review as a combination of subjects, LDA, on the other hand, finds latent thematic patterns, such as flight delays, customer service, or comfortable sitting. Instead of treating the review as a single, homogeneous text unit, this topic-based approach aids in capturing more detailed sentiment cues related to service features.

To enhance the classification performance, instead of only using feature extraction, this study employs feature deletion by using Information Gain (IG) filter approach. IG evaluates the contribution of each feature in reducing uncertainty (entropy) of sentiment labelling. The model can have a focus to work on the most informative characteristics by discarding features that contribute very little. Information Gain is an efficient feature selection technique of filter-based approach, which reduces the computational complexity and improves the model accuracy, particularly for tree-based classifiers (Bohani et al., 2024). Bohani's dataset results on airline customer satisfaction revealed that the J48 classifier outperforms Naive Bayes based on the data precision and F1-scores (0.33% comparing with Naive Bayes) when 10-fold CV strategy was used and 0.29% when 20-fold CV strategy was used after eliminating low importance features through IG.

By integrating TF-IDF or LDA with IG-based feature elimination, this study employs a hybrid approach that strengthens both the representation quality and feature relevance. This leads to more efficient training, reduced noise, and improved robustness in sentiment classification.

3.3 Classification

This study applies four supervised machine learning algorithms Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, and Random Forest on samples for training using their default parameter settings to classify reviews into sentiment categories. Default parameters are applied as a standard practice for establishing baseline results before tuning, and several recent studies confirm the effectiveness of these algorithms in text classification tasks with such setups (Venkatesh et al., 2024). SVM classification is evaluated using a confusion matrix and 5-fold cross-validation (Ipmawati et al., 2024). Then, Naïve Bayes is evaluated via standard tests and evaluation of the model (Pristiyono et al., 2021) while Random Forest uses k-fold cross validation (Khomsah, 2021).

Each model is then critically evaluated using the 80 to 20 ratio of training verses testing samples in order to ascertain generalizability and reliability using four primary metrics in the confusion matrix which include: Accuracy, Precision, Recall and F1-score (Lakshmanarao et al., 2022). Accuracy is the ratio of predicting sentiment labels with correct positive and correct negatives of all reviews examined. Although it gives a rough picture of the performance, accuracy is erroneous in situations where a single sentiment prevails in the dataset. As an example, when most of the reviews about an airline are negative, a model that always predicts the negative may have high accuracy without successfully discriminating positive reviews. Therefore, other measures will be required to counter this shortcoming.

Precision is used to measure the percentage of reviews that are supposed to be positive but instead, they are negative. High precision in airline sentiment analysis is important in that when an airplane system identifies a review as positive, there is high probability that the customer was actually satisfied. This is vital

https://doi.org/10.24191/mij.v7i1.11578

to process such applications as an airline recommendation system or a brand monitoring tool, where false positives may give the wrong impression of customer satisfaction.

Recall, or sensitivity, measures the classifier's ability to pick up positive reviews in all cases. A high recall score indicates that the model is good at identifying satisfied customers - even when customers are a minority. In the airline industry, it is important to capture the positive sentiment accurately, which is important for understanding which service or routes contribute to customer satisfaction.

The F1-score is a combination of the precision and recall functions into one metric by calculating the harmonic mean of the two. It is particularly helpful to use when there is an imbalance of positive or negative reviews such that the model's ability to correctly identify sentiment can be assessed in a balanced way. In the case of sentiment analysis for Airlines, F1-score is used to see whether the model can accurately classify both satisfied and dissatisfied customers unbiasedly.

4. RESULTS AND DISCUSSION

The performance of the four supervised machine learning models Logistic Regression, Random Forest, Naïve Bayes, and SVM was evaluated using two different feature extraction techniques: Latent Dirichlet Allocation (LDA) and Term Frequency-Inverse Document Frequency (TF-IDF). The performance metrics assessed were Accuracy, F1-score, Recall, and Precision.

4.1 Classification on Dataset A (Skytrax dataset)

Each feature extraction technique, LDA and TF-IDF was used together with the machine learning approach, and the classification results are recorded as in Table 3.

Table 3. Results from classification of Dataset A

Machine Learning Approach	Accuracy	F1	Recall	Precision
Logistic regression & LDA	0.551	0.546	0.551	0.544
Logistic Regression & TF-IDF	0.572	0.567	0.572	0.565
Random forest & LDA	0.508	0.488	0.508	0.4824
Random Forest & TF-IDF	0.496	0.464	0.496	0.473
Naïve Bayes & LDA	0.415	0.428	0.415	0.458
Naïve Bayes & TF-IDF	0.508	0.513	0.508	0.5476

When analysing the first dataset A, the results largely clustered around the 50% accuracy mark, indicating that traditional machine learning models and frequency-based features face significant challenges in handling the complexity and nuances of raw data. Specifically, the Random Forest classifier combined with TF-IDF achieved the highest overall Accuracy (0.521) and Recall (0.521) for this dataset. In contrast, the Support Vector Machine (SVM) model utilizing LDA features provided the most balanced performance for that specific feature set, securing the highest F1-score (0.509) and Precision (0.521)

4.2 Classification on Dataset B (Kaggle dataset)

The classification results after using LDA and TF-IDF for Dataset B as the feature extraction technique are presented in Table 4. Performance improved significantly when the models were applied to the pre-labelled B dataset with the SVM model using TF-IDF features achieving the highest Accuracy of 0.760. For precision measures, all approach produced results above the 0.75 indicating an acceptable high precision, however, Naïve Bayes approach produced the lowest result for Recall and F1 measures.

https://doi.org/10.24191/mij.v7i1.11578

The comparative results of these two data sets were consistent with the known literature; that is, that ensemble methods such as Random Forest were robust when handling noisy text, while SVM was strong in high dimensional spaces such as those produced by LDA topic distributions.

Table 4. Results from classification of Dataset B

Approaches	Accuracy	F1	Recall	Precision
Logistic Regression & LDA	0.715	0.814	0.847	0.783
Logistic Regression & TF-IDF	0.722	0.815	0.836	0.796
Random Forest & LDA	0.735	0.839	0.942	0.757
Random Forest & TF-IDF	0.746	0.842	0.921	0.775
Naïve Bayes & LDA	0.483	0.534	0.404	0.7884
Naïve Bayes & TF-IDF	0.627	0.713	0.632	0.818
SVM & LDA	**0.755**	0.848	0.928	0.78
SVM & TF-IDF	**0.76**	0.847	0.909	0.7946

In general, the fact that TF-IDF has outperformed most classifiers indicates that lexical-level features prove to be more efficient than topic-level features in sentiment classification, especially when the polarity of sentiment is based on the use of certain opinion words and not on the more general subject matter organization. Despite these successes with the structured dataset, the limits encountered with the Skytrax data suggest the need for a large methodological step. To obtain much better classification results and generalization, the future study should explore deep learning models, such as Convolutional Neural Networks (CNNs) or Long Short-Term Memory network (LSTMs).

5. CONCLUSION

This study set out to determine the most effective machine learning approach for classifying sentiment in Malaysia Airlines reviews by comparing Logistic Regression, Random Forest (RF), Naïve Bayes, and Support Vector Machine (SVM) across two feature extraction techniques: LDA and TF-IDF. The findings revealed distinct variations in performance depending on both the model architecture and the nature of the dataset used. For the more challenging and unstructured Skytrax dataset (Dataset A), the Random Forest classifier paired with TF-IDF demonstrated a slight superiority in predicting sentiment polarity, achieving the highest Accuracy and Recall of 0.521. Meanwhile, the SVM model using LDA features offered the most balanced outcome, recording the highest F1-score (0.509) and Precision (0.521). In comparison, the application of these models to the pre-labelled Kaggle dataset (Dataset B) yielded significantly better results, with SVM using TF-IDF reaching an accuracy of 0.760.

By assessing and comparing different machine learning models for analysing airline reviews, this study can provide a structured and data-driven approach to business improvement for stakeholders. Rather than manually analysing large amounts of customer feedback, decision-makers can take advantage of the most accurate model to quickly identify the most important factors contributing to passenger dissatisfaction, such as flight delays, seating comfort or service quality. These insights help marketing teams proactively retain at-risk customers through targeted promotions, support operational teams to address recurrent problems at particular airports or routes and help executives prioritize investments that generate the greatest return. Overall, this research transforms unstructured textual data into actionable insights that can be effectively used as a strategic guide for improving competitiveness and strengthening customer loyalty.

These results support the current scholarly agreement that SVM is more appropriate for high-dimensional spaces whereas ensemble techniques like Random Forest are more effective at handling unstructured, noisy text. However, the fact that the Skytrax dataset's performance measures stayed close to 50% suggests that conventional machine learning models are more useful as a baseline than as a final solution for this kind of data. The intricacy of the reviews which are marked by subtle wording, irony, and

https://doi.org/10.24191/mij.v7i1.11578

class disparity indicates that conventional methods may not be able to adequately capture sentiment context. Therefore, to improve generalizability, future research must go beyond frequency-based feature extraction. It is strongly recommended to explore deep learning architectures, such as Convolutional Neural Networks (CNNs) or Long Short-Term Memory networks (LSTMs) and to incorporate multilingual datasets to better address local language patterns and the intricate nature of airline passenger feedback.

6. ACKNOWLEDGEMENTS/FUNDING

The authors would like to acknowledge the support of Universiti Teknologi MARA (UiTM), Shah Alam, Selangor for providing the facilities at Faculty of Computer and Mathematical Sciences, UiTM and financial support on this research.

7. CONFLICT OF INTEREST STATEMENT

The authors agree that this research was conducted in the absence of any self-benefits, commercial or financial conflicts and declare the absence of conflicting interests with the funders.

8. AUTHORS' CONTRIBUTIONS

Muhammad Irham carried out the research, wrote and revised the article. Haslizatul Fairuz anchored the review, revisions, and approved the article submission.

REFERENCES

Ashbaugh, L., et al. (2024). A comparative study of sentiment analysis on customer reviews using machine learning and deep learning techniques. *Computers, 13*(12), Article 340. https://doi.org/10.3390/computers13120340

Bhojani, J. (2023). Airline Reviews [Data set]. Kaggle. https://www.kaggle.com/datasets/juhibhojani/airline-reviews

Bohani, F. A., Mohamed Rashid, F. S., Mahmud, Y., & Yahya, S. R. (2024). Analyzing the impact of feature selection using Information Gain for airlines' customer satisfaction. Malaysian Journal of Computing, 9(1), 1673–1689. https://doi.org/10.24191/mjoc.v9i1.24163

Ipmawati, J., Saifulloh, S., & Kusnawi, K. (2024). Analisis Sentimen Tempat Wisata Berdasarkan Ulasan pada Google Maps Menggunakan Algoritma Support Vector Machine: Sentiment Analysis of Tourist Attractions Based on Reviews on Google Maps Using the Support Vector Machine Algorithm. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4(1), 247-256. https://doi.org/10.57152/malcom.v4i1.1066

Kang, H. W., et al. (2021). Malaysian airlines sentiment analysis using BERT approach. In *Proceedings of the International Conference on Digital Transformation and Applications (ICDXA 2021)*, 135–142.

Khomsah, S. (2021). Sentiment analysis on youtube comments using word2vec and random forest. Telematika: Jurnal Informatika dan Teknologi Informasi, 18(1), 61-72. https://doi.org/10.31315/telematika.v18i1.4493

https://doi.org/10.24191/mij.v7i1.11578

Lakshmanarao, A., Gupta, C., & Kiran, T. S. (2022). "Airline Twitter Sentiment Classification using Deep Learning Fusion," 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2022, pp. 1-4. https://doi.org/10.1109/smartgencon56628.2022.10084207

Liu, Y., & Zhang, Y. (2025). Study of Airline Passenger Preference Formation Mechanisms-An Application of a Hybrid Analysis Method Based on Skytrax Ratings and Reviews. In SHS Web of Conferences (Vol. 225, p. 03028). EDP Sciences. https://doi.org/10.1051/shsconf/202522503028

Patel, A., Oza, P., & Agrawal, S. (2023). Sentiment analysis of customer feedback and reviews for airline services using language representation model. *Procedia Computer Science, 218*, 2459–2467 https://doi.org/10.1016/j.procs.2023.01.221

Pristiyono, Ritonga, M., Ihsan, M. A. A., Anjar, A., & Rambe, F. H. (2021). Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm. In IOP Conference Series: Materials Science and Engineering (Vol. 1088, No. 1, p. 012045). IOP Publishing. https://doi.org/10.1088/1757-899x/1088/1/012045

Teja, P. (2025). A Comprehensive Study of Deep Learning and Traditional Machine Learning Models for Twitter Sentiment Analysis. International Journal for Research in Applied Science and Engineering Technology. 13. 2010-2015. https://doi.org/10.22214/ijraset.2025.73303.

Venkatesh, J. D., Jaiswal, A., & Nanda, G. (2024). Comparing human text classification performance and explainability with large language and machine learning models using eye-tracking. Scientific reports, 14(1), 14295. https://doi.org/10.1038/s41598-024-65080-7

https://doi.org/10.24191/mij.v7i1.11578