Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6040 – Data Mining Application**

**Assignment:**

Module 2 Final Project - Milestone 1: EDA

**Submitted on:**

April 23rd, 2023

**Submitted to:**  **Submitted by:**

Professor: BEHZAD AHMADI ARCHIT BARUA

BHAGYASHRI KADAM

NIKSHITA RANGANATHAN

**ABSTRACT**

Sentiment analysis is the process of identifying and extracting the emotional tone or attitude expressed in text data, such as social media posts, comments, and reviews. It employs methods – like natural language processing and machine learning.

Social media companies can benefit greatly from sentiment analysis because it helps them to monitor and understand how their brand or products are being perceived by customers and the public. Companies can track changes in sentiment over time, identify trends and patterns in customer feedback, and quickly address any negative comments or complaints by analyzing the sentiment of social media posts and comments related to their brand.

Sentiment analysis can also be used to identify key influencers and advocates for a brand or product and to track how sentiment towards a particular topic or event changes over time. This can help companies to make informed decisions about their marketing and public relations strategies, and to develop more effective communication and engagement strategies with their customers.

Overall, sentiment analysis is an important tool for social media companies to gain insights into how their brand is perceived by customers, and to make data-driven decisions about how to improve customer satisfaction and engagement.

**INTRODUCTION**

**About this Dataset:**

**The Twitter US Airline Sentiment** dataset contains sentiment analysis information about the problems of each major US Airline. It was scraped from Twitter in **February 2015**. It includes tweets about six major US airlines.

The dataset contains **14640 instances and 15 features** which are tweets submitted by individual travelers. Each instance of sentiment is classified as either **positive, neutral, or negative.**

As part of the final project, we are interested in finding the relationship between the Airline sentiment (positive/negative/neutral) and the airline, analyzing the reasons for the negative tweets, and eventually building, training, and testing a model that can predict whether the tweet is positive, negative, or neutral.

**Understanding the dataset:**

This dataset was originally gathered by **Crowdflower's Data for Everyone library** and taken from the Kaggle platform.

Both numerical and categorical data types are present in this dataset. Below are the basic statistics of the Raw dataset:

A picture containing table

Description automatically generated

***Figure 1 – Basic Statistics of the dataset***

Below is the Percentage wise Distribution for Rows, Columns & Observations used in this Dataset:

Chart, bar chart

Description automatically generated

***Figure 2 – Memory Usage***

Below is the Data Structure of the Twitter dataset:

**Chart, radar chart

Description automatically generated**

***Figure 3 – Data Structure***

Below are the data descriptions of each variable of the data that briefly describe the contents of the data set. The dataset's features are as follows:

|  |  |  |
| --- | --- | --- |
| **No** | **Feature** | **Dictionary** |
| 1. | **tweet\_id** | A unique identifier for each tweet. |
| 2. | **airline\_sentiment** | The sentiment expressed in the tweet (positive, negative, or neutral). |
| 3. | **airline\_sentiment\_confidence** | The level of confidence in the sentiment classification (ranging from 0.34 to 1.0). |
| 4. | **negativereason** | The reason for the negative sentiment. |
| 5. | **negativereason\_confidence** | The level of confidence in the negative reason classification (ranging from 0.0 to 1.0). |
| 6. | **airline** | The name of the airline being referenced in the tweet. |
| 7. | **airline\_sentiment\_gold** | A field that is no longer used. |
| 8. | **name** | The Twitter handle of the user who posted the tweet. |
| 9. | **negativereason\_gold** | A field that is no longer used. |
| 10. | **retweet\_count** | The number of times the tweet has been retweeted. |
| 11. | **text** | The content of the tweet. |
| 12. | **tweet\_coord** | The latitude and longitude coordinates of the tweet's location (if available). |
| 13. | **tweet\_created** | The date and time the tweet was posted. |
| 14. | **tweet\_location** | The location of the tweet (if available), as indicated by latitude and longitude. |
| 15. | **user\_timezone** | The timezone of the user who posted the tweet. |

*Table 1: Dictionary with the Features of the Twitter US Airline Sentiment Dataset*

Below are the questions we plan to answer through our Analysis:

**Business Questions -**

1. How do different airlines compare in terms of sentiment distribution? Are there any airlines that consistently perform better or worse than others?
2. When are customers most likely to tweet about airlines? Are there any specific patterns observed in the data?
3. What are the most common reasons for negative sentiment towards different airlines? Are there any areas that need improvement?
4. How does the volume of tweets vary over time and across different airlines, and what factors might be driving these trends?
5. Which locations have the highest number of tweets?
6. What are the most frequently used words in tweets, and do they suggest any common issues that airlines need to address?
7. How does tweet length vary by sentiment and what can airlines learn from this information?

**DATA CLEANING AND MANIPULATION**

* **Checking the number of missing values for each Attribute in the dataset**

**Table

Description automatically generated**

***Figure 4 – Missing data***

From the above missing plot, we can see that there are a lot of missing values present only in the Negative Reason confidence column (around 28%, i.e 4118 records). The other variables/attributes do not seem to have any missing records.

* **Dropping the tweet\_id column**

As we can see from the dataset, column – **twitter\_id**  is not required or significant for our analysis, hence we have dropped it.

* **Replacing Missing values with their Mean - Negative Reason Confidence**

In order to handle the missing values for Negativereason\_confidence, we have imputed the missing values with the respective mean of the column. This approach can help to minimize the loss of information and can be a simple way to handle missing values without introducing any bias.

* **Removing duplicate rows**

Duplicate rows in the dataset can lead to inconsistencies and affect the accuracy of data analysis. It is essential to find and remove any duplicate rows from the dataset before starting the analysis.

In this project, we used the Duplicated () and AnyDuplicated() functions to check for duplicate rows in the Twitter Airline Sentiment dataset. After confirming that there were 36 duplicate values, we removed them. A clean dataset of 14604 records was produced as a result, which was then used for additional analysis.

* **Extracting hour values from time after separating date and time columns**

In order to analyze tweet data more effectively, the date and time information was separated into two distinct columns in the dataset. This was achieved using the "separate" function in R, which split the original "tweet\_created" column into two columns. Subsequently, the "hour" value was extracted from the "time" column using the "hour" function from the "hms" package.

* **Replacing column values**

The "Hours" column values were replaced with categories of Morning, Afternoon, Evening, and Night using the "mutate" and "recode" functions. This transformation allows for easier interpretation and analysis of time-related data in the dataset.

* **Creating a day column**

A 'day' column was added to the dataset and its values ​​were derived from the 'date' column. Dates from '2015-02-16' to '2015-02-24' have been replaced with the corresponding day of the week.

* **Changing datatypes**

The columns (airline sentiment, airline, day, and hour) were converted from one data type to factor using as.factor().

**DESCRIPTIVE CHARACTERISTICS OF THE DATASET**

Table

Description automatically generated with medium confidence

***Figure 5 – Descriptive statistics***

The table includes various descriptive statistics for each variable, such as the number of observations (n), mean, standard deviation (sd), median, trimmed mean, median absolute deviation (mad), minimum, maximum, range, skewness, kurtosis, and standard error of the mean (se).

While the others are continuous variables, the variables marked with an asterisk (\*) are categorical variables.

Some key descriptive statistics for the dataset are:

* The mean sentiment score is 1.53, indicating that the majority of tweets in the dataset are negative. The median sentiment score is 1.00, which is also the score for a neutral tweet, indicating that there are a significant number of neutral tweets in the dataset as well. The sentiment score's standard deviation, which is relatively high at 0.76, shows that the sentiment scores range widely.
* The confidence level of the sentiment classification is quite high, with a mean of 0.9 and a standard deviation of 0.16.
* The retweet count is low, with an average of 0.08.
* The text of the tweet is on average 7,168 characters long, and there are few tweets with coordinates and locations.

Chart, treemap chart

Description automatically generated

***Figure 6 – Correlation Plot***

The correlation plot illustrates that the variables "airline\_sentiment\_confidence" and "negativereason\_confidence" have a strong positive correlation (0.603), which means that as one variable rises, the other tends to follow suit.

On the other hand, the relationship between "airline\_sentiment\_confidence" and "retweet\_count", as well as between "negativereason\_confidence" and "retweet\_count", is weakly positive with coefficients of 0.013 and 0.019, respectively. This suggests that there is no strong association between these variables.

**EXPLORATORY DATA ANALYSIS (EDA)**

**Chart, histogram

Description automatically generated**

***Figure 7A – Histogram 1***

The above histogram shows the distribution of Airline sentiment confidence. We can observe airline sentiment confidence does not follow any positive or specific distribution as the sentiment confidence values are scattered between 0.6 to 0.75. Additionally, there are few values observed which are between 0.3 to 0.4. Most of the values with a count greater than 10,000 are in the range of 1.0.

**Chart, bar chart

Description automatically generated**

***Figure 7B – Histogram 2***

The distribution of Negative Reason Confidence is shown by the histogram in Figure 7B. We can observe negative reason confidence does not follow any positive or specific distribution as the negative reason confidence values are scattered majorly between 0.6 to 0.75 with a count greater than 6000. Additionally, there are a few values observed which are between 0.25 to 0.4 and with 1.0. The least distribution can be seen in the range of 0.0 to 0.1.

**Chart, histogram

Description automatically generated**

***Figure 7C – Histogram 3***

The above histogram shows the distribution of the Retweet Count for the tweets done. We can observe that the retweet count does not follow any positive or specific distribution as the retweet count values are majorly between 0.0 to 0.5 with a count greater than 10,000.

**Chart, scatter chart

Description automatically generated**

***Figure 8 – Scatterplot***

The above line dot plot is an interactive plot that explains the daily cumulative tweets received from the date 16th Feb to 24th of Feb. From the above plot, we can understand that the overall daily cumulative count shows an upward trend for the given period. Starting with 16th feb, the daily cumulative count was observed less than 5, however until the 20th of Feb, the daily cumulative count was almost 5x and accounted to 5632.

Finally, on the 24th of feb, the daily cumulative count was observed to be 14604.

**Chart, histogram

Description automatically generated**

***Figure 9 – Multiple density plots***

The above density graph/plot represents the distribution of negative tweets that were received for each of the Major US airlines. The plot consists of several density curves ​​and provides insight into the distribution of negative tweets by the airline, with different airlines having different peaks and shapes of density distributions.

The x-axis represents the count for negative tweets and the y-axis represents the density of the negative tweets for each airline. From the graph, we can understand that American Airlines shows a flatter distribution with less pronounced peaks suggesting that negative tweets for this airline are more evenly distributed across counts. Virgin America has the highest concentration of negative tweets with the count range of 0-100.

Further, Delta & Southwest Airline have density curves almost over-lapping with less concentration of negative tweets as compared to American airlines and a count range of 80-180.

Lastly, United & US Airways have less peaks suggested that the negative tweets for these airlines are evenly distributed across the counts up to 600.

**Chart, radar chart

Description automatically generated**

***Figure 10 – Circular barplot***

The above circular bar chart represents the highest number of tweets received from the top 10 locations in the USA as per count. The bars are colored based on the tweet location and the length of the bars represents the tweet count for each location. The circular layout of the bars creates a visually appealing display, and the use of polar coordinate system can be useful for comparing the values of each location.

From the above chart, we can understand that the highest number of tweets are from New York (638), followed by Washington, D.C (335) & New York (213). The least number of tweets have been received from Houston, TX (84) & Dallas, TX (97).

Chart

Description automatically generated

***Figure 11 – Animated line graph***

The above-animated line graph represents the timeline of daily tweets that were received for each US Airline from the period Feb 16th, Feb 24th, 2015.

The graph displays a line for each airline, with the date on the x-axis and the number of tweets on the y-axis.

The animation feature of the graph reveals the daily tweet counts for each airline over time, allowing us to see how the tweet volume for each airline changes over time. This can be helpful in identifying patterns or trends in the data, such as spikes or dips in tweet volume, as well as any seasonal or cyclical trends.

We can understand the below for each Airline -

**American**: The line for American is orange, and it starts out with a low number of tweets on Feb 18th. The tweet volume gradually increases showing the largest spike ever across all airlines on Feb 23rd.

**Delta**: The line for Delta is yellow, and it starts out with a moderate number of tweets on Feb 16th. The tweet volume for Southwest also gradually increases on Feb 22nd with a few spikes and dips again along the way.

**Southwest**: The line for Southwest is green, and it starts out with a higher number of tweets after United on Feb 17th. The tweet volume for Southwest shows certain spikes and dips along the way and gradually ends with the 2nd lowest dip on 23rd Feb.

**United**: The line for United is sky blue, and it starts out with a very low number of tweets on Feb 16th. The tweet volume for United gradually increases showing the 2nd largest spike ever across all airlines on Feb 23rd.

**US Airways**: The line for US Airways is cobalt blue, and it starts out with a moderate number of tweets on Feb17th showing few spikes and dips along the way with its highest daily peak on Feb 22nd and gradually showing a dip on Feb24th.

**Virgin America**: The line for Virgin America is purple, and it starts out with the 2nd lowest number of tweets on Feb17th. The tweet volume for Virgin America is mostly flat throughout the rest of the dates, with a few small spikes and dips and ending with the lowest no of tweets on 24th Feb.

Chart, treemap chart

Description automatically generated

***Figure 12 – Treemap***

This above treemap visualization represents the negative reasons for tweets about airlines. Each negative reason is represented by a rectangle of varying sizes on the treemap, and the size of the rectangle is proportional to the number of tweets associated with that negative reason.

The plot shows that the most common negative reason for tweets about US airlines is "Customer Service Issues," followed by "Late Flight" and "Can't Tell," which indicates that customers are dissatisfied with the airlines' customer service and flight punctuality. Other negative reasons include "Cancelled Flight", "Lost Luggage", "Bad Flight", "Flight Booking Problems", "Flight Attendant Complaints", and "longlines".

Chart, pie chart

Description automatically generated

***Figure 13 – Multiple pie charts***

The pie charts give the distribution of sentiments (positive, negative, and neutral) across six major airlines in the US. This is a great and simple way to compare the results and draw conclusions. It can be seen that all the airlines have the majority of tweets expressing negative sentiment, with US Airways and American Airlines having the highest proportion of negative tweets at over 70%.

These findings suggest that airlines need to address the underlying issues causing negative sentiment on Twitter to improve their reputation and customer satisfaction.

**Chart, bar chart

Description automatically generated**

***Figure 14 – Stacked bar graph***

This column chart represents the sentiment analysis of tweets about different airlines. The chart is grouped by airline and sentiment, and the count of tweets for each sentiment is shown using a stacked bar chart. The x-axis represents the different airlines, and the y-axis represents the total count of tweets.

We can see that United Airlines has the highest number of total tweets, followed by US Airways and American Airlines. However, when we look at the sentiment breakdown, we see that most tweets for these three airlines are negative, with more than 50% of tweets having a negative sentiment.

Furthermore, United Airlines has the highest number of negative tweets among all the airlines, with 2633 negative tweets. This suggests that United Airlines has a higher volume of negative sentiment on Twitter compared to other airlines.

Overall, this graph highlights the importance of sentiment analysis for businesses, particularly in the airline industry, as it provides insights into how customers feel about their products and services.

Chart, bar chart

Description automatically generated

***Figure 15 – Stacked bar graph 1***

From the output, we can see that the highest number of tweets for any day of the week is 3079 tweets on Sunday. The next highest number of tweets is on Monday, with 3032 tweets. The lowest number of tweets is on Wednesday, with only 1344 tweets.

This information can be useful for analyzing patterns and trends in the data, such as which days of the week have the highest or lowest engagement on social media.

Additionally, we can see that the highest number of tweets for any day and hour of the day is 1465 tweets on Tuesday morning. The next highest number of tweets is on Sunday afternoon, with 1410 tweets. The lowest number of tweets is on Wednesday night, with only 138 tweets.

This information can be useful for analyzing patterns and trends in the data, such as which times of the day and which days of the week have the highest or lowest engagement on social media. This can help you determine the best times to post content and engage with your followers.

Chart, bar chart

Description automatically generated

***Figure 16 – Stacked bar graph 2***

The graph shows the counts of negative reasons for tweets about airlines, grouped by the specific airlines. The x-axis shows the different negative reasons, and the y-axis shows the count of tweets.

From the graph, we can see that the most common negative reason for tweets is "Customer Service Issue", with 2904 tweets. The second most common negative reason is "Late Flight", with 1660 tweets, followed by "Cancelled Flight" with 1190 tweets, and "Can’t Tell" with 843 tweets. The least common negative reason for tweets is "Damaged Luggage", with only 74 tweets.

We can see that different airlines have different areas of opportunity for improvement. US Airways has the highest number of tweets related to customer service issues, with 811 tweets, while American Airlines follows closely with 762 tweets. United Airlines had the highest number of tweets related to late flights, with 525 tweets, followed by US Airways with 453 tweets. Additionally, American Airlines had the highest number of tweets related to canceled flights, with 242 tweets, followed by US Airways with 189 tweets.

Airlines can improve their operations or customer service in these areas by analyzing the results and determining where they need to make changes. The airline's reputation and financial performance may ultimately benefit from higher customer satisfaction and loyalty as a result.

Text

Description automatically generatedChart, bar chart

Description automatically generated

***Figure 17 – Wordcloud and Barplot***

We used word cloud and barplot to understand the most frequently used words in tweets related to airlines.

In the word cloud, larger words indicate a higher frequency of use. On the other hand, a bar chart is interactive, and hovering over the bars shows the exact count of the words in the tweets.

The results show that the most commonly used words in tweets are "flight" followed by "united," "usairways," and "americanair." This indicates that people are primarily talking about their flight experiences and mentioning the airline they flew with.

Both plots can provide valuable insights for airlines to understand the topics that are discussed by their customers on social media, which in turn can help airlines improve their services and address any issues that their customers are facing.

Chart

Description automatically generated

***Figure 18 – Ridge Chart***

The Ridge chart displays the density distribution of tweet text length across the three sentiment categories: negative, neutral, and positive. The x-axis represents the text length, and the y-axis represents the sentiment category.

The chart highlights the relationship between tweet text length and sentiment. It also suggests that negative sentiment tweets are typically longer than neutral and positive tweets, and people tend to express negative experiences with more details and explanations.

**TEXT PRE-PROCESSING**

In this section of the assignment, we have preprocessed a collection of tweets (related to airline customer experience). This includes cleaning, formatting, and transforming raw text data into a format that is suitable for analysis. The goal of text preprocessing is to prepare the text data for analysis, making it easier for ML algorithms to learn from the data and provide accurate results.

We begin by converting the tweets to UTF-8 encoding and creating a Corpus object from the text data. We then use various functions from the "tm" package to clean and normalize the text data. We convert all text to lowercase, remove punctuation, remove numbers, remove stop words, remove airline-specific terms such as flight and plane. Extra white space is also removed from the text, and each word is reduced to its base form using stemming. This is done to reduce the number of unique words and to group similar words together.

Chart, histogram

Description automatically generatedWe then created a TermDocumentMatrix from the pre-processed text and removed sparse terms from it. Sparse terms refer to terms in the matrix that occur with a very low frequency. Removing sparse terms from the matrix can help reduce its size and improve the efficiency of analysis.

***Figure 19 – Most used words in tweets***

Next, we wanted to know which words occur the most in the tweets. Therefore, we calculated the frequency of each word and selected words with a frequency greater than 500. We visualized the results by sorting the selected words and creating a bar chart. The resulting plot can provide insights into the topics or themes being discussed and can be useful for businesses to understand customer sentiment or for researchers to analyze public opinions and trends.

Text

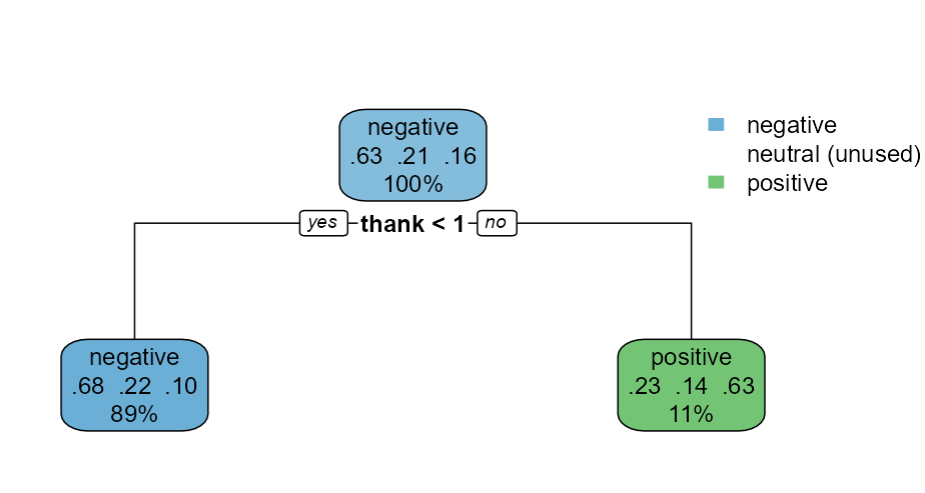
Description automatically generated

***Figure 20 – Baseline accuracy of test and train set***

Finally, the data is split into training and testing sets with 80:20 ratio (meaning that 80% of the data is used for training the model and 20% is used for testing the model's performance), and baseline accuracies for both sets are computed. Data splitting is an important step in machine learning to evaluate how well a model performs on new data.

The baseline accuracy is the proportion of the most frequent class in the dataset. In this case, the majority class is "negative" with a proportion of approximately 0.63 in the training set and 0.62 in the testing set.

**DECISION TREE**

****

***Figure 21 – Classification tree***

The decision tree is a powerful machine learning algorithm that is used for both classification and regression tasks. Decision tree is a tree-structured model that divides the dataset into smaller and smaller subsets based on the values of selected features, and then it predicts the target variable by analyzing the values of those features. It is also useful for handling both numerical and categorical data.

Text, letter

Description automatically generated

***Figure 22 – Summary of decision tree***

* The CP is a tuning parameter that controls the size of the tree, and a smaller CP value results in a larger tree with more splits. In this case, the CP value of 0.1171357 resulted in two splits, with a final tree containing two terminal nodes.
* The "thank" variable has an importance of 99 and the highest "improve" value, indicating that it was the most important variable in predicting sentiment. The tree has two levels, with the root node splitting on the feature "thank".
* Node 1 represents the root node, and it has 11,756 observations. The predicted class for this node is negative. The node is split into two child nodes: Node 2 and Node 3.
* Node 2 represents the left child node of Node 1, and it has 10,446 observations. The predicted class for this node is also negative. On the other hand, Node 3 represents the right child node of Node 1, and it has 1,310 observations. The predicted class for this node is positive.
* The other variables - primary splits (hour, great, delay, usairway) and surrogate splits (twitter, prompt, safe, southwestair) can also be observed in the outcome. However, their influence on the classification (improve value) is lower than that of "thank".

A picture containing text

Description automatically generated

***Figure 23 – Accuracy of decision tree***

The model correctly predicted the sentiment of around 67% of the tweets in the test set.

In a classification problem, it is important for a model to be able to accurately classify all classes present in the dataset. In this case, the model is unable to provide any classification for the neutral class, which means it is not able to distinguish between neutral tweets and tweets with a positive or negative sentiment. This is a major limitation as the model fails to provide a complete picture of the sentiment analysis task.

**RANDOM FOREST**

|  |
| --- |
| After the implementation of decision tree, we have applied the Random Forest model to predict the airline sentiment. We have built the Random Forest model using the Train set and 20 trees so that we can run the model in limited time.  Table  Description automatically generated  ***Figure 24 – Random forest model***  From the above output, we can observe the following –   * + - * For each split in each tree, a random subset of 33 variables were considered for splitting.       * The estimated out-of-bag (OOB) error rate is 26.2%       * From the confusion matrix, we can see that the model was able to correctly predict the 6935 Negative tweets, 1151 neutral tweets and 1069 positive tweets. While misclassifying 651 negative tweets, 1077 neutral tweets and 500 positive tweets.       * Lastly, from the class.error metric shows us the misclassification error rate which is highest for the neutral tweets – 53.73%   Further, we have made predictions on both the Training & Test data using the above random forest model and evaluated their accuracy respectively.  Graphical user interface, text, application  Description automatically generated  Graphical user interface, text, application  Description automatically generated  ***Figure 25 – Accuracy of test and train set (Random forest)***  Finally, we have evaluated the Model’s performance using the confusion matrix.    ***Figure 26 – Confusion matrix***  From the above results, we can say that the Random Forest Model has an accuracy of 74% which tells us that the model correctly predicted the sentiment of 74.35% of the tweets in the test set.  Further, we can see that the model correctly predicted 1588 negative tweets (true negatives) and 277 positive tweets (true positives). However, the model did not perform as well on the neutral tweets, with only 266 (true positives) being correctly predicted.  The model incorrectly predicted 264 instances of negative sentiment, 140 instances of neutral sentiment, and 75 instances of positive sentiment.  Lastly, for this model, the sensitivity values for the negative, neutral, and positive classes are 0.8808, 0.45470, and 0.57950, respectively. This means that the model can correctly identify 88.08% of negative cases, 45.47% of neutral cases, and 57.95% of positive cases.  The specificity values for the negative, neutral, and positive classes are 0.6275, 0.90837, and 0.94556, respectively. This means that the model can correctly identify 62.75% of negative cases, 90.84% of neutral cases, and 94.56% of positive cases. |

**CONCLUSION**

In conclusion, sentiment analysis is a valuable tool for social media companies to monitor and understand how their brand or products are being perceived by customers and the public. The analysis of the Twitter US Airline Sentiment dataset has provided insights into the sentiment distribution of different airlines, the most common negative reasons for tweets, and the most frequently used words in tweets related to airlines. These insights can help airlines improve their services, address customer issues, and enhance the customer experience. Additionally, analyzing the volume of tweets over time and across different airlines can provide valuable information for businesses to identify patterns and trends and determine the best times to engage with their followers. Overall, sentiment analysis can help social media companies make data-driven decisions about how to improve customer satisfaction and engagement, ultimately leading to increased customer loyalty and improved business outcomes.

* The dataset contains 14,640 instances of tweets about US airlines.
* Among these tweets, 62.7% were negative, 21.2% were neutral, and 16.1% were positive.
* United Airlines had the highest number of total tweets, with 3,032 tweets, followed by US Airways and American Airlines.
* However, United Airlines also had the highest number of negative tweets, with 2,633 negative tweets.
* The most common negative reason for tweets was "Customer Service Issue," with 2,904 tweets, followed by "Late Flight" with 1,660 tweets.
* The word "flight" was the most frequently used word in tweets related to airlines, with a count of 5,904, followed by "united" with a count of 3,263.
* The Ridge chart showed that the average text length for negative tweets was 118 characters, compared to 96 characters for neutral tweets and 98 characters for positive tweets.

**FUTURE ANALYSIS**

Data mining techniques such as prediction and classification, clustering and association mining, linear discriminant analysis, support vector machines, and text mining can be applied to the Twitter US Airline Sentiment dataset to gain deeper insights and answer more complex questions.

* Prediction and Classification Techniques (e.g., Decision Trees, Logistic Regression): These techniques can be used to predict the sentiment of a tweet based on various features such as airline, location, time of day, and keywords. This can help companies identify the variables that are most strongly linked to either positive or negative sentiment and create strategies to enhance the customer experience.
* Clustering and Association Mining: These techniques can be used to identify patterns and trends in the data that may not be immediately apparent, such as common groups of keywords or customer demographics that are associated with certain sentiment categories. This can help businesses develop targeted marketing and engagement strategies.
* Linear Discriminant Analysis: This method can be used to determine the key characteristics or aspects that are connected to either positive or negative feelings. This can assist companies in focusing their efforts on the most crucial areas for development.
* Support Vector Machines: This technique can be used to classify tweets into sentiment categories based on various features such as keywords and time of day. Businesses may benefit from automating the sentiment classification process in this way to save time and money.
* Text Mining: This technique can be used to analyze the text of the tweets to identify common themes, topics, and sentiments. This can help businesses understand the specific issues that customers are facing and develop targeted strategies to address them.

|  |
| --- |
| **APPENDIX** |

|  |
| --- |
| # Load Packages & Libraries library(tidyverse) library(caret) library(highcharter) library(RColorBrewer) library(ggcorrplot) library(corrplot) library(psych) library(dplyr) library(ggplot2) library(gtools) library(ggfortify) library(GGally) library(readr) library(readxl) library(knitr) library(modelr) library(scales) library(sqldf) library(car) library(ggpubr) library(grid) library(gridExtra) library(lattice) library(hrbrthemes) library(ISLR) library(caret) library(pROC) library(psych) library(olsrr) library(naniar) library(DataExplorer) library(formattable) library(glmnet) library(Metrics) library(MLmetrics) library(GGally) require(grid) library(lubridate) library(gganimate) library(babynames) library(hrbrthemes) library(gifski) library(plotly) library(leaflet) library(ochRe) library(tidytext) require(grid) library(treemapify) library(ggridges) library(tidyr) library(plotly) library(viridis) library(hrbrthemes) library(gganimate) library(wordcloud) library(magrittr)  # Importing the dataset  Tweets.data <- read.csv("Tweets.csv",header=TRUE,sep=",")  # Understanding the dataset str(Tweets.data) summary(Tweets.data) glimpse(Tweets.data) headTail(Tweets.data)  # Data Pre-Processing & Data Cleaning  # Checking data is clean? colSums(is.na(Tweets.data)) # check & Returns the number of missing values in each column sum(is.na(Tweets.data)) # Counts missing values in entire data frame colSums(Tweets.data==0) #Using colSums function to find the total number of Zero records in each column plot\_missing(Tweets.data, title="Missing Data Profile",geom\_label\_args = list("size" = 2, "label.padding" = unit(0.1, "lines")))  # Dropping tweet\_id column (numeric) drop1<-c("tweet\_id") Tweets.data<-Tweets.data[,!(names(Tweets.data) %in% drop1)]  # Replacing NA values with the mean of the column Tweets.data$negativereason\_confidence[is.na(Tweets.data$negativereason\_confidence)]<-round (mean(Tweets.data$negativereason\_confidence,na.rm=TRUE),2)  # Checking for duplicated rows and removing them duplicated(Tweets.data) anyDuplicated(Tweets.data) Tweets.data<-Tweets.data[!duplicated(Tweets.data), ]  # Seperating date and time into two columns Tweets.data<-Tweets.data %>%  separate(tweet\_created, c("date", "time"), " ")  # Hour from time Tweets.data$hour <- hour(hms(Tweets.data$time))  # Replacing hours to Morning, Afternoon,Evening and Night # 0-5 night # 6-11 morning # 12-17 afternoon # 18-24 evening Tweets.data <- Tweets.data %>%   mutate(hour = ifelse(hour %in% c(0:5), "Night",  ifelse(hour %in% c(6:11), "Morning",  ifelse(hour %in% c(12:17), "Afternoon", "Evening"))))  # Creating day column Tweets.data$day<-Tweets.data$date Tweets.data$day[Tweets.data$date == '2015-02-16'] <- 'Monday' Tweets.data$day[Tweets.data$date == '2015-02-17'] <- 'Tuesday' Tweets.data$day[Tweets.data$date == '2015-02-18'] <- 'Wednesday' Tweets.data$day[Tweets.data$date == '2015-02-19'] <- 'Thursday' Tweets.data$day[Tweets.data$date == '2015-02-20'] <- 'Friday' Tweets.data$day[Tweets.data$date == '2015-02-21'] <- 'Saturday' Tweets.data$day[Tweets.data$date == '2015-02-22'] <- 'Sunday' Tweets.data$day[Tweets.data$date == '2015-02-23'] <- 'Monday' Tweets.data$day[Tweets.data$date == '2015-02-24'] <- 'Tuesday'  # Relocating the columns Tweets.data<-Tweets.data %>% relocate(day,.after = date) Tweets.data<-Tweets.data %>% relocate(hour,.after = time)  # Changing the datatypes Tweets.data$airline\_sentiment<-as.factor(Tweets.data$airline\_sentiment) Tweets.data$airline<-as.factor(Tweets.data$airline) Tweets.data$day<-as.factor(Tweets.data$day) Tweets.data$hour<-as.factor(Tweets.data$hour) str(Tweets.data)  # Descriptive Statistics for entire dataset formattable(describe(Tweets.data),   caption = "Descriptive statistics summary of the Twitter US Airline Sentiment Dataset")  # Histogram Distribution  # Airline Sentiment Confidence ggplot(Tweets.data, aes(x=airline\_sentiment\_confidence)) +   geom\_histogram(color="black", fill="orange", position="identity")+  labs(title="Airline Sentiment Confidence Distribution",x="Airline Sentiment Confidence", y = "Count")  # Negative Reason Confidence ggplot(Tweets.data, aes(x=negativereason\_confidence)) +   geom\_histogram(color="black", fill="orange", position="identity")+  labs(title="Negative Reason Confidence Distribution",x="Negative Reason Confidence", y = "Count")  # Retweet Count ggplot(Tweets.data, aes(x=retweet\_count)) +   geom\_histogram(color="black", fill="orange", position="identity")+  labs(title="Retweet Count Distribution",x="Retweet Count", y = "Count")  # Correlation Matrix corr<-Tweets.data %>% select(airline\_sentiment\_confidence,negativereason\_confidence,retweet\_count) corr\_matrix <- cor(corr) corr\_matrix  ggcorrplot(corr\_matrix, hc.order = TRUE,   type = "lower",   lab = TRUE,   lab\_size = 3.5,   colors = c("#6D9EC1", "white", "#E46726"),   title = "Correlation Plot of Selected Variables")  # Data Profiling Report create\_report(Tweets.data)  # Data Visualizations # Graph 1 - Sentiment analysis by airline df1<-Tweets.data %>% group\_by(airline, airline\_sentiment) %>% summarize(count=n())  In2 <- hchart(df1, 'column',  hcaes(x = 'airline', y = 'count', group = 'airline\_sentiment'),  stacking = "normal") %>%  hc\_colors(c("#CD0000", "#EEE8CD", "#698B69")) %>%  hc\_xAxis(title = "Airline") %>%  hc\_yAxis(title = "Count") %>%  hc\_title(text = "Sentiment Analysis by Airline") In2  # Graph 2 interactive - Tweet Count by Day and Hour df2 <- Tweets.data %>% group\_by(day,hour) %>% summarize(count=n())  In2 <- hchart(df2, 'column',  hcaes(x = 'day', y = 'count', group = 'hour')) %>%  hc\_colors(c("#B0E2FF", "#FF8247", "#FDE725", "#00008B")) %>%  hc\_xAxis(title = "Day") %>%  hc\_yAxis(title = "Count") %>%  hc\_title(text = "Tweet Count by Day and Hour") In2  # Graph 3 interactive - Negative Reason Counts by Airline df3 <- Tweets.data %>%  filter(!is.na(negativereason) & negativereason != "") %>%  group\_by(negativereason, airline) %>%  summarize(count = n()) %>%  filter(count > 0) %>%  arrange(desc(count))  In3 <- hchart(df3, 'bar', hcaes(x = negativereason, y = count, group = airline)) %>%  hc\_colors(c("#1874CD", "#B22222", "#FFD700", "#00BFFF","#8B8682","#FF0000")) %>%  hc\_xAxis(title = "Negative Reasons") %>%  hc\_yAxis(title = "Count") %>%  hc\_title(text = "Negative Reason Counts by Airline") In3  # Graph 4 - Timeline of Daily Cumulative Tweets  dailyTweetscum <- Tweets.data %>% group\_by(date) %>% dplyr::summarise(count = n()) %>%  mutate(cuml = cumsum(count))  p1<- ggplot(data=dailyTweetscum , aes(x = date, y = cuml)) +  geom\_point(size=2.0 ,color = "skyblue")+  theme\_classic()+  ggtitle('Daily Cumulative Tweets') +  scale\_y\_continuous(labels = scales::comma, breaks = scales::pretty\_breaks(n=7)) +  theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))+  theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(),  text=element\_text(size=12, family="Comic Sans MS", color= "black"))  ggplotly(p1)  # Graph 5 - Density graph of Negative tweets  negativeTweets <- Tweets.data %>% filter(airline\_sentiment=="negative") negativeTweets <- negativeTweets %>% group\_by(airline , date) %>% dplyr::summarise(count = n())   p2 <- ggplot(negativeTweets, aes(x = count, fill = airline)) +  geom\_density(alpha = 0.5) +  ggtitle("Distribution of Negative Tweets by Airline") +  theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1)) p2  # Graph 6 - Timeline of Daily Tweets for Each Airline dailyTweets <- Tweets.data %>% group\_by(airline,date) %>% dplyr::summarise(count = n()) dailyTweets$date <- as.Date(dailyTweets$date)  p3 <-dailyTweets %>% ggplot(aes(x = date, y = count,   group = airline,  color = airline)) +  theme\_bw()+  geom\_line() +  geom\_point() +  ggtitle("Daily Tweets per Airline") +  theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1)) +  transition\_reveal(date)  animate(p3, renderer = gifski\_renderer()) rstudioapi::viewer("http://localhost:24784/session/filecd4843c75819.gif")  anim\_save(filename = "Animated.gif")  # Graph 7 - Location Wise Tweets - Circular Bar chart location <- Tweets.data %>% group\_by(tweet\_location) %>%  dplyr::summarise(count=n()) %>% arrange(desc(count)) %>% filter(!is.na(tweet\_location) & tweet\_location != "") %>% top\_n(10) location  ggplot(location, aes(tweet\_location, count, fill = tweet\_location)) +  geom\_col(position = "dodge") +  coord\_polar() +  geom\_text(aes(label = count), vjust = -0.5, size = 3) +  scale\_fill\_viridis\_d() +  ggtitle("Circular Bar chart- Top 10 Location Wise Tweets") +  xlab("Location") +  theme(axis.text.x = element\_text(hjust = 1))  # Graph 8 - Treemap of Negative reasons  data = Tweets.data %>% group\_by(negativereason) %>% summarise(count = n())%>% filter(!is.na(negativereason) & negativereason != "")  ggplot(data, aes(area = count, fill = negativereason, label = negativereason)) +  geom\_treemap() +  scale\_fill\_brewer(name = "Negative Reasons",palette = "Set3") +  geom\_treemap\_text(family = "Comic Sans MS", colour = "black", place = "centre", reflow = T) +  theme\_void() +  ggtitle("Treemap - Negative reason") +  theme(plot.title = element\_text(hjust = 0.5, size = 18, family = "Comic Sans MS", color = "black")) +  guides(fill = FALSE) +  coord\_fixed(ratio = 0.75)  # Graph 9 - Pie charts with Distribution of Sentiments by Airline sentiment\_summary <- Tweets.data %>%  group\_by(airline, airline\_sentiment) %>%  summarise(n = n()) %>%  mutate(percent = n / sum(n) \* 100) pie\_charts <- ggplot(sentiment\_summary, aes(x = "", y = percent, fill = airline\_sentiment)) +  geom\_bar(stat = "identity", width = 1, color = "white") +  coord\_polar(theta = "y") +  facet\_wrap(~ airline, nrow = 2) +  theme\_void() +  labs(x = NULL, y = NULL, fill = NULL, title = "Distribution of Sentiments by Airline") +  theme(plot.title = element\_text(hjust = 0.5, size = 18, family = "Comic Sans MS", color = "black"),  strip.text = element\_text(size = 16, family = "Comic Sans MS", color = "black"),  text = element\_text(family = "Comic Sans MS", color = "black")) +  scale\_fill\_manual(name = "Sentiment",  values = c("#CD0000", "#EEE8CD", "#698B69"),  labels = c("Negative", "Neutral","Positive")) +  geom\_text(aes(label = paste0(round(percent), "%")), position = position\_stack(vjust = 0.5))  print(pie\_charts)  # Graph 10 - Wordcloud of words used in tweets Tweets.data %>%  select(text) %>%  unnest\_tokens(word, text) %>%  dplyr::count(word, name = "n") %>%  arrange(desc(n)) %>%  with(wordcloud(word, n, max.words = 100, colors = brewer.pal(8, "Dark2")))  # Graph 11 - Top 20 words in Tweets colour <- c(  "#E31A1C",  "green4",  "#6A3D9A",   "#FF7F00", "gold1",  "skyblue2","dodgerblue2", "#FB9A99",  "palegreen2",  "#CAB2D6",   "#FDBF6F",   "khaki2",  "maroon", "green1", "orchid1","deeppink1", "steelblue4",  "darkturquoise", "yellow3",  "brown" )  bar <- Tweets.data %>%  unite(text\_combined, text, sep = " ") %>%  select(text\_combined) %>%  unnest\_tokens(word, text\_combined) %>%  filter(!grepl("^[0-9]+$", word)) %>%  dplyr::count(word, sort = TRUE) %>%  top\_n(20) %>%  ggplot(aes(x = word, y = n)) +  geom\_col(color = "black", fill = colour) +  coord\_flip() +  labs(x = "Words", y = "Frequency", title = "Top 20 Most Frequent Words in Tweets") +  theme\_bw() +  theme(text = element\_text(family = "Comic Sans MS", color = "black"))  ggplotly(bar)  # Graph 12 - Ridge chart - Tweet Text Length by Sentiment Tweets.data$text\_length <- nchar(Tweets.data$text) Tweets.data %>%  ggplot(aes(x = text\_length, y = airline\_sentiment, fill = airline\_sentiment)) +  geom\_density\_ridges(scale = 3, rel\_min\_height = 0.01,alpha=0.5) +  scale\_fill\_manual(name = "Sentiment",labels = c("Negative", "Neutral","Positive"),values = c("#CD0000", "#EEE8CD", "#698B69")) +  labs(x = "Text Length", y = "Sentiment", title = "Tweet Text Length by Sentiment") +  theme\_minimal() +  theme(text = element\_text(family = "Comic Sans MS", color = "black"))  # Text Pre-processing  # Convert Tweets data to UTF-8 encoding  tweets\_utf8 <- iconv(Tweets.data$text, to = "UTF-8")  # Create a Corpus object from the text data  corpus\_text <- Corpus(VectorSource(tweets\_utf8))  # Inspect the first five documents in the Corpus object  inspect(corpus\_text[1:5])  # Convert all text to lowercase  corpus\_lc <- tm\_map(corpus\_text, tolower)  # Remove all punctuation from the text  corpus\_np <- tm\_map(corpus\_lc, removePunctuation)  # Remove all numbers from the text  corpus\_nn <- tm\_map(corpus\_np, removeNumbers)  # Remove common English words from the text  corpus\_sw <- tm\_map(corpus\_nn, removeWords, stopwords('english'))  # Remove additional words such as airline-specific terms from the text  corpus\_asw <- tm\_map(corpus\_sw, removeWords, c('flight','get','plane','flights','flightl','i', 'day', 'im', 'cant', 'can', 'now', 'just','will', 'dont', 'ive', 'got', 'much'))  # Remove any extra white space from the text  corpus\_ws <- tm\_map(corpus\_asw, stripWhitespace)  # Reduce each word to its base form  corpus\_stem <- tm\_map(corpus\_ws, stemDocument)  # Create a TermDocumentMatrix from the pre-processed text  tdm <- TermDocumentMatrix(corpus\_stem)  # Remove sparse terms from the TermDocumentMatrix  tdm\_clean <- removeSparseTerms(tdm, sparse = 0.999)  # Convert the resulting matrix to a data frame  tdm\_df <- as.data.frame(as.matrix(tdm\_clean))  # Compute the frequency of each word in the Corpus object  word\_freq <- rowSums(tdm\_df)  # Select words with a frequency of more than 500  freq\_words <- subset(word\_freq, word\_freq >= 500)  # Sort the selected words in decreasing order  sorted\_freq\_words <- sort(freq\_words, decreasing = T)  # Create a bar plot of the selected words with their frequency  barplot(sorted\_freq\_words, col = rainbow(38,alpha=0.7), las = 2,  main = "Words with frequency more than 500",  xlab = "Words", ylab = "Frequency")  # Create a DocumentTermMatrix from the pre-processed text  dtm <- DocumentTermMatrix(corpus\_stem)  # Remove sparse terms from the DocumentTermMatrix  dtm\_clean <- removeSparseTerms(dtm, 0.999)  # Convert the resulting matrix to a data frame  dtm\_df <- as.data.frame(as.matrix(dtm\_clean))  # Add the airline sentiment labels as a factor variable to the data frame  dtm\_df$sentiment <- Tweets.data$airline\_sentiment  # Rename any duplicate variables in the training set  colnames(dtm\_df)[duplicated(colnames(dtm\_df))] <- paste0(colnames(dtm\_df)[duplicated(colnames(dtm\_df))], "\_2")  # Convert the airline sentiment labels to a factor variable  dtm\_df$sentiment <- as.factor(dtm\_df$sentiment)  # Split the data into a training set and a testing set  set.seed(222)  split <- sample(2, nrow(dtm\_df), prob = c(0.8,0.2), replace = TRUE)  train\_set <- dtm\_df[split == 1,]  test\_set <- dtm\_df[split == 2,]  # Compute the baseline accuracy for the training set  train\_set\_baseline\_acc <- prop.table(table(train\_set$sentiment))  # Compute the baseline accuracy for the testing set  test\_set\_baseline\_acc <- prop.table(table(test\_set$sentiment))  # Display the baseline accuracies for both sets  train\_set\_baseline\_acc  test\_set\_baseline\_acc  # Removing Duplicacy  # Display column names of train\_set data frame  names(train\_set)  # Rename any duplicate variables to ensure unique names  colnames(train\_set)[duplicated(colnames(train\_set))] <- paste0(colnames(train\_set)[duplicated(colnames(train\_set))], "\_2")  names(test\_set) <- names(train\_set)  # Decision tree  # Fit a decision tree classifier using the training set  dt\_classifier <- rpart(sentiment ~ ., data = train\_set, method = "class")  # Plot the decision tree  rpart.plot(dt\_classifier)  # Display a summary of the decision tree  summary(dt\_classifier)  # Predict the airline sentiment using the decision tree classifier and the testing set  dt\_predict <- predict(dt\_classifier, newdata = test\_set, type = "class")  # Compute the accuracy of the prediction on the testing set  dt\_accuracy <- accuracy(dt\_predict, test\_set$sentiment)  # Random forest  rf\_model <- randomForest(sentiment ~ ., data = train\_set, ntree = 20)  print(rf\_model)  #Predict on Train set  rf\_predict1 <- predict(rf\_model, newdata=train\_set, type="class")  accuracy(rf\_predict1,train\_set$airline\_sentiment)  # Predict on test data  rf\_predict2 <- predict(rf\_model, newdata=test\_set , type="class")  accuracy(rf\_predict2,test\_set$airline\_sentiment)  # Evaluate model performance  conf\_mat <- confusionMatrix(table(rf\_predict2, test\_set$airline\_sentiment))  conf\_mat |