Analysis for Airlines using Tweets

Yash Chungade

*Artificial Intelligence and Data Science*

*Vishwakarma Institute of Technology*

*Pune,411037,Maharashtra*

[yash.chungade20@vit.ed](mailto:yash.chungade20@vit.edu)u

### 

### 

***Abstract***— Social media has become one of the most important sources of information. Whether you want to excel in marketing or you want to work on crucial decision making, social media can be used to get all types of data which would fulfill the need. Applications like twitter, facebook, instagram etc are the apps on which people freely can express their opinions, choices and feelings.

The paper discusses the analysis done for the six famous airlines, on the basis of twitter application. The dataset is based on the suggestions, opinions, experience, good and bad reviews posted by the users on twitter. The analysis is done with the help of machine learning algorithms.

***Keywords— Machine learning, Natural language processing, Sentiment analysis. Airlines review system, Data visualization.***

### **INTRODUCTION**

Today’s generation is too cautious about doing and choosing anything. For example while purchasing a new mobile phone, the buyer would first do a deep analysis as to which phone to buy. Reference is to prove that we as a human would take a decision on the basis of other person’s experience or on other’s opinions. Now just think how much time and mental strength the person spends studying for purchasing it.

This paper discusses a machine learning model which is trained for doing a detailed analysis of a particular airline. When a person wants to travel to different parts of the world and he doesn't know which airline is best suited for him, then this model can help that person to give a basic analysis of different airlines. Basically the model has used a dataset which contains tweets of twitter users posted for six reputed US airlines. The airlines are Delta, Southwest, Virgin America, United, US Airways and American Airlines.

The model has predicted from all the tweets whether it has a positive, negative or neutral opinion about a particular airline. Task of opinion mining is done by the model discussed by the paper. It is also known as sentiment analysis. Sentiment analysis refers to analyzing an opinion or feelings about something using data like text or images, regarding almost anything. In this concept the useful information in the form of opinions and sentiments is extracted and used for decision making. It has a huge application in various sectors. For example it helps companies in their decision-making process. For instance, if public sentiment towards a product is not so good, a company may try to modify the product or stop the production altogether in order to avoid any losses.

The main concept of sentiment analysis is Natural language processing. NLP is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

A total of four different supervised machine learning algorithms were used for testing the model and the result has been derived by doing a comparative study of these used algorithms.

### **LITERATURE REVIEW**

“Sentimental Analysis of Airline Twitter Data''[1] It is very easy for the airline industry to obtain feedback from insightful data sources like Twitter in order to conduct sentiment analysis of individual customers. Consumers who want to know who is who and what is what in their daily lives can't help but turn to Twitter's sentiment analysis. This article classifies the sentiment of Twitter posts by presenting the results of machine learning algorithms using R and Rapid Miner. Tweets are pre-processed, extracted, and categorized into neutral, negative, and positive sentiment before results are aggregated. They used the Naive Bayes algorithm to classify attitudes expressed in recent tweets from different airlines.

"Topic Modeling and Sentiment Analysis of Online Airline Reviews"[2] The data collected consisted of over 14,000 online reviews from 27 airlines. They combined topic modeling and sentiment analysis with the data they collected to identify the types of key phrases that appear in online reviews. Topic modeling revealed that 'seat', 'service', and 'meal' were key flight issues through frequency analysis.

"Algorithms and Methods for Sentiment Analysis Using Text and Emoticons."[3] The main contribution of this study is an algorithm, methodology, and emoticon dictionary for evaluating sentiment in social media data (both text and emoticons). . For example, airline data collected by Twitter. This study demonstrates the effect of considering emoticons in addition to text when assessing emotion.

"Review of Sentiment Analysis Methodology, Practice, and Applications."[4] The basics and levels of sentiment/opinion mining are the main topics of this study. There are many strategies and techniques for extracting emotion from information. The analysis here serves as a presentation of machine learning techniques. Sentiment analysis uses various classification techniques to categorize data into positive, negative, and neutral ratings. This study shows that dictionary-based methods can be very successful, and machine learning techniques such as SVM, naive Bayes, and neural networks have the highest accuracy and can be considered as the underlying learning technique.

"Sentiment Analysis and Opinion Mining Applied to Scientific Essay Review"[5] This article examined the use of sentiment analysis techniques in the field of essay evaluation. Specifically, a review of 382 research papers (in non-empty Spanish) at an international conference evaluated supervised methods (NB and SVM), unsupervised methods (scoring algorithms), and a hybrid approach (HS-SVM).

"Sentiment Analysis of Airline Tweets Using Mutual Information for Feature Selection"[6] This study uses a dataset from a selection of tweets about US airlines. The dataset is hosted on Kaggle and lots of metadata is available so feature selection experiments can be completed quickly. Feature selection in this study uses a mutual information approach. They chose this method because, unlike previous references, it successfully measures correlations between different attributes. Naive Bayesian classification in this study yields faster training data on the airline dataset, among other derivable factors. For this airline dataset, the SVM linear classifier gives the best classification accuracy.

In this article, “Sentiment Analysis of Tweets from Gulf Airlines Using Machine Learning”[7] they build a tweet sentiment analyzer that classifies tweets as positive or negative. The proposed sentiment analyzer can be used by various companies, groups or individuals. To achieve this, they use machine learning to create a tweet classifier and natural language processing to analyze the text. First they preprocess the dataset, Then the preprocessed tweets were they used to extract features using TF-IDF and bag-of-word algorithms. To determine which features the classifier had the highest predictive accuracy, they further ranked them by 1-gram, 2-gram, and 3-gram. When using the trigram feature, logistic

regression outperforms all other classification algorithms in terms of accuracy on validation and test data.

In this article, “New Deep Learning-Based Sentiment Analysis of Twitter Data for US Airlines”[8] they proposed a new form of deep learning with different capabilities for classifying sentiment. In this study, they proposed a new deep learning model for analyzing a dataset consisting of tweets from six major US airlines using multi-class sentiment analysis and multi-word embedding techniques. Their selection system combines these features with various deep learning techniques for term embedding and emotional document classification. The first step in their methodology is the extraction of raw DNN data and preprocessing of CNN tweets and then Interpretations of positive/negative/neutral tweets using a 3-class data set and data set accuracy scores are generated from the test set product.

“An Effective Way to Predict Polarity of Airline Tweets Using Sentiment Analysis”[9] In this paper, they consider three well-known airline tweets to show how to efficiently and easily detect attitudes on Twitter. Score calculation served as the basis for identifying positive, negative, and neutral moods. The tweet about Emirates was more positive than the other two, while the tweet about Indigo was more negative. Tweets about Qatar Airways were more neutral. The downside of this strategy is that applying it to sarcastic tweets may not yield the intended results. Because different conclusions can be drawn depending on the placement of positive and negative terms.

The paper, “Sentiment Analysis of US Airlines Tweets Using LSTM/RNN”[10] examines aviation industry social media tweet data and shows that the training set performs better. In addition, bidirectional LSTM networks can improve accuracy. They investigated sentiment analysis using Recurrent Neural Network (RNN) models and Long-Short Term Memory (LSTM) networks (Bi-LSTM). It can handle long-term relationships by transforming memory into network models and introducing visualizations for prediction. The results showed significantly improved classification accuracy of 80% on the training set and 20% on the test set, demonstrating the reliability of the model for future predictions.

“Airline Sentiment Visualization, Consumer Loyalty Measurement and Prediction using Twitter Data”[11] The goal of this study is to forecast consumer loyalty by analyzing tweets about airlines with bases in Europe, India, Australia, and America. TextBlob analyzer is used to do sentiment analysis. The mean sentiment ratings for each airline, as well as their variations over time, are calculated from the tweets and visually shown. Using visualization techniques, the phrases for complaints and praises are shown. The data acquired from Twitter is intended to be used in a revolutionary approach to gauge customer loyalty. Here they use Random Forest, Decision Tree, and Logistic Regression classifiers. Random Forest exhibits a maximum classification accuracy of 99.05% based on 10-fold cross validation.

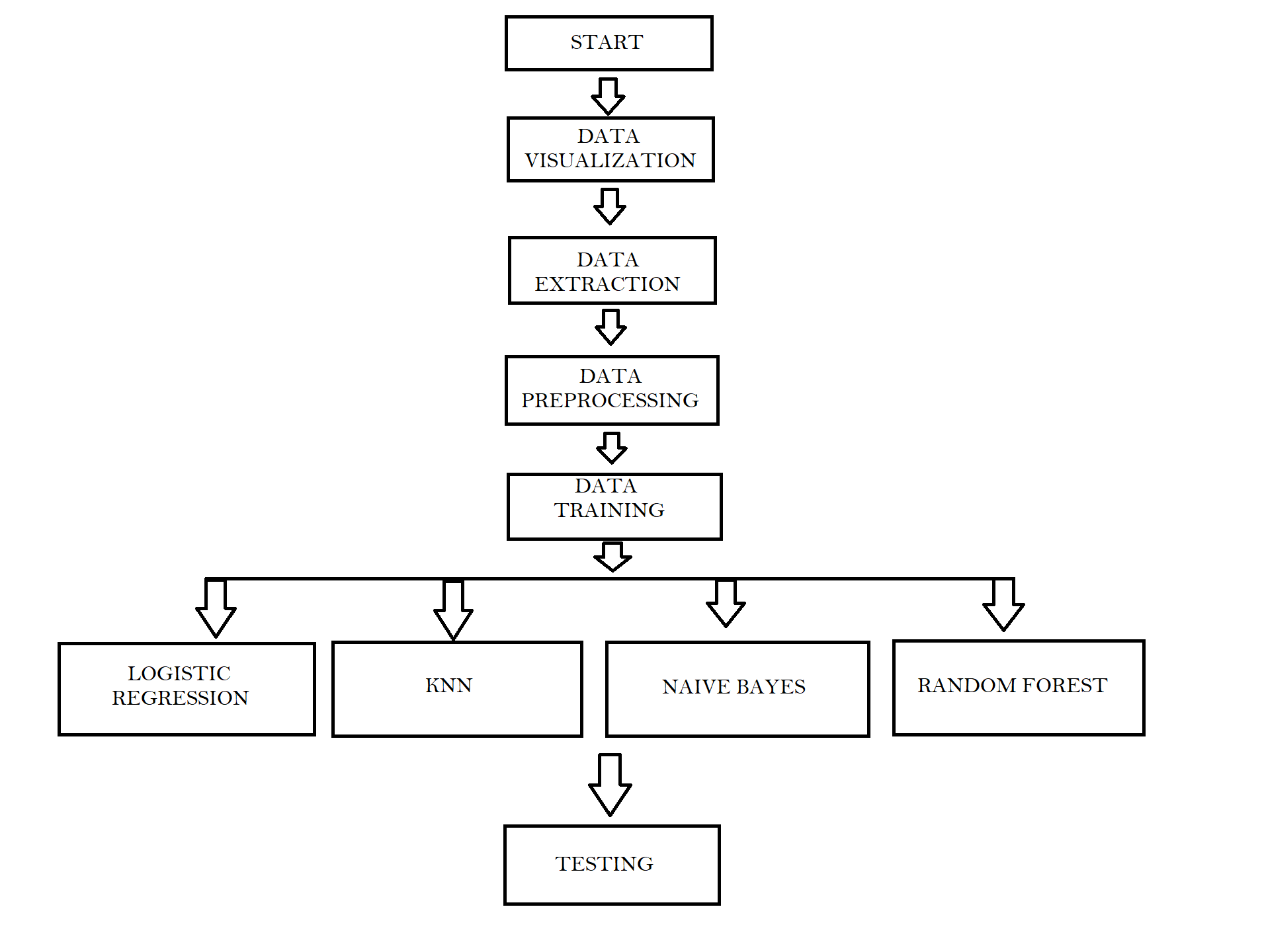
The authors Deb Dutta Das, Sharan Sharma, Shubham Natani, Neelu Khare and Brijendra Singh [12] have categorized the sentiment of Twitter posts by showcasing the findings of a machine learning system that makes use of R and Rapid Miner. The tweets are pre-processed, extracted, and then categorized into neutral, negative, and positive attitudes before the findings are summarized overall. The Naive Bayes algorithm was used to categorize the attitudes expressed in recent tweets from the various airlines.

Rana Al-Qahtani in their paper “Predict Sentiment of Airline Tweets Using ML Models”[13], Through sites with a wealth of opinion-rich data, such as Twitter, sentiment analysis may be utilized to analyze people' attitudes or sentiment. This study looks at the use of two machine learning (ML) and four deep learning (DL) techniques to predict sentiment from tweets from US airlines. It analyzes the content tweets text of consumers' emotions in US airline business services.

“Application of Sentiment Analysis: Assessing the Reliability and Validity of the Global Airlines Rating Program” [14] the author Guohuibin Li in this paper divides airlines into five categories and refreshes the data once a year. The outcome has been used by a growing number of airlines for marketing initiatives. The validity and dependability of this programme are under scrutiny by researchers, and little literature has addressed these concerns. The reliability, discriminating validity, and external validity of this software will be evaluated in this research using sentiment analysis based on traveller evaluations from Skytrax and linked.According to airlines' expertise on Twitter, the results show that the worldwide Airlines Rating programme has a low degree of dependability, a weak discriminating validity, and a weak external validity. The text mining paradigm put forward by Kayser & Blind(2016) is theoretically further enhanced by this study. The existing research framework will be very important for future research. Practically speaking, this theory raises concerns about how this programme may be applied to travelers and airlines.

“An Exploration of Airline Sentimental Tweets with Different Classification Model”[15] This paper's primary goal is to analyze the Twitter airline dataset in order to identify the best and worst airlines as well as to forecast the most frequent problems encountered when using an airline service. Next, negative tweet word clouds are generated, and using geographic analysis, the location of the consumer who tweeted adversely is predicted and shown. After that, the dataset was trained and tested using seven different classifiers, including the Logistic Regression classifier, the KNeighbors classifier, the SVC classifier, the Decision Tree classifier, the Random Forest classifier, the AdaBoost classifier, and the GaussianNB classifier. The outcomes of our tests show that when it comes to categorizing twitter data by sentiment, the Random forest method performs well in practice.

**III. PROJECT FLOW DIAGRAM**



**IV. ALGORITHM**

For achieving the problem statement a typical machine learning pipeline is followed. As it is using ML techniques, it has been implemented using pyt hon. First the useful libraries and dataset was needed to be imported. . Then an exploratory data analysis was performed. It was done for checking and finding any type of trends in the dataset. Next, the textual data is needed to be converted to the numeric data. For this, text preprocessing was done so that data can be used by a machine learning algorithm. Finally, different machine learning algorithms can be used to train and test the proposed sentiment analysis models.

Now lets see step by step how the model discussed in the paper is proposed-

1. First all the important libraries were installed. Then the twitter dataset which was freely available online was imported.
2. Then after importing the dataset, it is very important to understand the dataset. So a proper data analysis was performed for better understanding of the dataset. Also various data visualization tools were used for loting the dataset in a graphical manner.
3. Using the Seaborn library to view the average confidence level for the tweets belonging to three sentiment categories.
4. Next step is to perform some preprocessing on the data and then convert the numeric data into text data.
5. Now the dataset is divided into feature and label sets. Our feature set will consist of tweets only.Our label set will consist of the sentiment of the tweet that we have to predict. To create a feature and a label set, we can use the ‘loc’ method off the pandas data frame.
6. Then the data cleaning step is performed. The tweets from the dataset may contain many slang words, unwanted words and punctuation marks. So it is important to clean the tweets before they can be used by the machine learning model for training purpose.
7. Now data preprocessing is done. First, removing all the special characters from the tweets is done. Then removal of all the single characters which was left as a result of removing the special character was done. At last replacement of all the multiple spaces with single spaces was performed.
8. Next step consists of converting the textual data into the numeric data. This is needed to be done as statistical algorithms use mathematics to train machine learning models.
9. Now the data is all ready to be given to ML model for training. Before that, the data is needed to be splitted into testing and training sets. So it is done.The training set will be used to train the algorithm while the test set will be used to evaluate the performance of the machine learning model.
10. At last the model is trained. The proposed model is trained using four different ML models.
11. Then the training set is used for predicting the output and for checking the performance of trained models.
12. Evaluation of all the trained models is done and the best model is suggested on the basis of performance of all models.

**V. EXPERIMENTATION PART**

The best model which is prepared for predicting the opinion about a particular airline base upon the experiments done by using different machine learning algorithms. In total, four supervised algorithms were tested and four models were created. Below are the four models which were used for the testing purpose:

1. Logistic Regression:

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no and so on.

The model which was trained by using this algorithm had an accuracy of 78%.

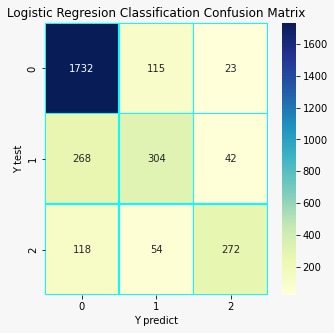


Fig 2: Confusion matrix for Logistic Regression

1. K-nearest Neighbour:

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

The model which was trained by using this algorithm had an accuracy of 69%.

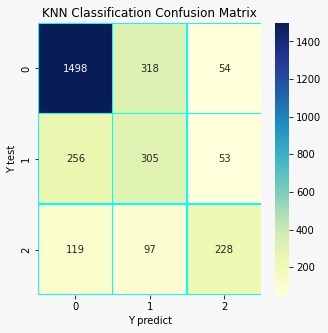


Fig 3: Confusion matrix for KNN Classification

1. Naive Bayes Algorithm:

Naive Bayes models are a group of extremly fast and simple classification algorithms that are often suitable for very high-dimensional datasets. Because they are so fast and have so few tunable parameters, they end up being very useful as a quick-and-dirty baseline for a classification problem.

The model which was trained by using this algorithm had an accuracy of 41%.

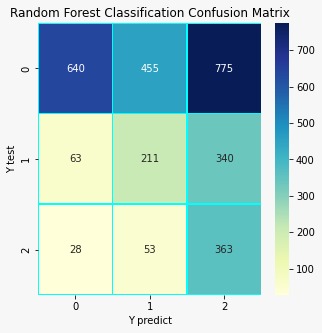


Fig 4: Confusion matrix for Naive Bayes

1. Random Forest Algorithm:

As the name suggests, ***"***Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.***"*** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The model which was trained by using this algorithm had an accuracy of 75%.

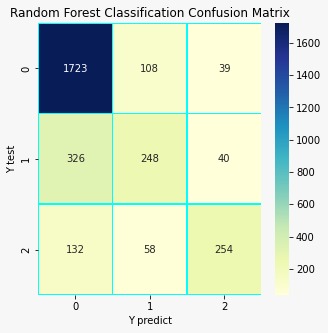


Fig 5: Confusion matrix forRandom Forest

**VI. RESULTS**

As a first part, data visualization was performed. So below is the image which shows that the dataset used contains how much tweets for which airlines.

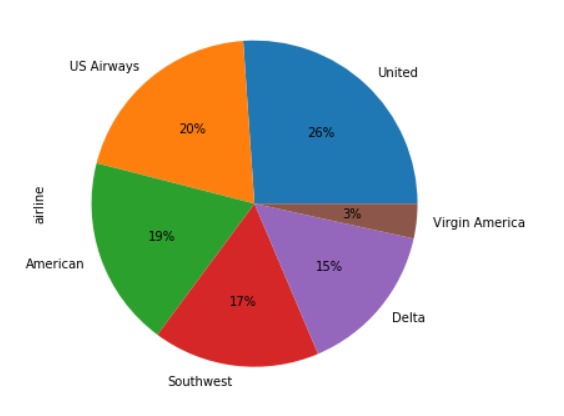
****

Fig 6: Pie chart for analysis of dataset.

Then using natural language processing data was cleaned and then it was analyzed for number of negative, positive and neutral tweets as of overall dataset.

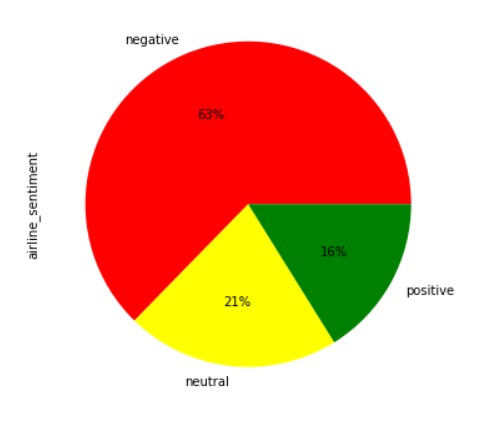
****

Fig 7: Pie chart for opinion mining

Then the specified airline based analysis was done.

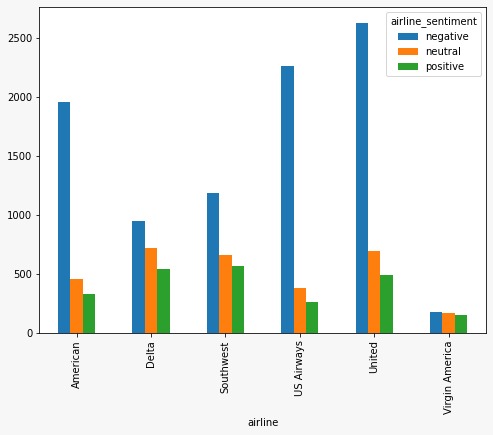
****

Fig 8: Bar chart showing type of tweets for all airlines

**VII. CONCLUSION**

This paper discusses a machine learning model which is trained for doing a detailed analysis of a particular airline. Four different machine learning algorithms which are linear regression,naive bayes classification, k-nearest neighbor and random forest algorithm are used to train models. For better understanding of the dataset various data visualization techniques were used to represent the data in different types of charts and graphs. different data preprocessing and data cleaning are practiced on the data and data is fed to train the model.

Results then are obtained from the model. This model can be used by a person who wants to know which airline is best for his traveling as dataset of this model used is based on the experience of previous customers. Four algorithms are used currently in this model, but in future more advanced algorithms can be used in order to obtain more accuracy.

To conclude our discussion, it can be stated that for the datasheet used for performing analysis, Logistic Regression has given the best accuracy. Then Random forest, followed by K-NN, and last

Naive Bayes.

**VIII. REFERENCES**

1. Deb Dutta Das, Sharan Sharma, Shubham Natani, Neelu Khare and Brijendra Singh, “Sentimental Analysis for Airline Twitter data”, 2017.
2. Hye-Jin Kwon 1 , Hyun-Jeong Ban 2 , Jae-Kyoon Jun 1 and Hak-Seon Kim, “Topic Modeling and Sentiment Analysis of Online Review for Airlines”, 2021.
3. Mohammad Aman Ullah , Syeda Maliha Marium, Shamim Ara Begum, Nibadita Saha Dipa, “An algorithm and method for sentiment analysis using the text and emoticon”, 2020.
4. Pooja Mehta, Dr.Sharnil Pandya, “A Review On Sentiment Analysis Methodologies, Practices And Applications”,2020.
5. Brian Keith Norambuena∗ , Exequiel Fuentes Lettura and Claudio Meneses Villegas, “Sentiment analysis and opinion mining applied to scientific paper reviews”,2019.
6. Hastari Utama, “Sentiment Analysis in Airline Tweets Using Mutual Information for Feature Selection”, 2019.
7. Mazen M. Hrazi, Abdulrahman M. Althagafi, Abdullah T. Aljuhani, Jenifar Rahman, Md. Mahfuzur Rahman, Mohammad Shorfuzzaman, “Sentiment Analysis of Tweets from Airlines in the Gulf Region Using Machine Learning”, 2021.
8. Khan Md. Hasib , Md. Ahsan Habib , Nurul Akter Towhid , Md. Imran Hossain Showrov, “A Novel Deep Learning based Sentiment Analysis of Twitter Data for US Airline Service”, 2021.
9. Adarsh M J, Dr. Pushpa Ravikumar, “An Effective Method of Predicting the Polarity of Airline Tweets using Sentimental Analysis ”, 2018.
10. Ms.R.Monika, Dr.S.Deivalakshmi, Dr.B.Janet, “Sentiment Analysis of US Airlines Tweets using LSTM/RNN ”, 2019.
11. Rida Khan, Siddhaling Urolagin, Airline “Sentiment Visualization, Consumer Loyalty Measurement and Prediction using Twitter Data”, 2018.
12. Deb Dutta Das, Sharan Sharma, Shubham Natani, Neelu Khare and Brijendra Singh, “Sentimental Analysis for Airline Twitter data”, 2017.
13. Rana Al-Qahtani, “Predict Sentiment of Airline Tweets Using ML Models”, 2021.
14. Guohuibin Li, “Application of Sentiment Analysis: Assessing the Reliability and Validity of the Global Airlines Rating Program”, 2017.
15. M.Vadivukkarasi, N.Puviarasan, P.Aruna, “An Exploration of Airline Sentimental Tweets with Different Classification Model”, 2018.