World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

Sentiment analysis of passenger feedback on U.S. airlines using machine learning classification methods

Md Nurul Raihen [1,*] and Sultana Akter [2]

[1] *Department of Mathematics and Computer Science, Fontbonne University, Saint Louis, MO, USA.*
[2] *Institute for Data Science and Informatics, University of Missouri Columbia, Columbia, MO, USA.*

World Journal of Advanced Research and Reviews, 2024, 23(01), 2260–2273

Publication history: Received on 12 June 2024; revised on 18 July 2024; accepted on 20 July 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.1.2183

Abstract

Twitter, a platform for micro-blogging, has contained as a novel information architecture. Everyday People worldwide publish about 200 million status messages, known as tweets. Twitter users express their opinions by posting concise text messages. Twitter data is useful for sentiment analysis and consumer feedback tweets. This study employed multi-class sentiment analysis to analyze tweets from 6 major US airlines (American, United, US Airways, Southwest, Delta and Virgin America). Airlines are essential for travel, and this study has helped people choose the best ones. Classification model with the lowest error rate could help airline companies improve their business by figuring out why information is being misclassified. This analysis of airline evaluations can help us identify good airlines and apply this model to our own journeys. This helps the airline identify its weaknesses so they can improve them. A technique of natural language processing (NLP) known as sentiment analysis (or opinion mining) classifies the tone of data as positive, negative, or neutral. The analysis was conducted with seven distinct classification strategies: Linear Discriminant Analysis, Quadratic Discriminant Analysis, Decision Tree, Random Forest, K-Nearest Neighbors, Gradient Boosting, and AdaBoost to utilize the split validation (80% as train data set, 20% as test data set) and 10 folds cross validation process. The suggested model demonstrates superior accuracy and efficiency compared to all others, achieving an accuracy score of 90.13% for the Random Forest with 10 folds cross validation approach. The project aims to utilize machine learning techniques to estimate the reasons for misclassified information since the lowest error rate means the airline sentiment provides less wrong prediction.

Keywords: Twitter; Airlines; Classification; Error Rate; Validation

1. Introduction

This guide Classification algorithms can help airline companies to segment customers into different groups for targeted marketing campaigns. Classification methods are a subset of machine learning techniques (NLP) that are used to categorize data into predefined classes. There is a growing demand to extract tweets pertaining to a certain topic of interest from the vast number of tweets available. The airline business is a highly competitive field that has experienced significant growth during the previous two decades. Airline companies rely on conventional customer feedback forms, which can be extremely laborious and time-consuming. The success of a service business depends heavily on happy customers. Because of this, understanding the customer's mental and emotional state is crucial. Often, textual data is subjected to sentiment analysis so that organizations can comprehend customer requirements and track the sentiment of customers regarding their products and brands.

The objective of the test set is to determine the sentiment of the tweets (positive/negative/neutral) based solely on the Text variable (customer feedback). Twitter is a gold mine of data with over 1/60th of the world's population using it

* Corresponding author: Md Nurul Raihen

which nearly amounts to 100 million people, more than half a billion tweets are tweeted daily and the number keeps growing with every passing day. With the rising demand and advancements of Big Data technologies in the past decade, it has become easier to collect tweets and apply data analysis techniques on them [1]. Twitter is a much more reliable source of data as the users tweet their genuine feelings and feedback thus making it more suitable for investigation [2]. The company can conduct sentiment analysis on product-related tweets as part of their market research to enhance their product.

After gathering the airline tweets, they are subjected to pre-processing in order to eliminate extraneous information. Subsequently, sentiment categorization methods are employed on the processed tweets. This provides data scientists and airline firms with a more comprehensive understanding of the sentiments and viewpoints expressed by their customers. The primary objective of this article is to offer the airline industry a more extensive understanding of the feelings expressed by their consumers and to cater to their demands in the most effective manner possible. This article examines various tweet pre-processing approaches and applies seven distinct machine learning classification algorithms to ascertain the sentiment expressed in the tweets. Next, the classifiers are compared to determine their respective accuracies.

In recent times, big data has been applied to datasets, and the essence of big data comprises a very huge amount of information. Therefore, with the growth of technology and the rising multitudes of data pouring out of a particular organization, it became necessary to develop new ways to manage this data in a faster and more effective manner. The traditional tools and software are unable to handle this enormous dataset: as a result, it became necessary to design new ways to manage this data. In order to uncover previously undiscovered facts, businesses and organizations consistently make efforts to investigate and gather data with a high level of structure. Big data can be characterized by three characteristics: volume, diversity, and velocity [3]. Every one of these characteristics is important. Volatility refers to the fluctuating rate of data or the frequency with which it is produced, while volume refers to the amount of the data and how enormous it is. Variety refers to the numerous formats, types, and different ways of using and evaluating the data. The size, the number of records, the transactions, the tables, or the files are the characteristics that distinguish big data [4]. It is more difficult to acquire, process, and analyze big data using typical data management abilities since big data is a sort of data that does not have any structure [5]. Big data is a substantial amount of information that originates from a variety of sources, such as logs, clickstreams, and social media [6].

First created in 2006, Twitter is a social media platform that allows users to interact with one another and communicate online through the use of computers [7]. Users of Twitter are able to share brief posts, which are referred to as "tweets," that are 140 characters in length. Twitter usage went from 20,000 tweets per day to 60,000 tweets per day after a year, and users began using it for daily updates, information about brands, to enhance their connection with people all over the world, and to share their thoughts on a certain product. Although Twitter users are now able to use up to 280 characters in a single tweet, the platform is still considered to be short. This is due to the fact that it was initially designed as a mobile phone service or a short message service, similar to the traditional SMS. For the same reason that businesses are becoming more and more popular, many brands and corporations are utilizing this platform in order to learn about the feedback of customers and to provide services to clients. Research indicates that 79% of Twitter users follow brands, and some businesses manage their Twitter accounts in order to provide help to their customers [8].

Unstructured data is text heavy or configured in way that it becomes difficult to analyze like twitter data and it contains opinions, features of recommendations etc. Structure data is predefined structure format, often time numerical, like excel, google sheets where people can add columns, rows with the predefined parameters. The semi structure is also text heavy but organized into categories and “meta tags” [9].

For unstructured data it is challenging to analyze, but it has its own internal structure, that includes text, audio, images. Particularly business interactions are unstructured. Over 80% of generated data is unstructured. These data can be analyzed by using natural language processing, which focuses on text and speech, social media data like twitter, computer vision, self-driving car. One of the most well-known natural language processing fields is sentiment analysis, which deals with mining of information which is related to sentiments and opinion of a group for a specific issue [10].

2. Methodology

The following section consists of a description of the dataset used for sentiment analysis, including its visual representation and the methodology suggested for conducting sentiment analysis on the chosen dataset.

2.1. Data Set

The initial source of this data was the Data for http://www.crowdflower.com/data-for-everyone. A modified version of the original source is what is available on Kaggle (Data source: Twitter US Airline Sentiment (kaggle.com). The scraping of Twitter data started in February 2015. Participants were provided to initially identify tweets as positive, negative, or neutral. This dataset is made up of reviews that people who have taken flights on different airlines have written. In this study, the dataset collected tweets for six airlines of the United States (US). The name of the data set is "twitter-airline-sentiment", which consists of 12 columns and a total of 14,640 observations, with one column indicating the sentiment (negative, neutral, and positive) of the tweet named "airline_sentiment" (target variable) and another column expressing the emotions of the tweet called "text". These two columns are our primary focus among the 12 columns. All other columns contain diverse information pertaining to the content of the tweet, its geographical origin, the time of posting, and the number of retweets, among other details. Table1 is a brief description of the variables of twitter-airline-sentiment data.

Table 1 Variable description of the twitter-airline-sentiment data

Name of Variables	**Description**
tweet_id	Id of the tweet
airline_sentiment (target)	Sentiment of the tweet (negative=-1, neutral=0, positive=1)
airline_sentiment_confidence	Confidence with which the given sentiment was determined.
Negative reason_confidence	Confidence with which the negative reason of tweet was predicted
name	Name of the person who tweeted
retweet_count	Number of retweets
Text (target features)	Text of the tweet whose sentiment has to be predicted.
tweet_created	Time at which the tweet was created
tweet_location	Location from where the tweet was posted.
user_time zone	Time zone from where the tweet was posted.
Negative reason	Reason for which user posted a negative tweet
airline	Airline for which the tweet was posted.

Table 2 Sample of customer review

Tweets	**keywords**	**Sentiment types**
united thnx for the info	thnx	positive
united after a Cancelled Flighted flight, and 2 delays, you lost my luggage	Cancelled, lost luggage	negative
united It was 3387. But we helped you with the weight issue, took your vouchers, and hopped on a @Delta flight instead. #winwin	Helped, instead flight	neutral

Table 2 displays a selection of terms that indicate various types of moods. When the keyword "thanks" is detected in the feedback, it is classified as a positive review. Conversely, if the keyword "cancelled" is present, it is categorized as a negative review. The term "instead flight" is used to represent neutral feedback.

2.2. Data Exploration

The code was completely implemented utilizing Spyder software, a powerful Python development environment that features sophisticated editing, testing, and numerical computing capabilities.

A word cloud, which is also called a tag cloud or wordle, is a picture of text data. It is often used to show keyword information on websites or to showcase free-form text [11]. This is a common way to show the frequency or how important words are in a text. We had used word cloud to visualize our text data.

Figure 1 Word cloud of Airline tweet data

Figure 1 is the visualization of Word Cloud of Airline data set. The tweets undergo various stages of preprocessing to get the cleaned tweets which can be used for further analysis. Data visualization gives us more information about our data set [12].

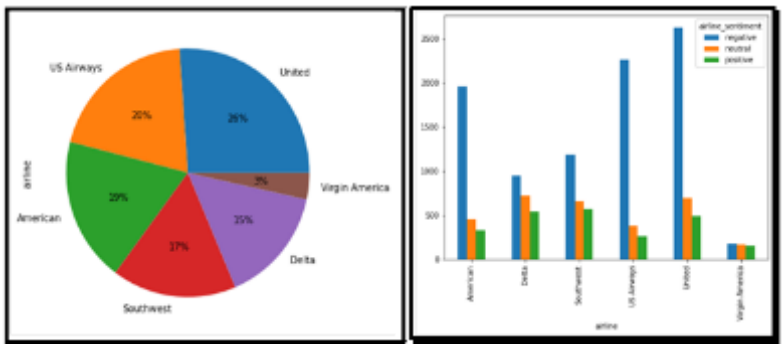


Figure 2 Percentage of 6 USA Airlines, and Frequency of airline_sentiment

Figure 2 shows us the percentage of six airline companies, as well as the frequency of each level of airlines_sentiment (negative, neutral, and positive) in all 6 airlines separately.

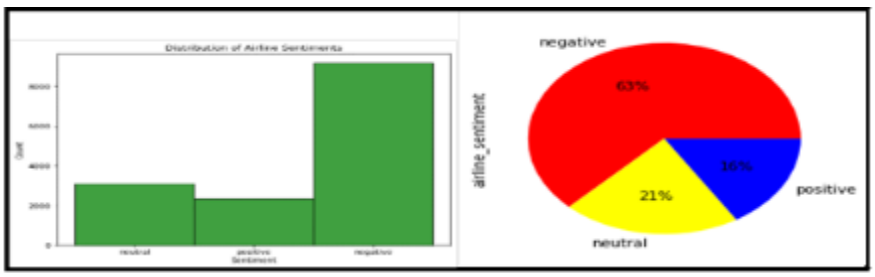


Figure 3 Plots for frequency of Sentiment levels of sentiment_airline_tweet data

Figure 3 shows us the percentage of each sentiment group. Figure 2 shows that, in our tweet airline data, American Airlines 19%, United Airlines 20%, United Airlines 26%, Virgin America 3%, Delta 15%, Southwest 17%. This pie chart tells us which level has the highest frequency of customer reviews. Negative review has the highest percentage value here. It is an indication of the most customers were not so much pleased with airline companies service of those 6 airline groups in our data we considered (United, US Airways, Southwest, Delta and Virgin America, American).

Table 3 Sentiment distribution of Tweets

Sentiment	Tweet Count
negative	9178
neutral	3099
positive	2363

Table 3 is the frequency of each level of sentiment feedback from airline customers. It will give more detailed information about the opinion of customers. The following Table 3 gives the tweets sentiment distribution.

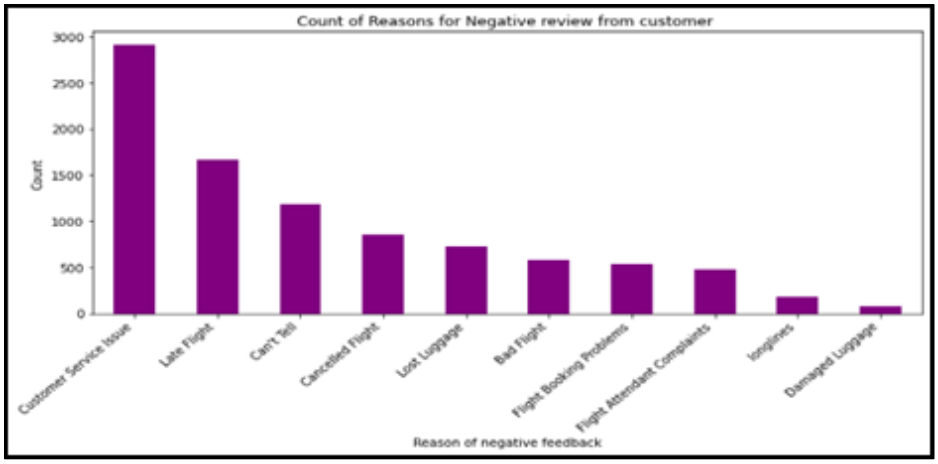


Figure 4 Reasons for negative feedback of customer

Table 3, and Figure 4 show that the number of customer's feedback (positive, negative, and neutral) about their experience in those six airlines in the United States. Both plots tell us which level has the highest frequency of customer reviews. Negative review has the highest percentage value here. It is an indication of the most customers were not so much pleased with airline companies service of those 6 airline groups in our data we considered (United, US Airways, Southwest, Delta and Virgin America, American).

- Positive sentiment: All reviews that show a positive are from travelers who had a good experience with airlines [13].
- Negative sentiment: All reviews that show a bad response describe problems travelers had [13].
- Neutral sentiment: All reactions include those that aren't clear about whether they should be satisfied or not [13].

On observing the graph, the majority of the tweets expressed negative sentiment, this may be because people generally use the social media platform to convey their dissatisfactory remarks. It is very clear here in this airline tweet data the customers had so many bad experiences that's why negative reviews have the highest frequency.

Moreover, Figure 4 will be very helpful for the airline companies to minimize their issues about customer service, flight schedule, luggage issues, flight environment, booking problem those will reduce the negative feedback from the customers.

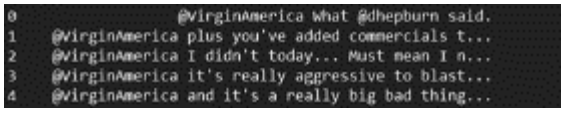
2.3. Analysis Plans

Tools: Spyder, which is a powerful development environment for Python language.

2.3.1. Data Pre-Processing

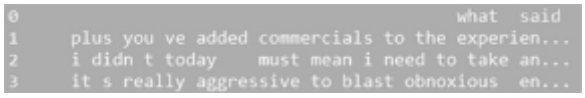
- For pre-processing our data, we checked whether there is any missing observation or not in this airline tweet data by using Python utilizing Spyder.

There is no missing observation in our target variable “airline_sentiment”, and our main features “text”.



It shows that the 1st 5 observations from text variables from the raw data. Data pre-processing steps to clean text data which need to be dealt with before performing any kind of analysis.

- Remove words which start with @ symbols.



- remove special characters except [a-zA-Z].
- remove link starts with https.
- remove stopwords from the text data (the stopwords are common words in a language (like "the", "is", "in", etc.))
- convert to lower case letter.
- convert text features into TF-IDF features vector.
- TF-IDF

it is the bag of words that has the same weight. The idea behind the TF-IDF approach is that the words that occur less in all the documents and more in individual documents contribute more towards classification.

TF-IDF is combination of two terms:

Term Frequency (TF) and Inverse Document frequency (IDF)

TF = (frequency of a word in the document)/ (total words in the documents)

IDF=log (total number of docs)/ (number of docs containing the word)

TF-IDF = TF ∗ IDF

Note: In python tf-idf values can be computed using TfidfVectorizer() method in sklearn module.

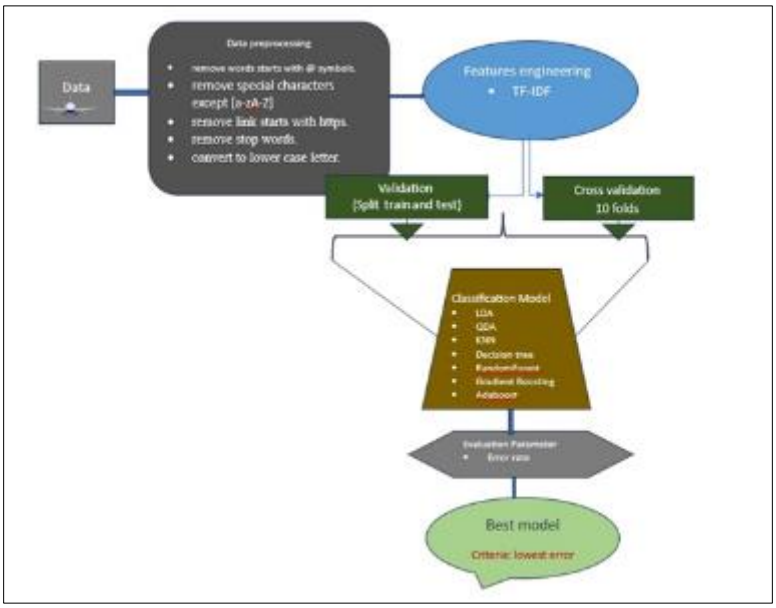


Figure 5 Analysis plan of sentiment analysis

2.3.2. Validation (Split Train/Test)

Statistical model validation involves assessing the competency of a selected statistical model. Model validation is the post-training process of evaluating the performance of a trained model using a separate testing dataset [14]. The testing data may or may not be a subset of the same dataset used for the training set.

- A training data set is a collection of instances that is used throughout the developing process to adjust the parameters.
- A test data set is a collection of data that is separate from the training data set yet adheres to the same probability distribution as the training data set.
- For this analysis, Train data contains 80%observations, and test data contains 20% observations.

2.3.3. Cross Validation

It is a sampling technique that includes excluding certain portions of the data during the fitting process. This allows us to assess how well the model predicts the excluded data points, determining whether they are in close proximity or significantly deviate from the predicted values. There are several types of cross validation approaches [15]. But here we used k-Fold cross validation process.

K-folds cross validation: K-fold cross-validation evaluates predictive models. The dataset is folded into k subsets. The model is trained and tested k times with a different validation fold. Each fold's performance indicators are averaged to estimate model generalization.

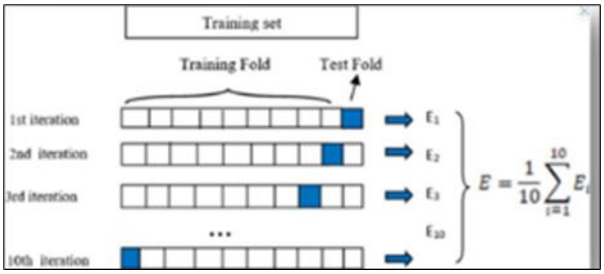


Figure 6 10-folds cross validation

To evaluate all classification models utilized 10 folds cross validation. Figure 6 is a representation of a 10-folds cross validation process.

2.4. Classification Model

In this subsection we will discuss broadly each classification model that we used in our analysis.

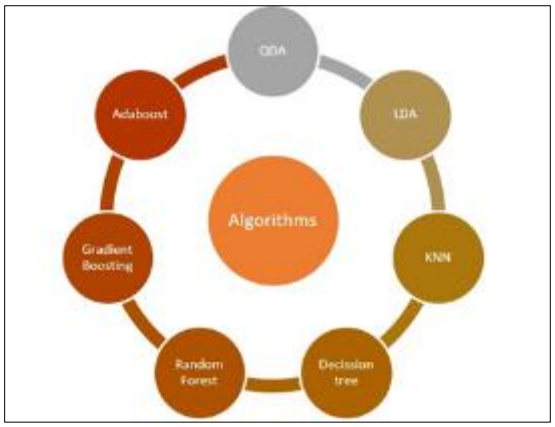


Figure 7 Model for Classification

2.4.1. Linear Discriminant Analysis (LDA)

It was used as a machine learning method for dataset analysis in this study that looked at the potential health hazards to mothers. LDA is a classification method that is widely used because it is good at reducing the number of dimensions while keeping the ability to tell classes apart [16]. This method works by finding a linear combination of features that best separates two or more classes of events. Here LDA performed with 3 different labels of target variable (RiskLevel). In this study, LDA was applied to the training dataset, with its parameters fine-tuned through a validation and cross validation process. LDA is useful because of its simplicity, efficiency in multi-class issues, and ability to reveal classification features.

2.4.2. Quadratic Linear Analysis (QLA)

It is used to assess maternal health risks. QDA is an LDA variant that separates classes non-linearly. QDA believes each class has its own covariance matrix, unlike LDA. This makes QDA more versatile for datasets without a shared covariance matrix. This study used QDA and cross-validation to optimize parameters. QDA was used in the study because it can simulate more complex feature-class connections and handle datasets with varying covariance patterns. A bigger sample size is needed to estimate covariance matrices accurately.

2.4.3. k-Nearest Neighborhood (kNN)

This algorithm was used to analyze the dataset in comprehensive maternal health risk research. A simple but powerful classification method used in data mining and machine learning is KNN. Here I utilized k=10 nearest neighbors. This idea is that similar data points are likely to be close in feature space. This method classifies a fresh sample by the majority class of its 'k' nearest neighbors in the training dataset. This work optimized the critical parameter 'k' using cross-validation. However, the choice of 'k' and the distance measure can affect its performance, and computing distances to all training samples can be computationally costly for large datasets [17]. Here I performed validation and cross validation approaches to fit the KNN classification model.

2.4.4. Decision Tree (DT)

A non-parametric supervised learning technique for regression and classification is called a decision tree (DT). The objective is to build a model that, by utilizing basic decision rules deduced from the data features, predicts the value of a target variable. They look like trees, which is how they got their name. In the case of classification, they begin at the tree's root and proceed through binary splits based on possible outcomes until they reach a leaf node and give the final binary result. Parameter used max_depth=2, random_state=1 to fit the decision tree classifier model in this study.

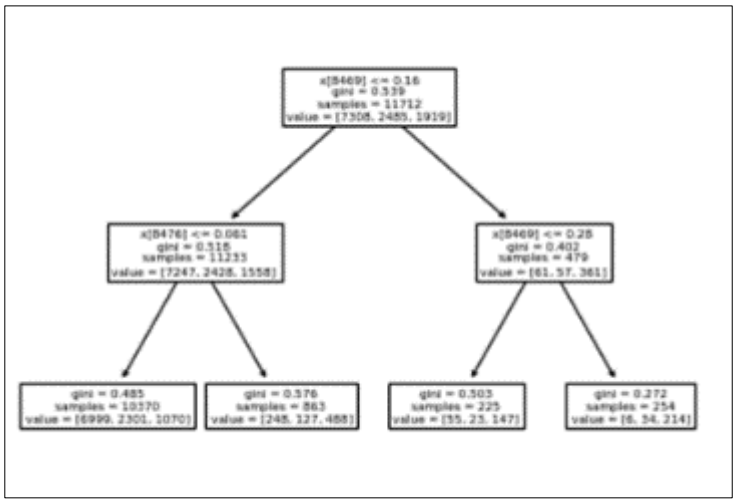


Figure 8 Decision Tree of sentiment_airline_tweet

2.4.5. Random Forest (RF)

The random forest algorithm is a modified version of the bagging method that combines bagging and feature randomness to generate a collection of decision trees that are not connected with each other. Random forests only select a subset of those features. When evaluating a split in a tree a new set of predictors is chosen, and it is common practice to select a number of predictors that is equal to the square root of the total number of predictors.

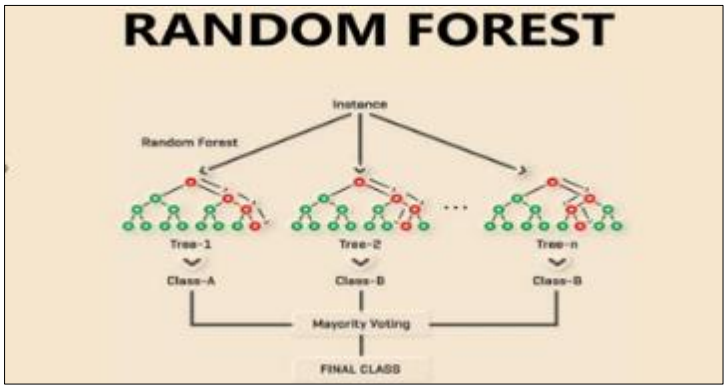


Figure 9 Sample Figure of Random Forest Model

2.4.6. Gradient Boosting Classifier (GBC)

Gradient boosting is a machine learning method employed in many applications such as regression and classification. The system provides a predictive model that consists of a collection of weak predictive models combined together. A gradient-boosted trees model is constructed iteratively, similar to previous boosting methods. However, it distinguishes itself by enabling the optimization of any differentiable loss function. It is a boosting technique that builds a final model from the sum of several weak learning algorithms that were trained on the same dataset. It operates on the idea of stagewise addition. In python, I utilized all of those parameters (max_depth = 3, n_estimators = 10, learning_rate = 1, random_state = 1) for Gradient Boosting Classifier algorithm in this study.

2.4.7. Adaboost Model (AM)

Adaboost is one of the earliest implementations of the boosting algorithm. It forms the base of other boosting algorithms, like gradient boosting and XGBoost. In Boosting we combine predictions that belong to different types. AdaBoost is a boosting technique that employs the stagewise addition approach to combine numerous weak learners and create strong learners. Here, we applied those parameters (n_estimators = 10, random_state = 1) in Spyder python to fit this Adaboost Model.

2.4.8. Criteria of Chosen the Best Model

There are many ways which we can use to evaluate the fitted model:

- Accuracy Score/ Error rate
- Null Accuracy
- Precision
- Recall
- f1 score
- ROC — AUC

Here, we used error rate for each classification model, and figured out the lowest error rate (which means of mis-correct predictions). Confusion matrix gave me an accuracy score (correct predictions) to calculate the diagonal elements of this matrix, and error rate is the inverse of accuracy score [18, 19]. Confusion matrices are tables used to describe the performance of a classification model on a set of test data that is already well-known. There are 4 variables that make up the confusion matrix. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are the four possible outcomes. The format of the 3*3 confusion matrix is displayed in below:

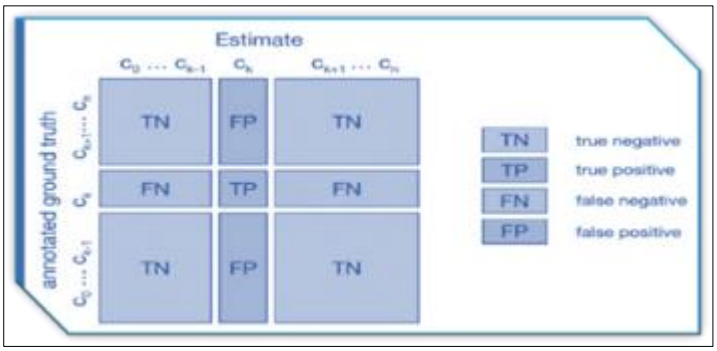


Figure 10 Figure of 3*3 confusion matrix

The formula to calculate the accuracy score:

$$Accuracy\ Score\ = \frac{(TN + TP + TN)}{(TN + FP + TN + FN + TP + FN + TN + FP + TN)}$$

And error rate = 1-Accuracy Score.

Fitting all models to our data, so our goal was classification to predict the best model by comparing it with the lowest test error rate to calculate the non-diagonal (false-negative and false-positive) elements from a confusion table for each predicted classification model (KNN classification, LDA, QDA, Decision Tree, Random Forest, Gradient Boosting, Adaboost etc.) [20].

3. Results and discussion

Table 4 contains the result of split validation approach to utilize 80% as train and 20% as test data set. It found that Graindent Boosting with parameters of max_depth = 3, n_estimators = 10, learning_rate = 1, random_state = 1 has the lowest error rate among all models.

Table 4 Result of split validation process

Validation Approach (Split test and train set)		
Model	**Error rate**	**Accuracy rate**
LDA	0.3261	0.67383
QDA	0.5119	0.4880
KNN(k=10)	0.7465	0.25341
Decision Tree	0.3254	0.67452
RandomForest	0.3254	0.67452
Gradient Boosting	0.1835	0.8165
Adaboost	0.3097	0.6902

Table 5 Result of cross validation approach

Cross Validation Approach (10 folds)		
Classification model	**Error rate**	**Accuracy rate**
LDA	0.3456	0.6543
QDA	0.5595	0.4404
KNN(K=10)	0.4125	0.5874
Decision Tree	0.3293	0.6745
RandomForest	0.0987	0.9013
Gradient Boosting	0.2969	0.7030
Adaboost	0.3120	0.6879

Table 5 implements all results of classification models with their error rate and accuracy rate to applied 10 folds cross validation. It presents RandomForest (n_estimators = 100, random_state = 1) contains the lowest error rate of 25.94%, setting the value of n_estimators to 100 indicates that the model will construct 100 decision trees. Every tree provides a vote towards the ultimate prediction in a classification problem. And the highest error rate is contained for Quadratic Discriminant Analysis model.

Table 6 Summary Result from all Classification model

Methods	**10 folds Cross Validation**		**Split validation Approach**	
Classification model	**Error Rate**	**Accuracy rate**	**Error rate**	**Accuracy rate**
LDA	0.3456	0.6543	0.3261	0.67383
QDA	0.5595	0.4404	0.5119	0.4880
kNN(k=10)	0.4125	0.5874	0.7465	0.25341
Decision Tree	0.3293	0.6745	0.3254	0.67452
Random Forest	0.0987 (lowest error rate)	0.9013	0.3254	0.67452
Gradient Boosting	0.2969	0.7030	0.1835	0.8165
Adaboost	0.3120	0.6879	0.3097	0.6902

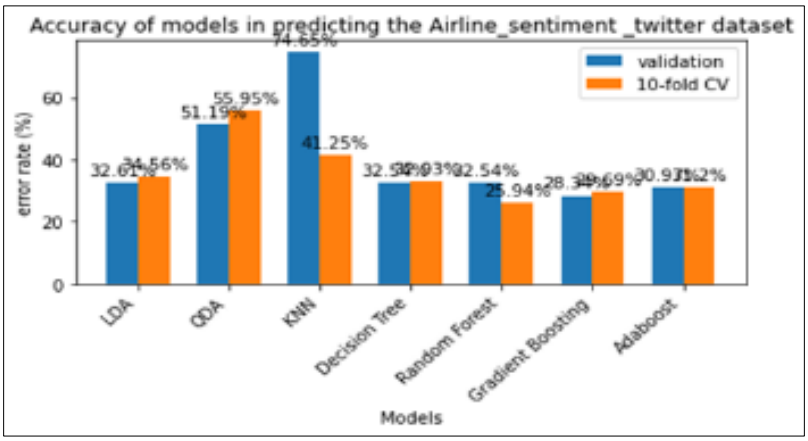


Figure 11 Comparable plot of error rates

Presented in this summary Table 6, and Figure 11 are making comparison of all of the outcomes with their respective accuracy scores for each and every model that was applied in this research. When Random Forest utilized the 10 folds cross validation strategy rather than the split validation method, it is evident that it achieved the lowest error rate of 9.87% among all of the methods.

Table 7 Best Model Selection

Model	**Random Forest**
Parameters	n_estimators = 100, random_state = 1
Validation Approach	10 folds cross validation
Error rate	0.0987
Accuracy Rate	0.9013

Table 7 displays the results of the final selected model, which achieved the lowest error rate. The model was trained using several combinations of parameters. In this twitter sentiment analysis Random Forest algorithm provides the most accurate prediction among all models, which contains the lowest error rate.

4. Conclusion

Both the discipline of data science and the field of sentiment analysis can benefit from the empirical contribution that this work makes. The purpose of this study is to examine and contrast the accuracy of a number of different classical classification methods. Regarding the field of sentiment analysis for airline services, there has been a remarkably limited amount of study conducted. Ensemble approaches, such as Random Forest, are among the classification methods that are utilized. These methods integrate a number of different classifiers to create a single powerful classifier that achieves an accuracy of 90.13%. The classifiers have achieved such high levels of accuracy that they can be utilized by the airline industry in order to carry out investigations aimed at ensuring overall client satisfaction. The most significant limitation of this research is the low number of tweets that were utilized in the training of the model. Despite this, there is still room for improvement in this analysis. Through the process of increasing the number of tweets, we are able to construct a more robust model, which ultimately leads to improved categorization accuracy. The methodology that is outlined in this analysis is one that airline firms can use to examine the data that is available on Twitter. Due to the extremely high volume of such reviews, it is necessary to have a significant number of specialists in order to do analysis and classification. Consequently, a number of different machine learning classifiers have been proposed, each of which has the potential to reduce the amount of human labor required to classify these reviews.

Future applications

The following future word is predicted for proposed solution.

- The proposed technique can be implemented utilizing big data technologies to handle enormous datasets.
- The proposed technique is applicable for classifying tweets into several categories, excluding those related to airlines.
- The proposed approach can be implemented with other classification algorithms, such as Support Vector Machines (SVM) and Naive Bayes, to enhance its better performance.

Limitations

Several limitations could be considered in this study-

- There are considerably more negative reviews than neutral and positive feedback, could be a reason why this study isn't very accurate.
- Improvements are still necessary to further increase classification accuracy.
- Sentiment analysis is sensitive and requires a sophisticated machine learning model, which is laborious and expensive to develop.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Cambria, E., Wang, H., & White, B. (2014). Guest editorial: Big social data analysis. Knowledge-based systems, 69(1), 1-2.

[2] Kamal, S., Dey, N., Ashour, A. S., Ripon, S., Balas, V. E., & Kaysar, M. S. (2017). FbMapping: An automated system for monitoring Facebook data. Neural Network World, 27(1), 27-57.

[3] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (Vol. 1, pp. 492-499). IEEE.

[4] Raihen, M. N., & Akter, S. (2024). Prediction modeling using deep learning for the classification of grape-type dried fruits. International Journal of Mathematics and Computer in Engineering.

[5] Raihen, M. N., Akter, S., Tabassum, F., Jahan, F., & Begum, S. (2023). A statistical analysis of excess mortality mean at Covid-19 in 2020-2021. Computational Journal of Mathematical and Statistical Sciences, 2(2), 223-239.

[6] Zakir, J., Seymour, T., & Berg, K. (2015). Big data analytics. Issues in Information Systems, 16(2).

[7] Shepherd, J. (2024, April 23). 23 essential Twitter (X) statistics you need to know in 2024. Herd 1. Retrieved from https://www.herd1.com.

[8] Shepherd, J. (2024, April 23). 23 Essential Twitter (X) Statistics You Need to Know in 2024. The Social Shepherd. Retrieved June 24, 2024, from https://thesocialshepherd.com/blog/twitter-statistics.

[9] MonkeyLearn. (n.d.). What is unstructured data? MonkeyLearn. Retrieved July 4, 2024, from https://monkeylearn.com/unstructured-data/.

[10] Datadition. (2024, June 3). Is Twitter structured data or unstructured data? Datadition. Retrieved July 4, 2024, from https://datadition.com/is-twitter-structured-data-or-unstructured-data/.

[11] Lamsal, R., Harwood, A., & Read, M. R. (2022). Twitter conversations predict the daily confirmed COVID-19 cases. Applied Soft Computing, 129, 109603.

[12] Raihen, M. N., Akter, S., Tabassum, F., Jahan, F., & Sardar, M. N. (2023). A statistical analysis of positive excess mortality at Covid-19 in 2020-2021. Journal of Mathematics and Statistics Studies, 4(3), 07-17.

[13] Batool, R., Khattak, A. M., Maqbool, J., & Lee, S. (2013, June). Precise tweet classification and sentiment analysis. In 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS) (pp. 461-466). IEEE.

[14] Raihen, M. N., & Akter, S. (2023). Forecasting Breast Cancer: A Study of Classifying Patients' Post-Surgical Survival Rates with Breast Cancer. Journal of Mathematics and Statistics Studies, 4(2), 70-78.

[15] Raihen, M. N., & Akter, S. (2024). Comparative Assessment of Several Effective Machine Learning Classification Methods for Maternal Health Risk. Computational Journal of Mathematical and Statistical Sciences, 3(1), 161-176.

[16] Arora, D., Li, K. F., & Neville, S. W. (2015, March). Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study. In 2015 IEEE 29th International Conference on Advanced Information Networking and Applications (pp. 680-686). IEEE.

[17] Shaheen, M., Awan, S. M., Hussain, N., & Gondal, Z. A. (2019). Sentiment analysis on mobile phone reviews using supervised learning techniques. International Journal of Modern Education and Computer Science, 10(7), 32.

[18] Nishida, K., Banno, R., Fujimura, K., & Hoshide, T. (2011, October). Tweet classification by data compression. In Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web (pp. 29-34).

[19] Rustam, F., Ashraf, I., Mehmood, A., Ullah, S., & Choi, G. S. (2019). Tweets classification on the base of sentiments for US airline companies. Entropy, 21(11), 1078.

[20] Sharma, N. K., Rahamatkar, S., & Sharma, S. (2019, December). Classification of airline tweet using naïve-Bayes classifier for sentiment analysis. In 2019 International Conference on Information Technology (ICIT) (pp. 70-75). IEEE.