BUDT758T: Data Mining & Predictive Analytics

Data Mining for Business (BUDT758T)

SENTIMENT ANALYSIS OF DELTA AIRLINE REVIEWS

Team Members:
Anjali Yamdagni
Aparna Raghavendra Rao
Haareca Chintala
Charishma Jaladi
Nishanthi Ravichandran

Table of Contents

EXECUTIVE SUMMARY

Sentiment analysis is the process of extracting subjective information from text, and it has become an increasingly popular tool for businesses to understand customer feedback.

By analyzing reviews of Delta Airlines, the company can gain insights into how customers feel about their experiences with the airline. This information can then be used to make data-driven decisions that improve customer satisfaction.

Problem Statement

In recent times, Delta Airlines has observed a sharp decline in its revenue. While the problem could be attributed to the global pandemic, the reasons could have unexplored factors such as passenger sentiments, etc. The aim here is to unearth the sentiment value of Delta's customers and help them gauge its impact on the organizational revenue.

Proposed Solution

- Observe trends depicted by customer sentiments, subject to other factors like the month of the year or the traveler type.
- Help forecast travelers based on the relative sentiment in the group
- Help identify the pain points by highlighting the frequently encountered problems by the passenger.

Proposed value

- Better Delta Airlines' customer base
- Refine marketing strategies
- Solution scalable to other service industries such as Hotel Management, etc.

I. DATA DESCRIPTION

The dataset used for this project was the Kaggle Delta Airline Review data for Sentiment Analysis Dataset.

The original dataset comprises 8 columns and 2689 observations. We split a part of the data into additional columns for ease of calculation. We chose the variable Traveler_types as our dependent variable. The data dictionary is defined below -

Variables:

1. **Traveler_types:** The type of traveler, i.e, Solo Leisure, Couple Leisure, Family Leisure and Business - **Categorical**
2. **Star rating:** A numerical rating from 1 to 10 stars given to Delta Airlines by the customer where 10 is the highest rating and 1 is the lowest rating - **Numerical**
3. **Date:** The date when the review was written. This variable was split into day, month and year for our analysis.
4. **Seat Type:** the type of seat the customer flew in, such as Economy Class, Premium Economy, First Class and Business Class - **Categorical**
5. **Routes:** The origin and destination of the flight. This variable was split into departure, arrival and via for our analysis.
6. **Country:** The country where the customer is from.
7. **Reviews:** The reviews a customer has written about their experience with Delta Airlines, including positive and negative aspects of their trip.

A sample of the data can be found below:

Customer_ID	star rating	date	Seat Type	routes	traveler_types	country	reviews
1	4	17th February 2023	Economy Class	New York to Tel Aviv	Solo Leisure	United States	âœ... Trip Verified | First, travel st
2	1	16th February 2023	Economy Class	Milwaukee to Ft Lauderdale	Business	United States	âœ... Trip Verified | Delta did not
3	1	10th February 2023	Economy Class	New York to Charlotte	Business	United States	âœ... Trip Verified | Per our pilot,
4	1	8th February 2023	Premium Economy	Atlanta to Cape Town	Family Leisure	United States	âœ... Trip Verified | We flew from
5	2	6th February 2023	Economy Class	Dallas to Atlanta	Solo Leisure	United Kingdom	âœ... Trip Verified | Checkin staff
6	3	6th February 2023	First Class	San Jose to Philadelphia via Atlanta	Solo Leisure	United States	Not Verified | After a horrible exp
7	9	5th February 2023	Economy Class	Los Angeles to Kona	Solo Leisure	United States	âœ... Trip Verified | Mixed bag. Cr
8	6	4th February 2023	Economy Class	Orlando to Lisbon via New York	Couple Leisure	United States	Not Verified | Our flight was to le

Delta Airlines reported its highest annual revenue of $44.438 billion in the year 2018, but saw a decline in the subsequent years, owing to covid. This project can assist the airlines

in bettering its huge customer base and refining its marketing strategies. The recommended solution can be scaled to other service industries such as Hotel Management etc. and thus, can have a farther reaching utilization.

II. RESEARCH QUESTIONS

The following are the questions we aimed to answer through our analysis.

- How are the current reviews by the customers on Delta Airlines? Is there any observable trend with respect to the overall sentiment?
- Can we forecast travelers into categories based on their reviews?
- Can the sentiment be improved? What are the major issues affecting the sentiment values?

III. METHODOLOGY

1. Data Preprocessing

- The gathered data was checked for unicode encoding, since R can not interpret these values).
- Duplication at all levels, across all rows and columns were checked and accounted for.
- The 'Date' variable was split into 3 variables, i.e, 'Day', 'Month', 'Year'. This was for the purpose of identifying trends based on quarters, etc.
- The routes variable was subjected to a split, wherein the 3 resulting variables were 'Departure', 'Arrival' and 'Via'.
- The purpose behind the split was to enable a deeper analysis regarding the issues that could possibly be faced during connectivity between the various locations.

After splitting the data we had a total of 12 columns excluding the CustomerID.
A sample of the data can be found below:

Customer	star rating	Day	Month	Year	Seat Type	Departure	Arrival	Via	traveler_types	country	status	reviews
1	4	17	February	2023	Economy Class	New York	Tel Aviv	Direct Flight	Solo Leisure	United States	Trip Verified	First, travel starts 6pm arrival at the airp
2	1	16	February	2023	Economy Class	Milwaukee	Ft Lauder	Direct Flight	Business	United States	Trip Verified	Delta did not issue any weather waivers
3	1	10	February	2023	Economy Class	New York	Charlotte	Direct Flight	Business	United States	Trip Verified	Per our pilot, there would be a slight del
4	1	8	February	2023	Premium Econo	Atlanta	Cape Tow	Direct Flight	Family Leisure	United States	Trip Verified	We flew from Atlanta to Cape Town and
5	2	6	February	2023	Economy Class	Dallas	Atlanta	Direct Flight	Solo Leisure	United Kingdom	Trip Verified	Checkin staff always seem unfriendly, no

2. Sentiment Analysis

The following packages were installed and employed for sentiment analysis of the Airline reviews:

- ‘stringr’
- ‘tm’
- ‘sentimentr’
- ‘ggplot2’
- ‘e1071’
- ‘SparseM’
- ‘lubridate’
- ‘cowplot’
- ‘quanteda’
- ‘stm’
- ‘stminsights’
- ‘tidyverse’
- ‘ggwordcloud’
- ‘corpus’
- ‘purr’

- Firstly, the preprocessed clean data was read into a dataframe ‘Deltadata’.
- Following this, the punctuations in the reviews were replaced with empty spaces.
- Thirdly, a corpus was created to proceed with further analysis, along with which we converted the text to lower case, removed the stop words, special characters and URLs.
- For the sake of convenience and faster execution, the lemmatization technique was used to convert the trimmed word to its root, rather than resorting to stem completion.
- Then, using the ‘sentiment_by’ function, the dataframe ‘sentiment’ was created which held the average sentiment scores for each review.
- The above data frame could be used to plot a histogram of the scores, and also calculate the average sentiment score per class of customers (for instance, score of customers of different seat types, different traveler types etc.)
- Further, each review was labeled ‘Positive’ or ‘Negative’ based on if the corresponding sentiment score is greater than or lesser than 0 respectively.

- The corpus was then converted into a Term Document Matrix (TDM), along with the label column from the previous step.
- Further, a transpose of the same was taken and converted to a Document Term Matrix which was then stored in a dataframe for the purpose of further analysis.

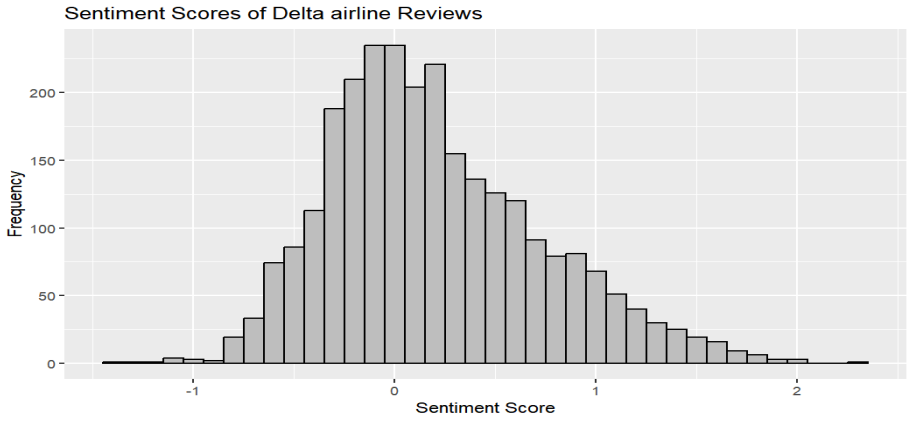
3. Data Mining Models

- Naive Bayes Model
- Logistic Regression Model:
- Random Forest Classification:
    - A need for improvement in accuracy prompted us to consider models that would aid in the best subset/ best features selection amongst the parameters in the data.
    - An alternative to random forest could have been regularization or best subset selection, however the random forest fared better in terms of information gain at each step.
    - As in the case of the previous models, the data was divided into training and test data sets which were then used to fit and predict the 'traveler_types' based on the sentiment label. This was done using the **'randomForest'** library.
- Hierarchical Clustering
- Topic Modeling

IV. VISUALIZATION

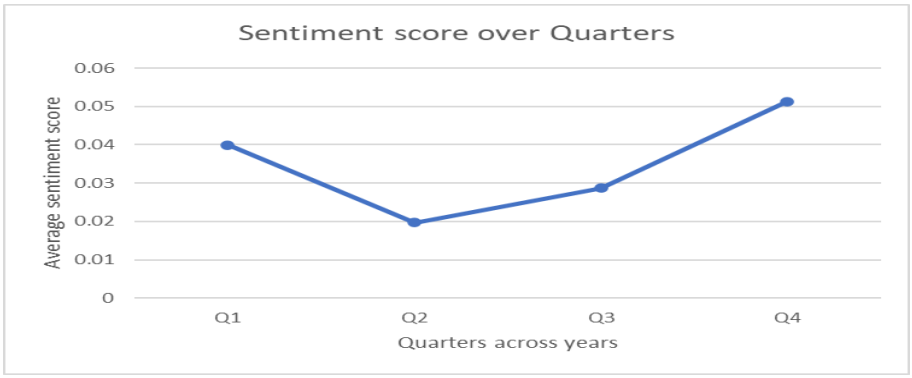
1. Sentiment scores - Histogram

This visualization provides insights into the sentiment distribution of the dataset. The x-axis represents the sentiment score, while the y-axis represents the frequency of reviews with that sentiment score. By analyzing this histogram, we can understand the overall sentiment of the dataset.



2. Sentiment scores - By Time of Travel

The line chart provides a concise and visual representation of average sentiment scores across quarters, allowing the audience to easily grasp the sentiment trends over the course of the year.. The sentiment scores are plotted on the y-axis, while the quarters of the year are plotted on the x-axis. sentiment was relatively higher in Q1 and Q4 compared to Q2 and Q3.From this we can understand that the sentiment scores are higher for holiday season

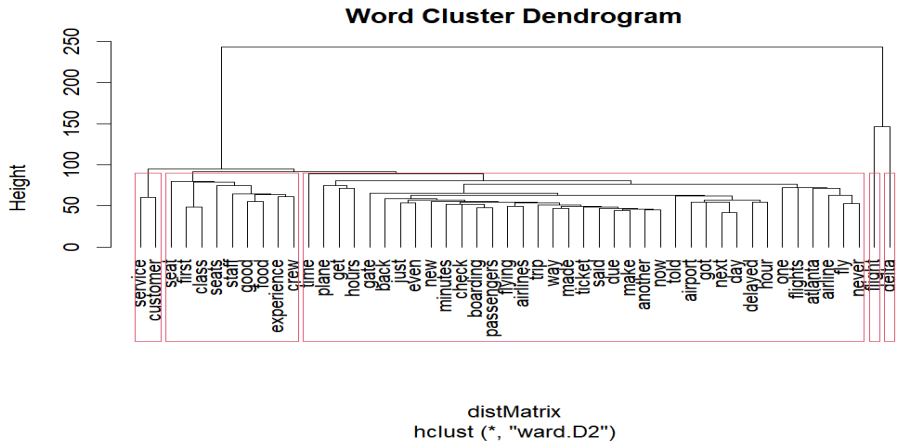


3. Hierarchical clustering - Dendrogram

This visualization helps us identify clusters and relationships among words in the dataset. It uses hierarchical clustering to group similar words together based on their co-occurrence patterns. The height of the branches represents the similarity between the words. By analyzing the dendrogram, we can discover that the terms customer and service occur together in the first cluster followed by the next cluster having a good set of terms like

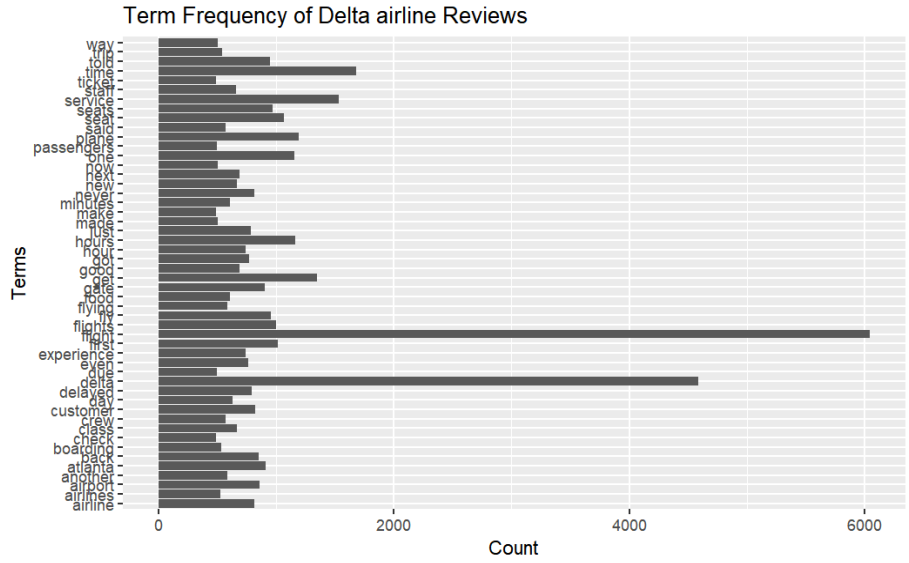
good, first, class,and experience.

In the third cluster it's mostly a set of negative words occurring together, the next two clusters contain just delta and flight.



4. Term Frequency - Bar Chart

This visualization allows us to understand the most frequent terms in our dataset. We can observe which words appear most often and gain insights into the key topics or themes present in the reviews. By analyzing the term frequency plot, we can identify the important terms that shape the content of our dataset. Delta and flights occur the most followed by time and service.

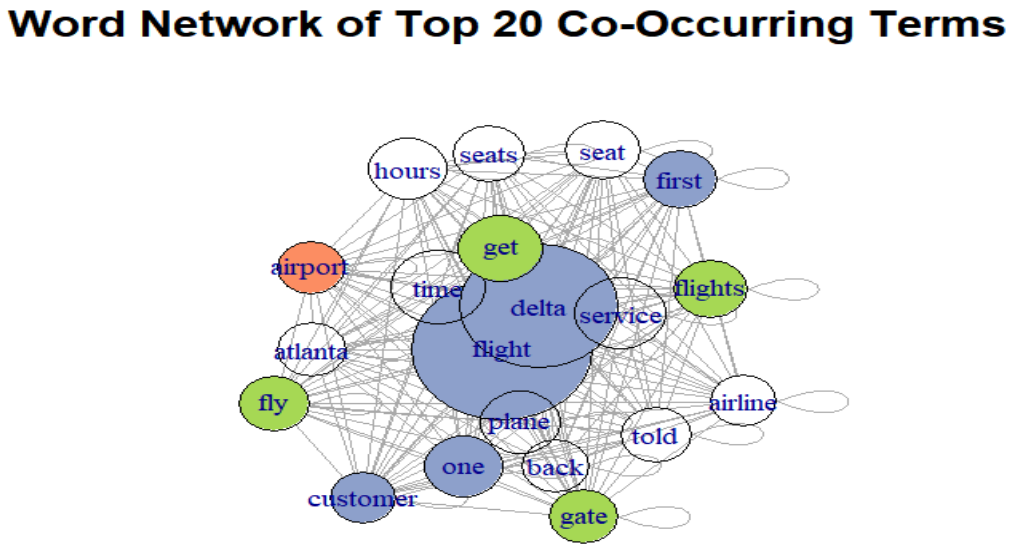


5. Word Cloud

This visualization provides a visually appealing representation of the most frequent words in our dataset. The size of each word in the cloud corresponds to its frequency, while the color palette adds an additional visual dimension.By analyzing the word cloud, we can quickly grasp the prominent terms and visually understand their significance.

6. Top 20 Co-Occurring Terms - Word Network

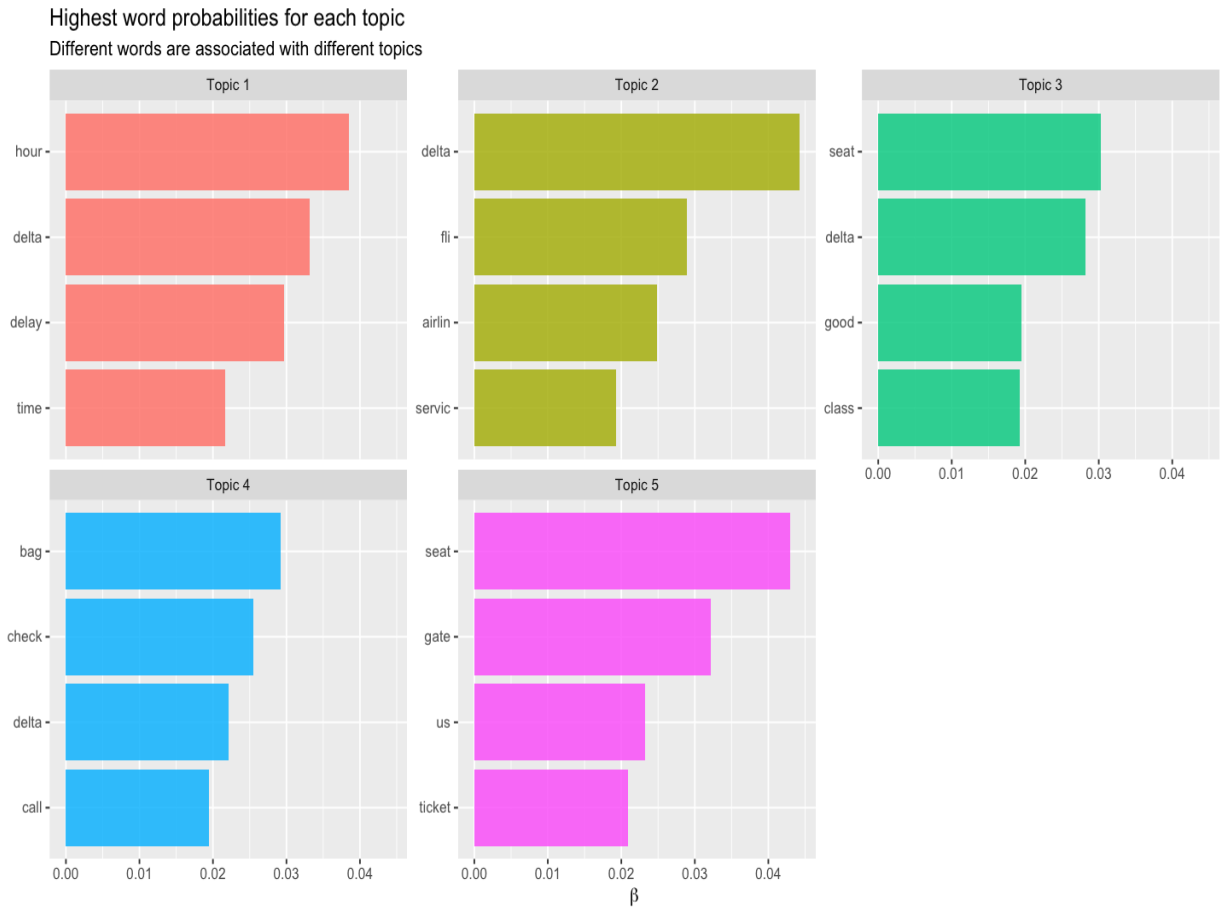
This visualization uncovers relationships between words by representing them as nodes connected by edges. Nodes represent individual words, while edges indicate co-occurrence patterns.The color of each node indicates the group to which the term belongs to, a group number is assigned to each term based on hierarchical clustering. Nodes with similar colors are likely to have similar co-occurrence patterns.



7. Topic Modeling

Another method under the unsupervised learning techniques which uses ‘Structure Topic Modelling’ to cluster similar groups of words. The visualization here is a depiction of the topics when the number of clusters defined is 5. The first graph gives a summary of the words grouped under each topic. The second plot is the visualization of topics based on their expected proportion within the dataset. The third plot is an insight into the most frequently occurring words within each group.

```
A topic model with 5 topics, 2680 documents and a 967 word dictionary.
Topic 1 Top Words:
         Highest Prob: delta, hour, delay, time, fli, day, get
         FREX: cancel, delay, hotel, day, hour, miss, refund
         Lift: florida, rental, cancel, strand, mechan, hotel, reschedul
         Score: florida, cancel, delay, hotel, refund, hour, day
Topic 2 Top Words:
         Highest Prob: delta, good, servic, crew, time, staff, great
         FREX: great, friend, crew, nice, good, profession, thank
         Lift: tast, heathrow, amaz, smooth, great, effici, london
         Score: tast, good, great, entertain, crew, ife, excel
Topic 3 Top Words:
         Highest Prob: bag, check, delta, told, call, luggag, get
         FREX: bag, luggag, baggag, check, call, claim, carri
         Lift: tag, suitcas, claim, carri, carry-on, tsa, bag
         Score: tag, bag, check, told, luggag, call, claim
Topic 4 Top Words:
         Highest Prob: seat, delta, class, first, comfort, economi, busi
         FREX: class, economi, upgrad, busi, premium, leg, comfort
         Lift: segment, coach, class, elit, reclin, narrow, flat
         Score: segment, class, seat, economi, comfort, entertain, food
Topic 5 Top Words:
         Highest Prob: seat, gate, plane, us, delta, ask, board
         FREX: gate, sit, rude, togeth, husband, ask, assign
         Lift: enter, assign, mask, togeth, wheelchair, young, stewardess
         Score: enter, gate, seat, sit, ask, togeth, rude
```


Highest word probabilities for each topic
Different words are associated with different topics
Topic 1
hour
delta
delay
time
Topic 2
delta
fli
airlin
servic
Topic 3
seat
delta
good
class
Topic 4
bag
check
delta
call
Topic 5
seat
gate
us
ticket
0.00
0.01
0.02
0.03
0.04
β

8. N-grams

N-grams is the study of most frequently occurring words together, which provides useful insight in terms of predicting the consequent word. In case of sentiment analysis, it could also help overcome one of the limitations that sentiment analysis is often troubled by, which is the presence of ‘negation’ or ‘sarcasm’, which in turn leads to a misclassification of sentiment. Frequently occurring words could be used to study the most commonly assigned sentiment , which could then be used to update the ‘bag of words’ and increase the domain knowledge of the industry.

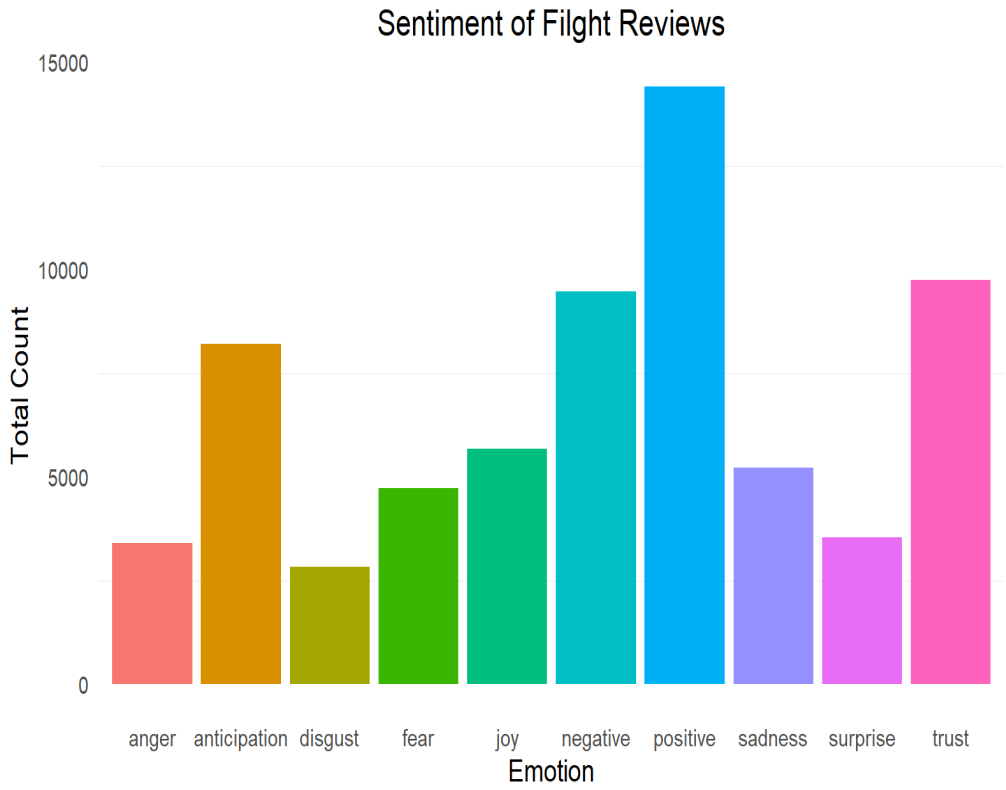
Bi-grams (2-grams):

customer_service	first_class	fly_delta	new_york	connecting_flight	flight_attendants	delta_air	air_lines
634	375	325	315	293	288	241	235
never_fly	flight_attendant						
214	208						

Tri-grams (3-grams):

delta_air_lines	never_fly_delta	salt_lake_city	new_york_jfk	delta_customer_service	first_time_flying
235	126	117	75	39	30
flight_got_delayed	poor_customer_service	first_class_ticket	via_new_york		
29	27	24	24		

9. Emotion Analysis



IV. RESULTS

1. Model Outputs

1) Naive Bayes Model

```
```{r}
# TRAIN NAIVE BAYES MODEL
model <- naiveBayes(traindata[,-1],traindata[,1]);

# PREDICTION
Predictions <- predict(model,testdata[,-1])
(confusion = table(testdata[,1],Predictions))


(Acc_test = (confusion[1,1]+confusion[2,2])/sum(confusion))

```
```

```
           Predictions
            Negative Positive
  Negative       250       84
  Positive       164      309
[1] 0.692689
```

The data frame comprising the Document Term Matrix is split

into train and test data sets for the purpose of training and prediction of sentiment label classification. The model above fits the train data to the 'naiveBayes' function, which is used against the sentiment label (computed earlier).
The trained model is then used for prediction on the test data which yields an accuracy of 69.2%.

This could be due to Naive Bayes's inability to accurately study the underlying patterns due to inherent low variance.

2) Logistic Regression

```
          Predicted
Actual     Negative Positive
  Negative      234      100
  Positive      101      372
[1] 0.7509294
```

As with the previous model, the train and test data sets are taken in order to analyze the accuracy of sentiment label classification, which is done using the 'glm' model.
The trained model is then used for prediction on the test data which yields an accuracy of 75.02%.

The ability to analyze complex patterns helps the Logistic Regression model outperform the Naive Bayes model, however a further improvement could help reduce the misclassification of the reviews.

3) Random Forest Classifier:

```
               predict_rf1
                Business Couple Leisure Family Leisure Not Specified Solo Leisure
  Business              5              9             15            11           97
  Couple Leisure        4             16             16            18          112
  Family Leisure        3             13             50            16           69
  Not Specified         2              6             12            62           34
  Solo Leisure          2             18             29            21          167
              Business Couple Leisure Family Leisure Not Specified Solo Leisure
class.error
Business            20             28             37            33          180
0.9328859
Couple Leisure      16             45             64            35          208
0.8777174
Family Leisure       8             26            122            32          216
0.6980198
Not Specified        4             16             22           144          105
0.5051546
Solo Leisure        15             50             66            59          331
0.3646833
```

The Random Forest Classification model takes the ‘traveler_types’ as the dependent variable, wherein, the aim is to predict the type of traveler according to the assigned sentiment label.
As evident from the results above, the accuracy is skewed where the classes ‘Not Specified’ and ‘Solo Leisure’ perform the best of the lot.

The results here could be used to further subset the dataset based on the classes that perform the best and gauge the impact of other factors in tandem with the sentiment label.

When viewed holistically, the model does not show a marked improvement over the other models in terms of accuracy, however there is a better insight into which are the subsets of the current data that could help improve the predictive accuracy.

This is comparable to if the data in question were to be regularized using either the Ridge or Lasso methods, thereby reducing the dimensions and increasing the predictive power of the model.

2. Comparison of the Unsupervised Clustering Methods used in the Model

Hierarchical Clustering Vs Topic Modelling

Despite the difference in the underlying algorithm used to compute both the models (Hierarchical uses ‘distance’ metric whereas Topic Modelling uses the Structure Topic Model to provide

insights), it is interesting to see the commonalities in the results produced by the two methods.

Both methods had a pre-defined cluster size of 5 and albeit a few differences, the theme of the clusters remained roughly the same. It was observed that the common words cropping up in both the results were:

'Flight'
'Delta'
'Hour'
'Time'
'Delay'

When grouped together, these words provide a negative connotation which can be corroborated by the fact that the weightage of both the negative and positive sentiments in the dataset is approximately the same.
The most expected topic amongst the reviews provided by travelers reflected the same sentiment which can further be backed by the best subset that was used to predict the Random Forest Model.

INFERENCES AND FUTURE SCOPE

1) Although the overall sentiment regarding Delta Airlines is positive, the difference is marginal and can be improved upon by reflecting on the results derived, i.e, actionable measures looking to minimize problems such as flight cancellations, flight scheduling, waiting time for passengers, etc.
2) The model can be further improved upon by incorporating more data which would give a better aggregate measure of the sentiment of the crowd.

V. CONCLUSION

Overall, sentiment analysis has proven to be a valuable tool to understand customer feedback and improve customer satisfaction. By identifying pain points and taking targeted actions, Delta Airlines can make data-driven decisions that have a positive impact on their customers.

While there are limitations to sentiment analysis, it remains a powerful tool for businesses looking to improve customer satisfaction and stay competitive in today's market.

VI. REFERENCES

1) https://ladal.edu.au/topicmodels.html
2) https://www.rdocumentation.org/packages/ngram/versions/3.2.2
3) http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
4) https://www.rdocumentation.org/packages/syuzhet/versions/1.0.6/topics/get_nrc_sentiment

Data source: *https://www.kaggle.com/datasets/datazng/delta-airline-review-dataset-sentiment-analysis?select=Delta_Airline_Review_Dataset-Asof02172023.csv*