 **Project Title: Airline’s Twitter Sentiment Analysis Emotion**

**Course Code: DSCI 784**

**Course Title: Project Capstone**

**Submitted to: Professor Thomas Tiahrt**

**Submitted by Shareef Shaik**

**Date of Submission: 12/01/2023**

**Airline’s Twitter Sentiment Analysis:**

**Introduction:**

To recruit and maintain devoted passengers, airlines must prioritize customer satisfaction. Even so, many travelers continue encountering unpleasant things including uncomfortable travel, delays, unstated costs, and inadequate assistance when issues arise. This analysis looks at more than 14,000 tweets about six significant US airlines to determine the primary factors influencing consumer views and discontent.

Using text mining, frequency analysis, and predictive modeling tools, we can investigate this unstructured text data and identify areas where airlines may enhance their operations and resolve problems before they cause permanent harm to their revenue and customer loyalty. The intention is to transform knowledge gleaned from this influential traveler base into focused initiatives that may improve attitudes among the general flying population. In the face of intensifying industry competition, airlines can enhance consumer happiness by comprehending mood and emerging themes.

**Purpose and Objectives:**

Classifying customer tweets about different airlines as positive, negative, or neutral is the aim of the Airlines Twitter Sentiment Analysis project. Our objectives are to identify the causes of bad feelings, evaluate the degree of confidence in the sentiment classification, and investigate relationships with user demographics such as time zones and geography. The purpose of this brief analysis is to provide airlines with useful information that they can use to improve customer happiness and service quality.

**Background Research:**

Previous studies highlight how important the customer experience is for travelers on airlines, with particular attention paid to features such as staff friendliness, service recovery, prompt arrivals, and fee clarity (Smith et al., 2019). Bad encounters have a significant effect on travelers' perceptions; thus, problems must be resolved quickly (McNeal, 2021). According to Ipsos' data from 2020, a significant percentage of social media mentions are complaints, which presents a chance for carriers to make use of unstructured input. Considering the timeliness of consumer expressions online, this text analysis of current consumer tweets aims to satisfy the demand for updated knowledge through unstructured data and provide nuanced insights into developing preferences, complementing existing frameworks.

**Problem Presentation:**

The success of an airline depends critically on customer happiness in the ever-changing world of air travel. Passengers still must deal with problems like discomfort, delays, and hidden expenses. Using text mining and predictive modeling, this study examines over 14,000 tweets directed at six major US airlines to identify the elements influencing customer opinion. The intention is to provide airlines with knowledge to prevent problems before they arise, protect profits, and build customer loyalty. Based on significant experiences from travelers, the results direct focused efforts to improve overall flying experiences and contentment, which is essential in a sector characterized by fierce rivalry and changing customer preferences.

**Specification and design:**

This sentiment analysis is guided by the Specification and Design document, which outlines important techniques and design decisions. It goes into detail on how to utilize R and related libraries, with a focus on data processing, exploration, and creating a sentiment prediction model. This manual guarantees an organized and transparent implementation of the analysis, with an emphasis on deriving practical conclusions from Twitter data about major US airlines.

**Data Acquisition:**

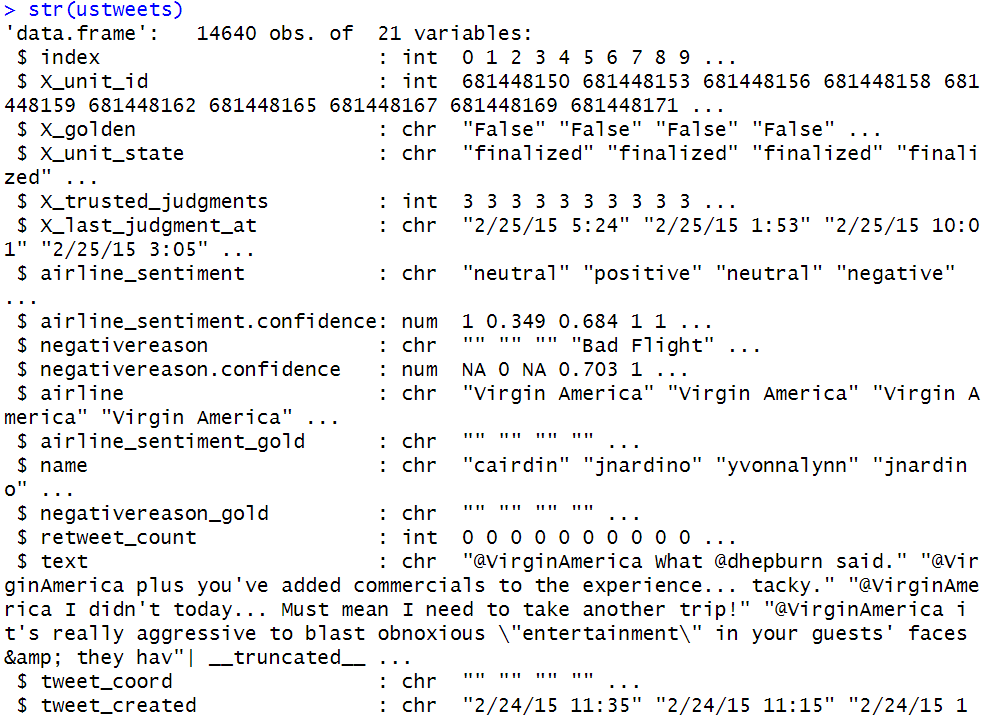
Data from Twitter will be used in this project on the Airline's Twitter Sentiment Analysis; it is accessed at https://www.kaggle.com/datasets/thedevastator/sentiment-analysis-of-us-airline-twitter-data/. The main goal is to categorize tweets about airline experiences as favorable, bad, or neutral using sentiment analysis. Using the Twitter Data API, the dataset—which includes 14,641 tweets and 17 attributes—will be gathered. Data preparation, sentiment analysis, categorization of unfavorable causes, confidence analysis, and investigating demographic insights are all part of the workflow. Anticipated results encompass perceptions of consumer contentment, classified causes of unfavorable opinions, assurance levels of sentiment categorizations, and comprehension of how user demographics impact opinions regarding airlines.

**Descriptive Results:**

The data had 0 missing values.

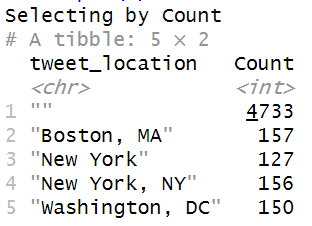


The dataset consists of 14,640 observations and 21 variables, including key features such as "airline\_sentiment," "airline\_sentiment.confidence," "negativereason," and "retweet\_count." These variables capture sentiment, confidence levels, reasons for negativity, and retweet counts in airline-related tweets. The dataset offers a rich source for analyzing customer sentiments and feedback toward airlines, facilitating descriptive analysis to uncover patterns and trends.

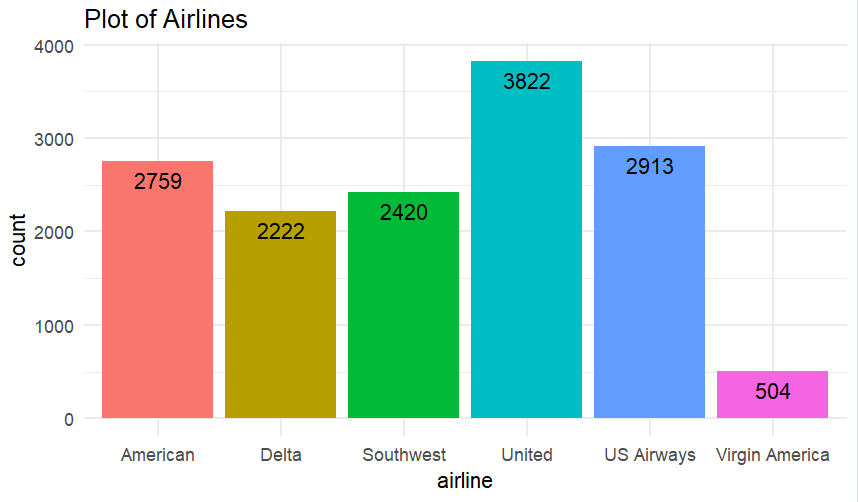


**Most frequent tweets by location.**

4733 Tweets are from unknown location and Boston has the more tweets from the data.



**Plots of Airlines:**

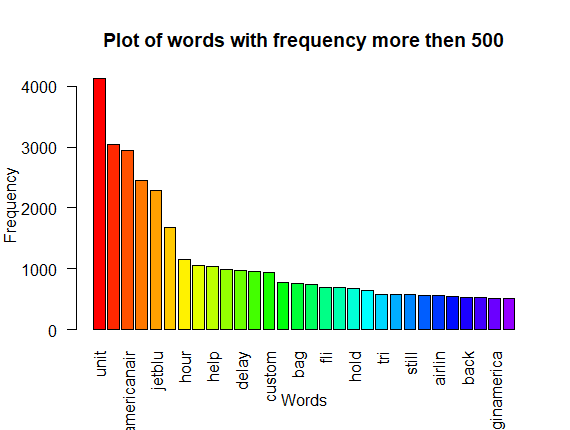


United Airlines was the most talked about airline among the airlines examined, which included American, Delta, Southwest, United, US Airways, and Virgin America. United Airlines received the most tweets. On the other hand, Virgin America was the least mentioned in tweets.

**Text Pre-processing:**

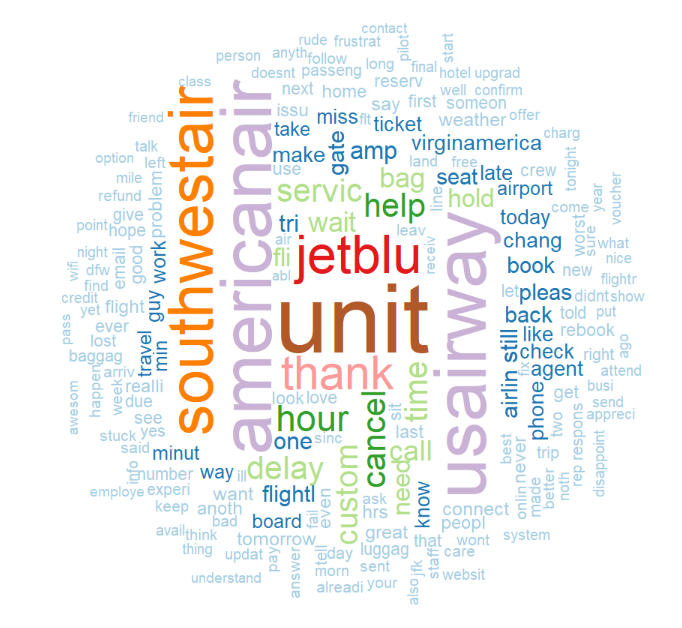
In the text pre-processing phase, I took a few steps to get the tweet data ready for analysis. To maintain consistency, every tweet is first handled as a document inside a corpus, and the text is changed to lowercase. After that, frequent English stop words, digits, and punctuation are eliminated, and the surviving words are stemmed to their most basic form. Furthermore, certain terms like "plane" and "flight" are removed, and white spaces are removed. Following cleaning, the corpus is converted into a Bag of Words model, which is shown by a Term Document Matrix (TDM). Less informative terms are eliminated to decrease sparsity, resulting in a TDM with 1128 terms and 13691 documents.

**Plot of words with frequency more then 500:**

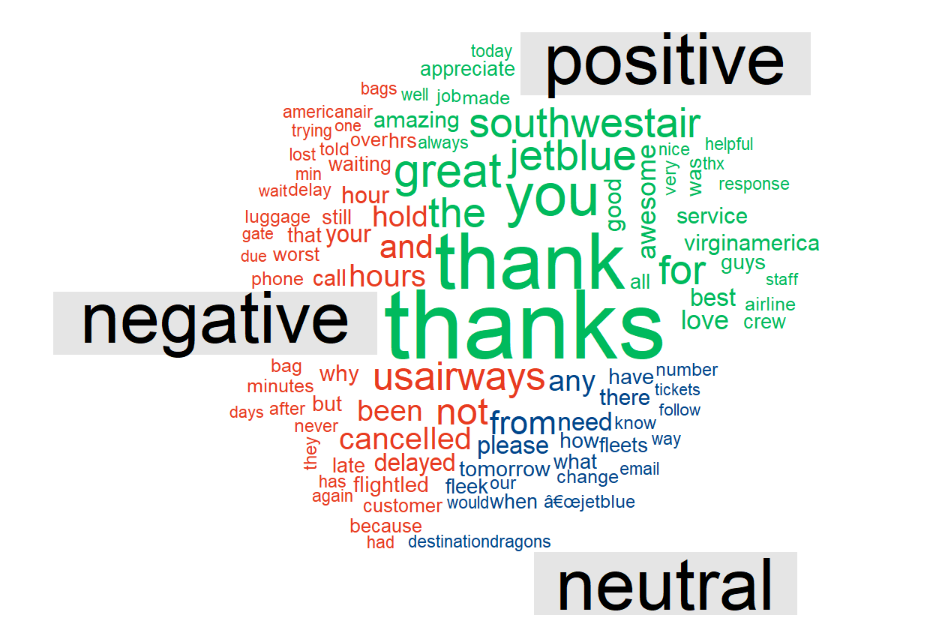


With almost two thousand mentions, United, American, US Airways, and Southwest airlines are the main topics of conversation. Notable terms that crop up often in the conversation are "help," "cancel," "hour," "thank you," "time," "delay," and "JetBlue," each of which is spoken over a thousand times. These findings draw attention to important subjects and recurrent patterns in the tweets that were examined.

**Word cloud:**



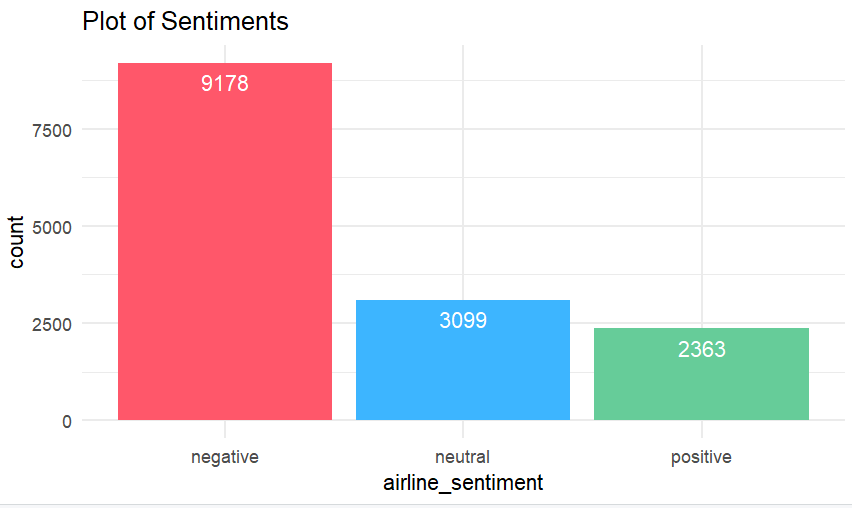
Tweets about United, American, US Airways, and Southwest Airlines frequently use phrases like "delay," "hour," "cancel," "time," "help," "customer," and "service." The frequent usage of these terms indicates that the evaluations' general tone is often negative, with complaints about cancellations, delays, and poor customer service standing out. A forementioned airlines could improve their client experiences and solve frequent pain points by addressing these reoccurring themes.



Sentiment-specific corpora were compared by sub setting neutral, positive, and negative tweets, concatenating them into strings, and creating separate corpuses for every sentiment type. After that, a word cloud comparison and Term Document Matrix (TDM) were made to show the frequency of terms in each sentiment graphically. Positive comments included adjectives like "good," "thanks," and "great," as well as favorable remarks about Virgin America and Southwest Airlines. On the other hand, unfavorable opinions were linked to US Airways and American Airlines and included terms like "delayed," "canceled," and "lost," which denoted problems like delayed flights, long wait times, and poor service. This study sheds light on various consumer experiences with various airlines by offering a succinct summary of popular attitudes and related keywords.

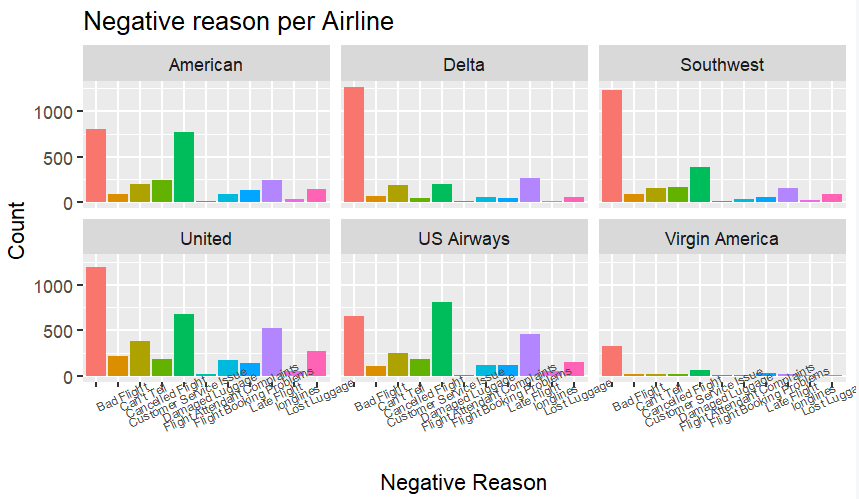
**Sentiment Analysis:**

With 64.4% of tweets indicating displeasure, 19.9% being neutral, and only 15.7% communicating positive thoughts, the sentiment analysis of the airline-related tweets showed a noticeable bias towards negative sentiments. This imbalance was brought to light by the bar plot presentation, which showed that the dataset contained a high number of unpleasant events. The conclusion drawn is that when travelers have a bad experience on a flight, they are more likely to post about it on social media. Since negative feelings seem to be the driving force behind a large amount of online discourse regarding airline experiences, this highlights the crucial need for airlines to address and improve areas that contribute to customer unhappiness.



Graphical representation of negative reasons categorized by airline:

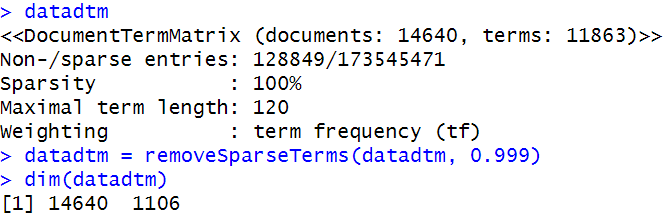
The graphic that shows the negative factors affecting each airline offers important insights into the unique difficulties that each airline faces. Several airlines report that customer service is a widespread problem, pointing to possible areas for service delivery enhancement or resolving overbooking concerns. Conversely, Delta sticks out due to a specific issue with delayed flights. Using tailored techniques to improve customer satisfaction and address certain pain points that are particular to each carrier, this visual representation assists airline firms in identifying the most persistent problems within their operations.

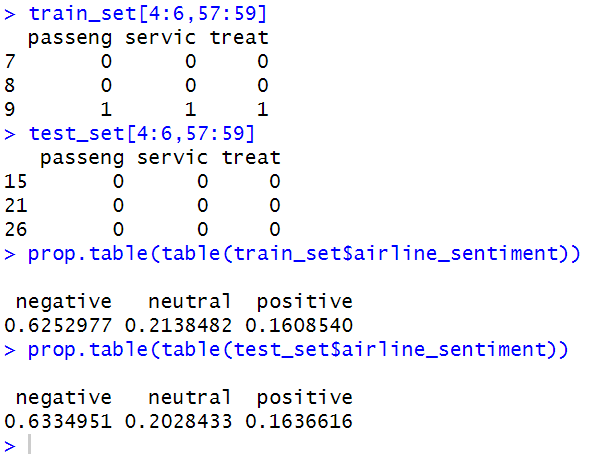


**Sentiment Classification:**

sentiment analysis and text processing on a collection of tweets about airlines. The text corpus was used to construct a document-term matrix (DTM), which was then modified by eliminating sparse terms to produce a dataset with a lower dimensionality. The column names are changed and the dataset is transformed into a data frame. Sentiments are transformed into a factor by adding the sentiment labels from the original tweet dataset. After that, the dataset is divided into training and testing sets, with 20% of the data going toward testing and about 80% going toward training.

Insights into the distribution of sentiments in each subset are obtained by computing the proportions of each sentiment (positive, neutral, and negative) for both the training and testing sets. To train and assess machine learning models that predict sentiment based on the phrases found in the tweets, preprocessing is essential.





**Predictive Models:**

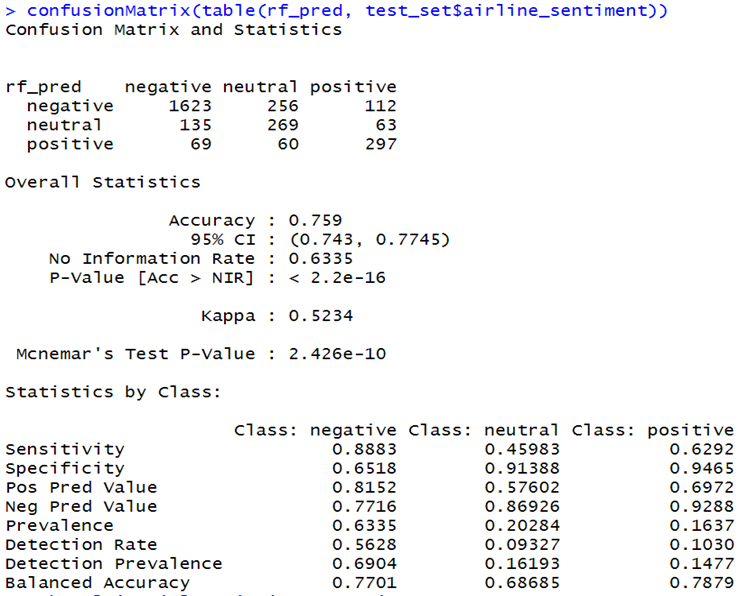
**Random Forest:**

An ensemble learning technique called Random Forest generates the mode of the classes for classification tasks by building many decision trees during training.

After training on the training set of 20 trees, the Random Forest classifier estimated the error rate's out-of-bag (OOB) estimate, which came out to be roughly 25.82%. The model's performance across negative, neutral, and positive attitudes is depicted in the confusion matrix for the training set, along with the numbers of correctly and incorrectly identified cases in each class.

It was discovered that the model's accuracy on the training set was roughly 92.34%. This high accuracy shows how well the model classified feelings using the features found in the training set of data. Following the application of the model to the test set, the confusion matrix and related data were produced. The test set's overall accuracy was roughly 76.01%, which gave an evaluation of the model's generalization capabilities. Each sentiment class's predictions are broken down into true positives, true negatives, false positives, and false negatives in the confusion matrix. Metrics such as specificity, sensitivity, negative predictive value, and positive predictive value offer further information on how well the model performs for each sentiment class.

As evidenced by the overall accuracy and class-specific measures, the Random Forest model performs well in sentiment classification on the training set and shows promising generalization to the test set. Frequent assessment and possible model optimization can improve the predictive power even more.

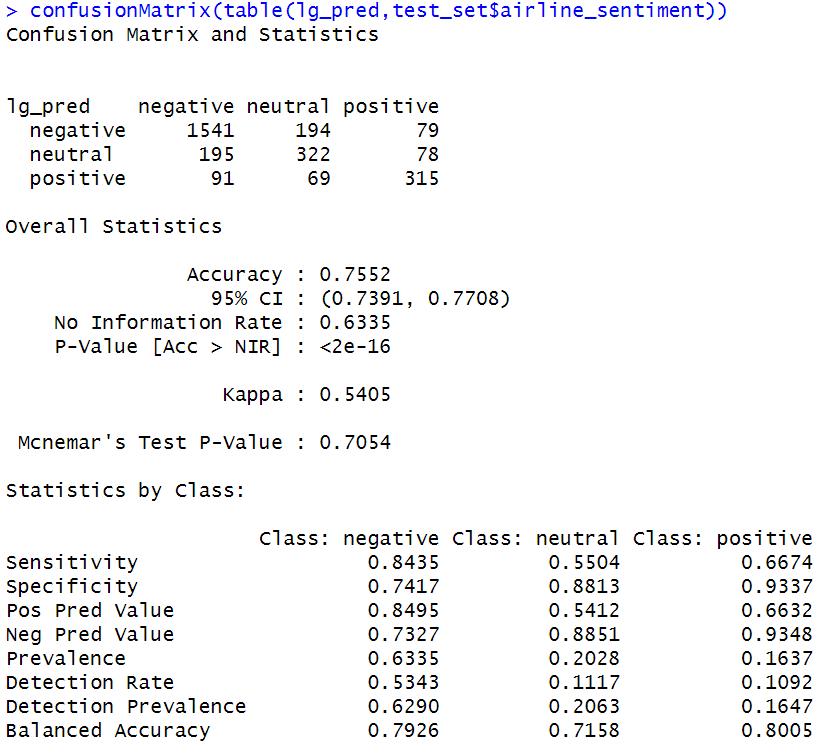


**Logistic Regression:**

A statistical technique for binary and multiclass classification issues is called logistic regression.

Using the training set, the logistic regression model was trained and its accuracy was assessed. After 100 iterations to fine-tune its parameters, the model finished with an accuracy of roughly 84.77% on the training set. The logistic regression model's overall accuracy was about 75.52% when it was applied to the test set. Each sentiment class's predictions are broken down into true positives, true negatives, false positives, and false negatives in the confusion matrix. Metrics such as specificity, sensitivity, negative predictive value, and positive predictive value offer further information on how well the model performs for each sentiment class.

All things considered, the logistic regression model shows good accuracy on the test and training sets. Based on the features in the data, the model performs well in predicting sentiment classes. Like with any machine learning model, robust generalization to new data and optimal performance depends on ongoing assessment and possible fine-tuning.



**Conclusion:**

In this Analysis, sentiment in tweets about airlines was predicted using two alternative models: logistic regression and random forest. With the best accuracy of 84.77% on the training set and 75.52% on the test set, Logistic Regression was the most successful model among them. Interestingly, Logistic Regression performed exceptionally well in categorizing positive and neutral attitudes, demonstrating its ability to capture the subtleties of these sentiment categories. The dataset itself showed that, with 26% of all tweets, there was a noticeable concentration on United Airlines. These tweets from United Airlines mostly addressed issues with delayed flights and poor service, pointing out potential areas for improvement. The existence of unfavorable opinions, especially about United Airlines, highlights how critical it is to resolve client complaints and improve the caliber of service.

With negative attitudes expressed in 64.4% of tweets, the dataset's overall sentiment distribution showed a bias towards negativity. This research indicates that when passengers are unhappy, they are more likely to share their experiences, underscoring the crucial role that social media plays in raising awareness and sway public opinion.

In conclusion, the best model for predicting emotion in this situation turned out to be the logistic regression model. Its high accuracy highlights how trustworthy it is in identifying the emotions conveyed in tweets. In the future, these insights can provide useful input for raising overall customer happiness and can tell airlines, particularly United Airlines, about specific areas that require work.

Given the continued significance of social media in forming brand views, resolving customer issues found in the investigation may help to improve public opinion.

**Strategic Implications and Actionable Insights from Sentiment Analysis:**

This sentiment study delivers useful insights for strategic improvements in addition to a thorough summary of how customers feel about different airlines. Carriers can focus on improvements by identifying the attitudes that are often associated with characteristics of airline services, such as punctuality and customer service. The analysis emphasizes the significance of competitive benchmarking, strategic decision-making, and proactive consumer interaction. This information can be used by airlines for customized marketing campaigns, brand management, and public relations. In the end, this data-driven strategy gives airlines a competitive edge in the market and improves brand loyalty by allowing them to better handle consumer problems and raise overall customer happiness.

**Limitations:**

One of the analysis's limitations is its dependence on Twitter data, which may not be entirely indicative of the whole landscape of customer mood. Since the dataset only includes tweets regarding airlines, opinions from other platforms or businesses could not be represented. Furthermore, text analysis was used to determine the sentiment labels, which might not always adequately portray the users' complex feelings. The investigation also concentrated on phrase frequency, which offers insightful information but may not adequately convey the sentiment's strength or context. Moreover, even with their remarkable accuracy, the sentiment prediction algorithms could have trouble processing newly discovered data. Finally, this investigation did not consider outside influences on sentiment, such as marketing efforts or current events.