Enhancing Airline Customer Experience Through Sentiment Analysis of Passenger Reviews

Table of Contents

1. Executive Summary

NLP is an important analysis that utilizes AI and Linguistics to help computers understand human written language. One of the most important aspects of NLP is sentiment analysis, which is used by researchers to mine opinions, such as positive or negative. Sentiment analysis can be performed by using ML techniques; however, with the promising nature of NN, this project utilized RNN to classify reviews from ten airlines as either positive or negative. Trends indicated that Eva Air had enjoyed a largely positive rating over the years, while Turkish Airlines has struggled with poor reviews. The reviews were pre-processed: tokenization, lower-casing, stopwords removal, punctuation removal, and white space removal. This was followed by the use of NRC to categorize each term as positive and negative while also comparing the aggregate rating to the actual reviews. A notable discrepancy was found between the two. However, in both cases, positive reviews dominated. The RNN was utilized to classify the labeled reviews as positive or negative. Despite the relatively low accuracy, it remains a crucial model in text analysis if provided with adequate computing power. LDA also proved to be a reliable method when investigating topics within textual data. Findings revealed that food, comfort, and services were topics associated with positive sentiments, and time, delay, and services were the most prevalent with negative sentiments.

2. Introduction and Background

NLP is a branch of artificial intelligence (AI) and Linguistics that allows computers to understand human written language. Linguistics is crucial in Natural Language Processing (NLP) as it provides the primary foundational principles for understanding and generating human language. For instance, there is phonology, lexical, syntactic, discourse, pragmatic, sentiment, and semantic analysis. Together, these linguistic components enable NLP systems and algorithms to understand and provide meaningful insights from textual data that may otherwise be time-consuming.

Notably, with the increased demand for processing big data, companies find themselves with human-written texts, either as reviews, books, articles, or social media posts such as Tweets. One notable way of gathering meaningful information from such texts is through sentiment analysis. Sentiment analysis involves processing and analyzing sentiments extracted from textual data. This analysis can be used to identify and categorize the emotions, opinions, or attitudes expressed in textual data. Most importantly, it has been used to categorize sentiments as being positive, negative, or neutral—towards a particular subject or topic. Sentiment analysis has become a critical tool for organizations, governments, industries (beyond tech), and even individual researchers.

Sentiment analysis has been largely done with ML. This is because ML models have proven to be reliable in the classification of texts into either positive, negative, or neutral. For instance, random forest classifier has had high accuracy rates in classifying Twitter reviews with accuracies as high as 96%. Additionally, the support vector model (SVM) is a popular text classification ML model that has been used to classify reviews based on sentiments. A comparison of the RFM model to an SVM model and shows that the SVM model has a higher accuracy.

However, SVM models have difficulties identifying or taking advantage of latent semantic information. An example is the identification of terms that have similar meanings, such as "king" and "queen," which mean royalty. With SVM, the relationship between words may not be captured well, hence the increasing reliance on neural network models such as recurrent neural networks (RNN). RNN can take advantage of the relationship between terms to increase the accuracy of sentiment analysis while also being faster than the SVM. RNN gives a highly accurate classification of sentiments, as high as 98%. Therefore, this project utilizes RNN to classify sentiment data for airlines with the incorporation of word embedding using the pre-trained GloVe model. The goals or objectives of the project include:

- Exploratory data analysis of the airline reviews to unearth trends
- Sentiment analysis: categorizing reviews as either positive, negative, or neutral and using RNN to classify these reviews
- Topic modeling to identify opinions that relate to entertainment, food, value for money, service, and comfort.

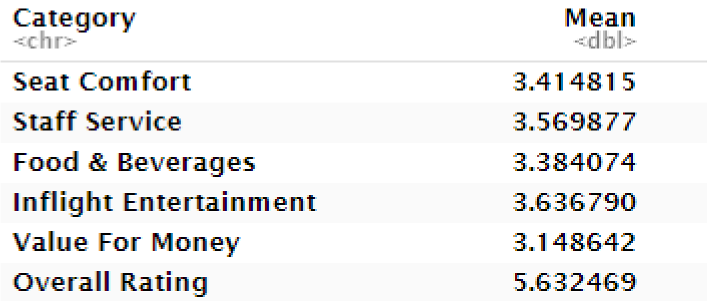
3. Solutions

3.1. Exploratory Data Analysis

It is critical to understand the data before processing as one can remove outliers and null values and, most importantly, understand basic trends. The columns for the airline data were: "Title," "Name," "Review Date," "Airline," "Verified," "Reviews," "Type of Traveller," "Month Flown," "Route", "Class," "Seat Comfort," "Staff Service," "Food & Beverages," "Inflight Entertainment," "Value For Money," "Overall Rating," "Recommended." However, based on the project's objectives, I chose to work with ten of these variables: "Airline," "Reviews," "Month Flown," "Seat Comfort," "Staff Service," "Food & Beverages," "Inflight Entertainment," "Value For Money," "Overall Rating," and "Recommended."

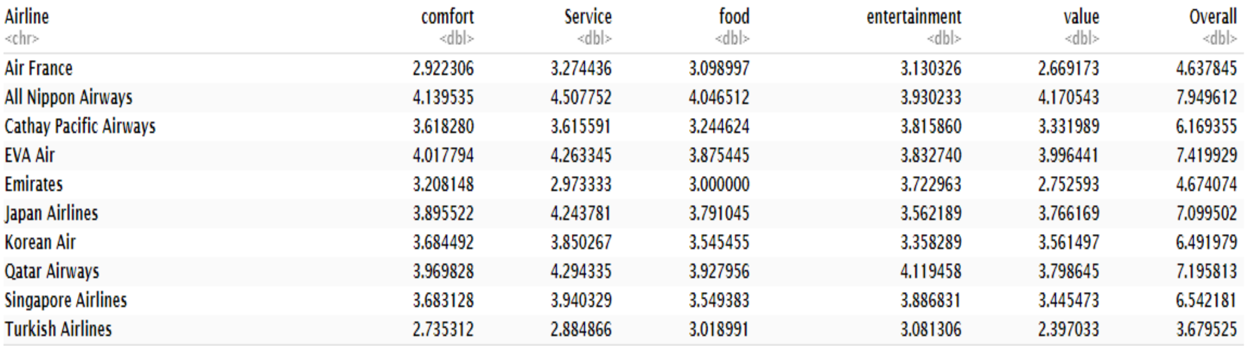
3.1.1. Mean

The main variables of interest were**:** "Seat Comfort," "Staff Service," "Food and beverages," "Inflight Entertainment," "Value For Money," and "Overall Rating." Their mean values are displayed below:



Category <chr>	Mean <dbl>
Seat Comfort	3.414815
Staff Service	3.569877
Food & Beverages	3.384074
Inflight Entertainment	3.636790
Value For Money	3.148642
Overall Rating	5.632469

The majority of reviewers were pleased with the entertainment; however, most of them seemed unsatisfied with the value they got from the money they spent on their flights.

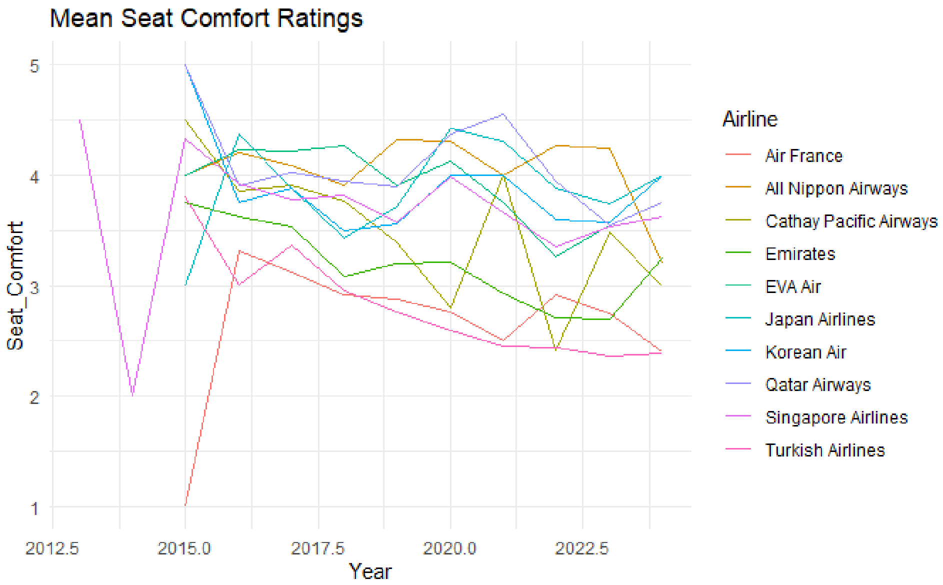


Airline <chr>	comfort <dbl>	Service <dbl>	food <dbl>	entertainment <dbl>	value <dbl>	Overall <dbl>
Air France	2.922306	3.274436	3.098997	3.130326	2.669173	4.637845
All Nippon Airways	4.139535	4.507752	4.046512	3.930233	4.170543	7.949612
Cathay Pacific Airways	3.618280	3.615591	3.244624	3.815860	3.331989	6.169355
EVA Air	4.017794	4.263345	3.875445	3.832740	3.996441	7.419929
Emirates	3.208148	2.973333	3.000000	3.722963	2.752593	4.674074
Japan Airlines	3.895522	4.243781	3.791045	3.562189	3.766169	7.099502
Korean Air	3.684492	3.850267	3.545455	3.358289	3.561497	6.491979
Qatar Airways	3.969828	4.294335	3.927956	4.119458	3.798645	7.195813
Singapore Airlines	3.683128	3.940329	3.549383	3.886831	3.445473	6.542181
Turkish Airlines	2.735312	2.884866	3.018991	3.081306	2.397033	3.679525

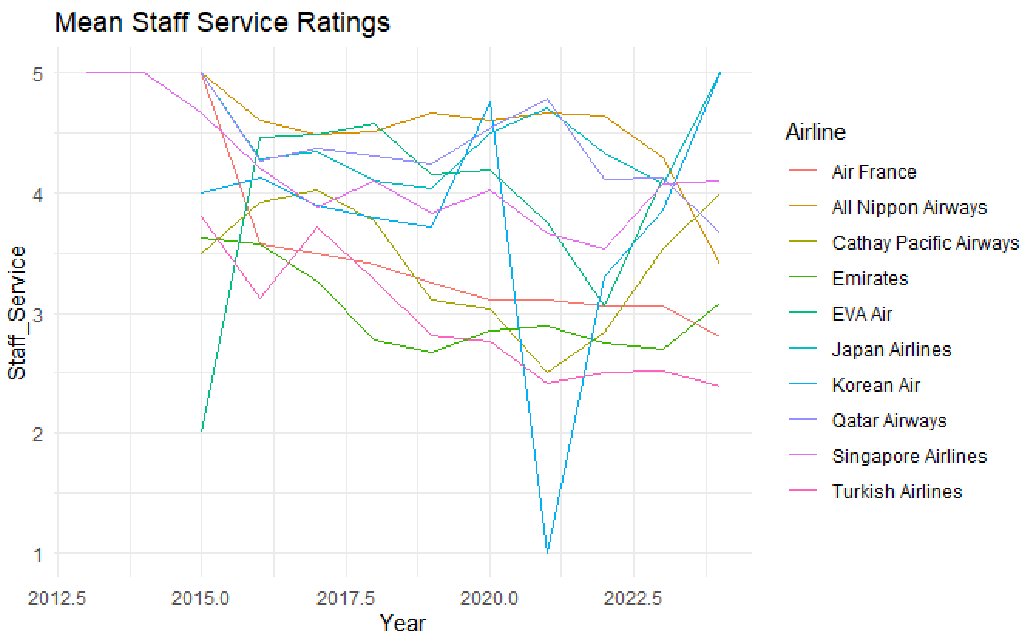
All Nippon and Eva Air had the best comfort; All Nippon, Eva Air, Japan, and Qatar Airs had the best services; All Nippon had the best food; Qatar Airways entertainment; and All Nippon value for money. All around, All Nippon was the best-rated. The worst rated was Turkish Airlines overall, with its poorest rating being its comfort, services, and value for money.

3.1.2. Trends

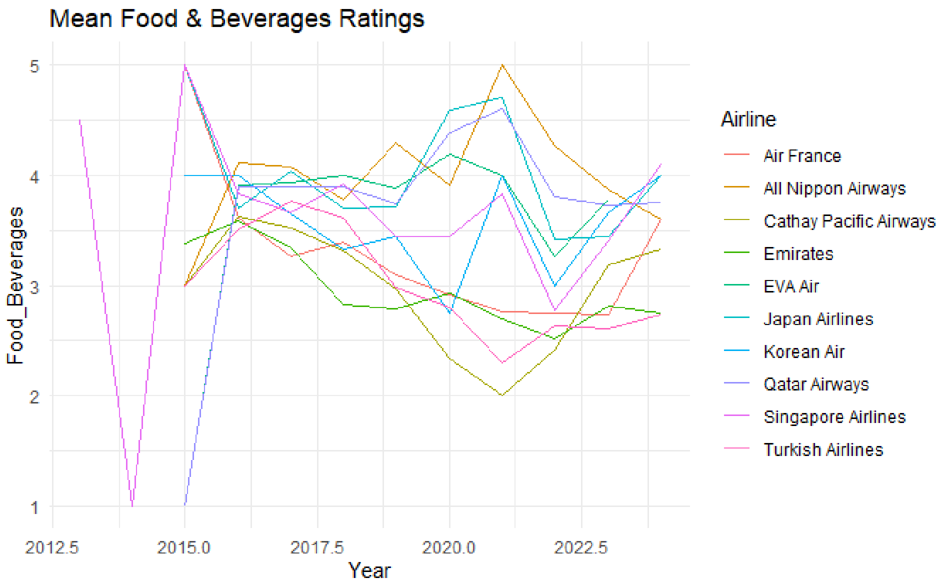
Turkish Airlines has seen a dip in reviews regarding seat comfort since 2015; the same trend has been observed with Air France.



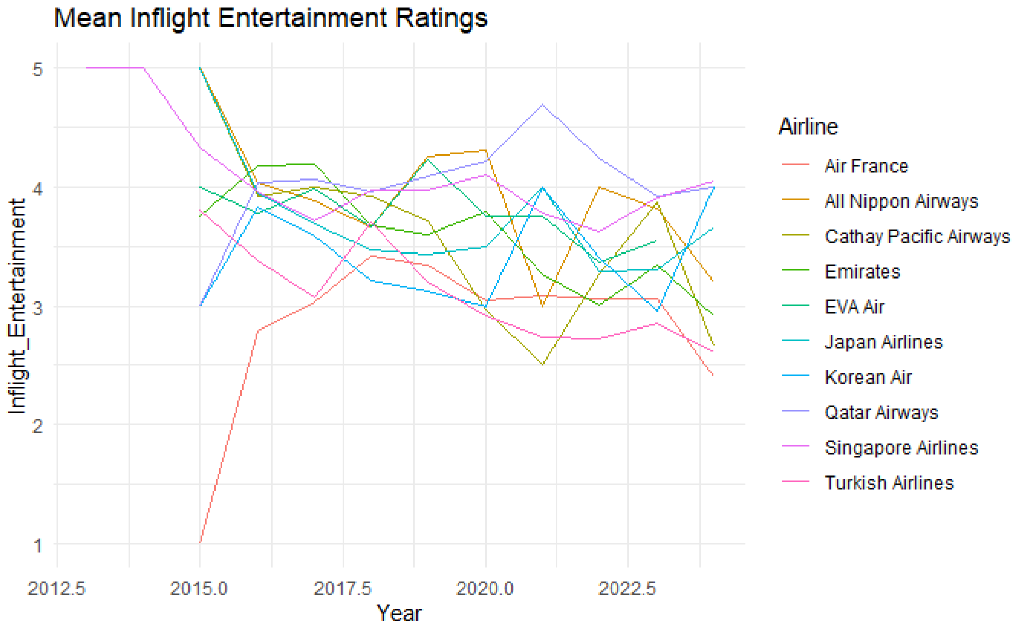
As for staff service, Turkish Airlines, Air France, and Emirates have dominated among the poorly rated in the last decade.



In terms of food and beverages, the best-rated has been predominantly All Nippon; however, recently, Singapore Airlines, Japan Airlines, and Korean Air have been the best-rated.



The best entertainment ratings have recently been given to Japan Airlines, Qatar Airways, and Singapore Airlines, with Qatar Airways being the most dominant.

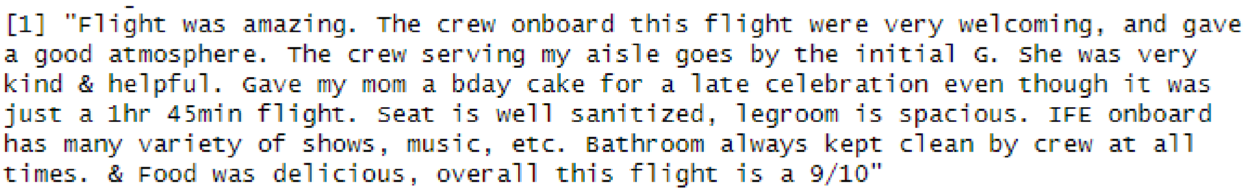


In late 2021, some airlines were rated poorly for value for money, but they seemed to have made efforts to improve this as they saw their ratings improve; such is the case for Cathay Pacific, Korean Air, and Eva Air. Korean Air, Singapore Airlines, Qatar Airways, Eva Air, and Japan Airlines have had the best ratings recently.

3.2. Natural Language Processing

3.2.1. NLP Pre-processing

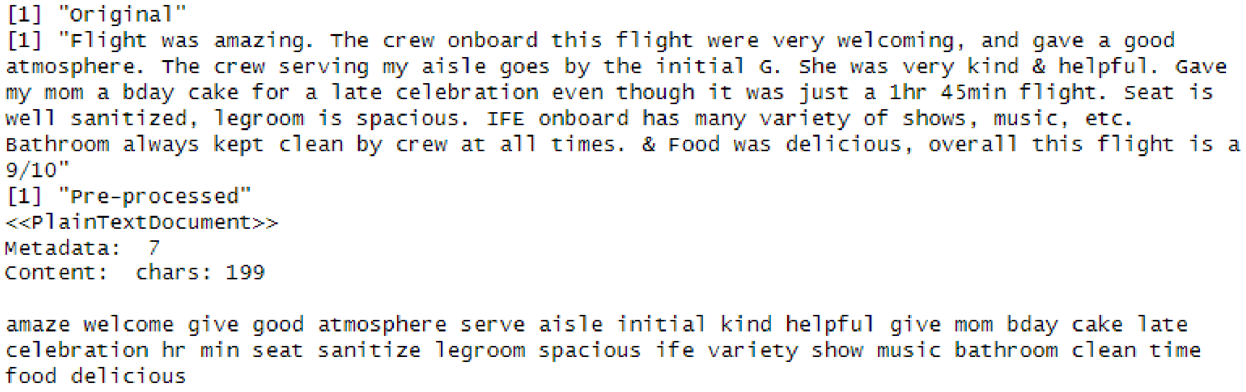
In the case of NLP, pre-processing steps need to be applied to the textual data before any opinion mining or sentiment analysis is done. The image below is an excerpt of the textual data or airline review:



```
[1] "Flight was amazing. The crew onboard this flight were very welcoming, and gave
a good atmosphere. The crew serving my aisle goes by the initial G. She was very
kind & helpful. Gave my mom a bday cake for a late celebration even though it was
just a 1hr 45min flight. Seat is well sanitized, legroom is spacious. IFE onboard
has many variety of shows, music, etc. Bathroom always kept clean by crew at all
times. & Food was delicious, overall this flight is a 9/10"
```

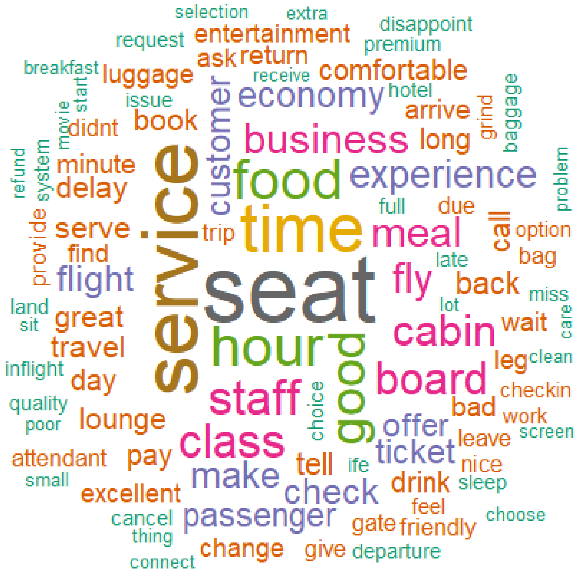
In this pre-processing step, the reviews need to be prepared for analysis by the computer, as in its textual form, it is raw data. Therefore, the following steps were taken to pre-process the reviews.

- **Tokenization:** tokenization is the splitting of sentences into either single words (unigrams), two words (bigrams), or multiple words (n-grams). In this case, the reviews were split into individual words (unigrams), which allowed for the next couple of steps where unwanted terms were removed.
- **Lower-casing**: transforming all texts to lower-case for uniformity.
- **Removal of punctuation** was essential in ensuring that unnecessary punctuation marks were removed.
- **Stop-words removal** was done. This is the removal of common terms such as "is" and "it," which do not add much meaning to the words.
- **Whitespace removal** was done to remove spaces between words.
- **Lemmatization** was performed to ensure that the basic or root form of words was utilized, such as "operational" to "operator." The excerpt below shows the end result of the pre-processed review compared to the original. It has fewer terms, and the terms preserved still preserve much of the meaning.



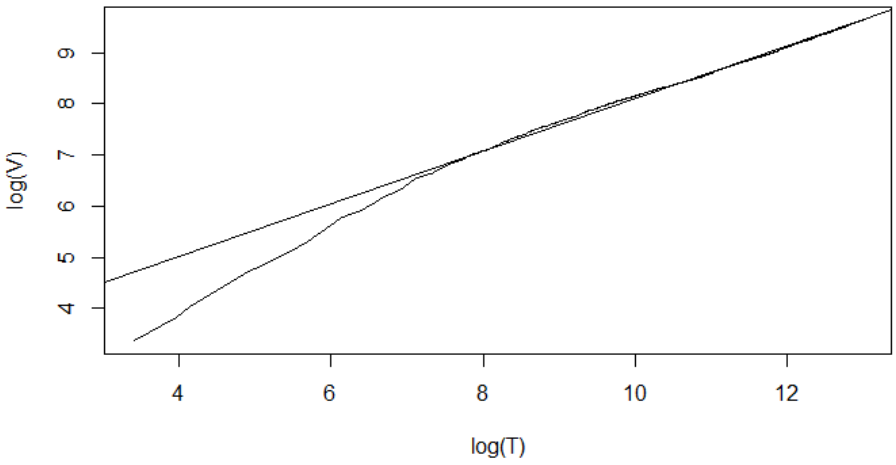
```
[1] "Original"
[1] "Flight was amazing. The crew onboard this flight were very welcoming, and gave a good
atmosphere. The crew serving my aisle goes by the initial G. She was very kind & helpful. Gave
my mom a bday cake for a late celebration even though it was just a 1hr 45min flight. Seat is
well sanitized, legroom is spacious. IFE onboard has many variety of shows, music, etc.
Bathroom always kept clean by crew at all times. & Food was delicious, overall this flight is a
9/10"
[1] "Pre-processed"
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 199

amaze welcome give good atmosphere serve aisle initial kind helpful give mom bday cake late
celebration hr min seat sanitize legroom spacious ife variety show music bathroom clean time
food delicious
```

Below, the terms in all reviews are displayed based on their frequency. As seen, the most common term, which is the largest in the word cloud, is "seat," suggesting that the majority of reviews mentioned something about comfort. There is also "food," "time," and "service," which were discussed a lot, highlighting possible discussions about food and beverages, travel time, and staff service. "Ife," which stands for inflight entertainment, also appears, and "economy" and "business," which might suggest value, also appear in the word cloud, suggesting that the five main topics of concern this project also aims to analyze are discussed in the reviews.

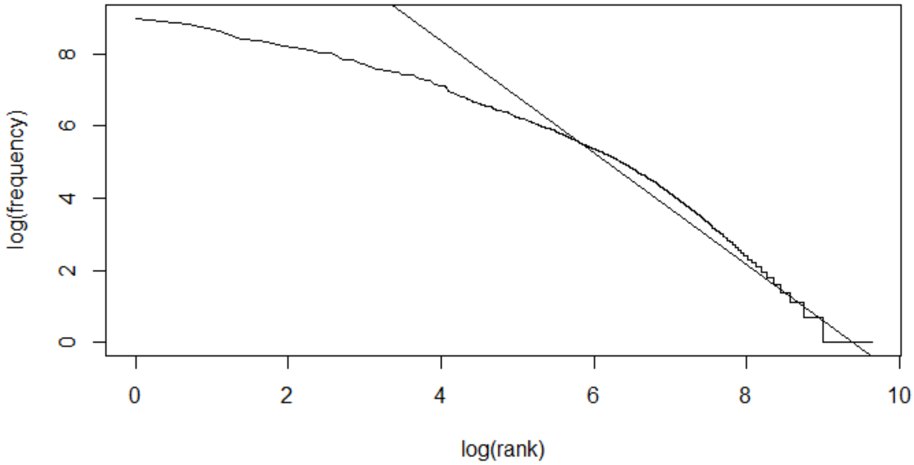


3.2.1.1. Heap's Law & Zipf's Law

These two laws help ascertain the importance of pre-processing, as they show terms or words appear in a bag of words, corpus, or document matrix. According to Heap's law, as the size of a corpus (a list of words) increases, the vocabulary size grows, but at a decreasing rate. As per Heap's Law, the curve shows that the terms saw a reduction in their growth rate as a result of pre-processing as the slope of the curve approaches 0.5, as shown below.



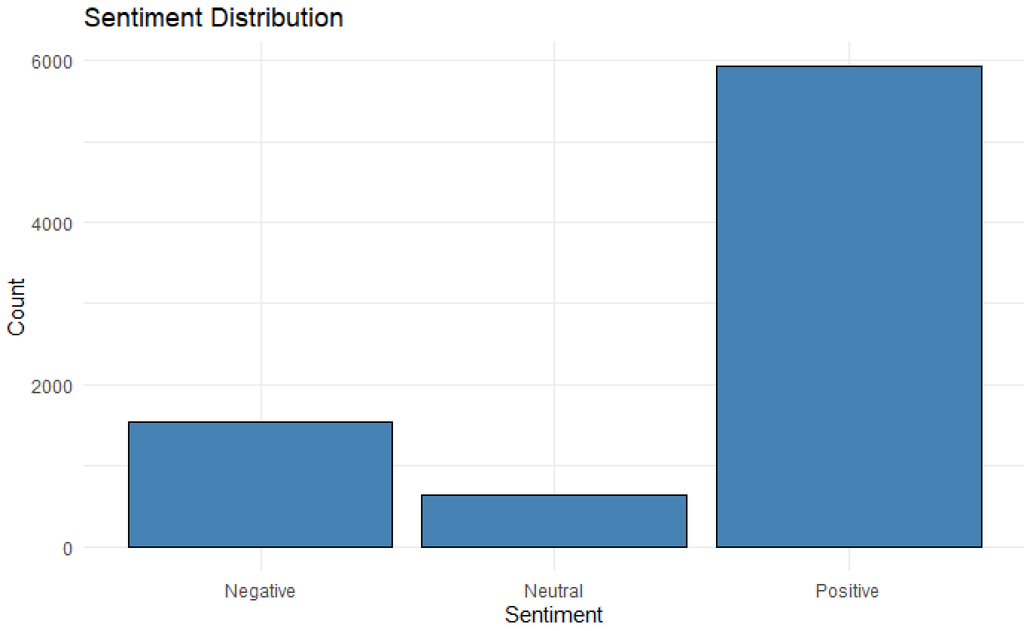
According to Zip's law, the most frequent word occurs almost twice as much as the second most frequent word, three times as much as the third most frequent word, and so on. As seen below, the terms approach a slope of -1.5, as expected by Zipf's law. Hence, it suggests that pre-processing is important in textual analysis, such as semantic analysis.



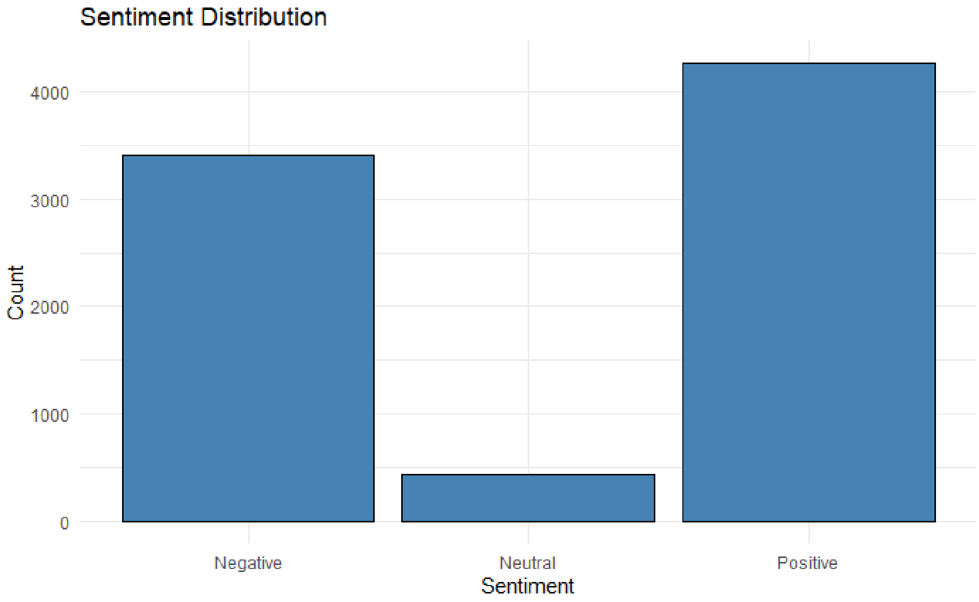
3.2.2. Sentiment Analysis

3.2.2.1. Categorizing Reviews to Positive, Negative, and Neutral

In semantic analysis, one of the most important steps is to categorise texts into either positive, negative, or neutral. This is usually performed by assigning a specific value to each term. The categorization was done through a dictionary, NRC, an openly available dictionary of positive and negative terms. Each term in a review was assigned a value, and the values were summed for each review, with values > 0 being positive, < 0 negative, and = 0 neutral. As seen below, the majority of the terms were neutral.



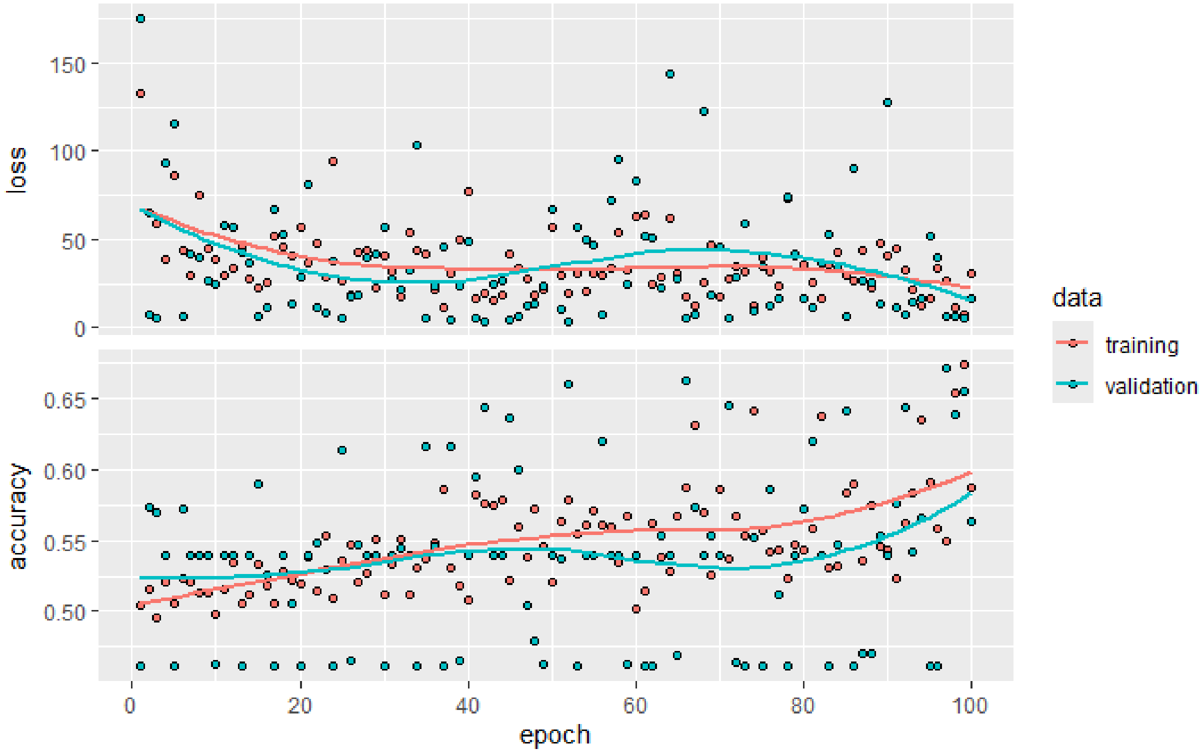
However, the data used had ratings in the "Overall Rating" column (1 to 10). This can also be used to categorise the reviews. Here, data is categorized as positive if > 5, neutral if = 5, and negative if < 5, and the results are as below. Comparing the use of values from the dictionaries and the overall ratings suggests there is a significant discrepancy; the dictionary does fail to accurately capture the negative sentiments, while it seems to over-exaggerate the positive ratings.



3.2.2.2. RNN

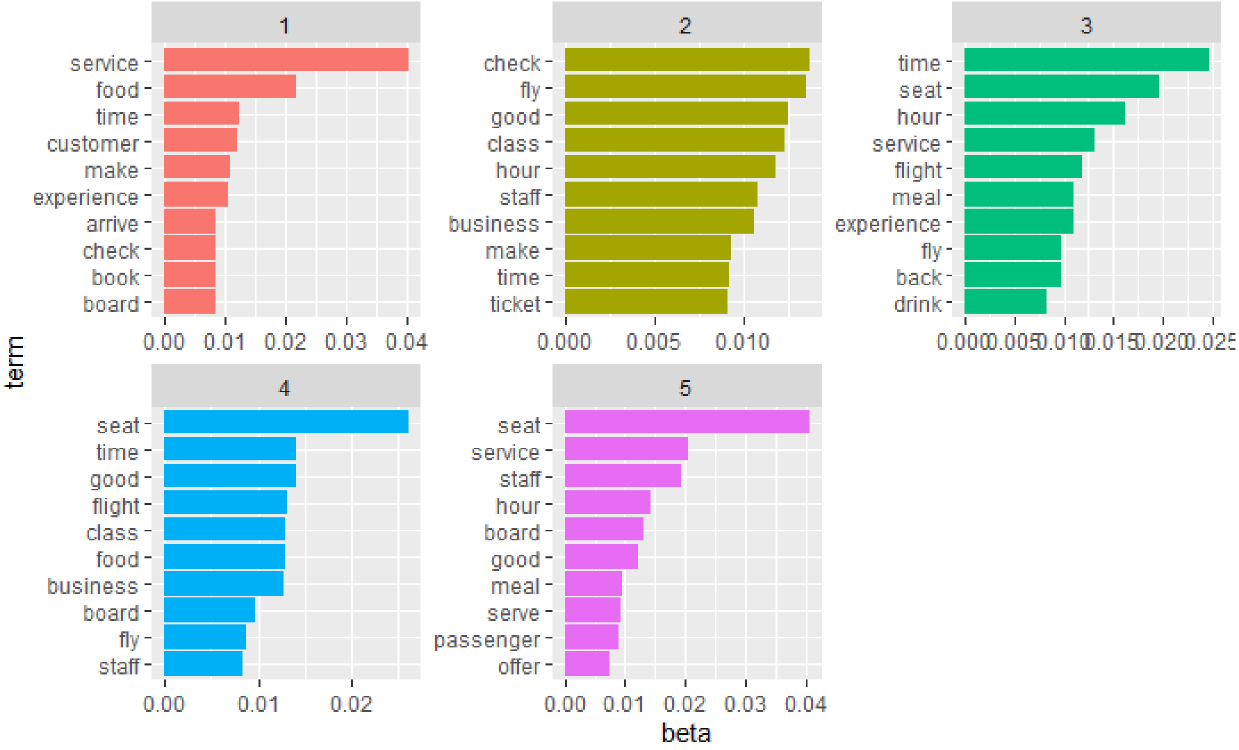
After categorizing the reviews, an RNN model was trained to classify them. This was done in a series of steps. The first step was to create a word-embedded matrix that would be used in the RNN since the RNN takes only numerical values. Word embedding is a technique that creates a vocabulary from the available terms. In this case, a pre-trained word embedding model, GloVe, was used. This model takes a list of terms and creates a vocabulary where each term is assigned a list of weights based on its relationship with surrounding terms. Thus, a term such as "king" is most likely to be assigned a similar weight to a term such as "queen" as they are used in almost similar settings within a sentence. Using such a matrix in an RNN model ensures that it also takes advantage of the relationship between terms, as terms do not have much meaning on their own.

The RNN model was first trained using 70% of the data, with 15% being used to validate the model. It was later tested using the remaining 15%. As the results show, RNN had a relatively low accuracy during training as it averaged around 55% for 100 epochs. However, as the below graph shows the model’s accuracy increased as the loss decreased with more epochs, which suggests potential for better accuracy. Following accuracy assessment of the predicted 15%, the model got an AUC value of 0.5221722, suggesting that it was quite close to random that correct.

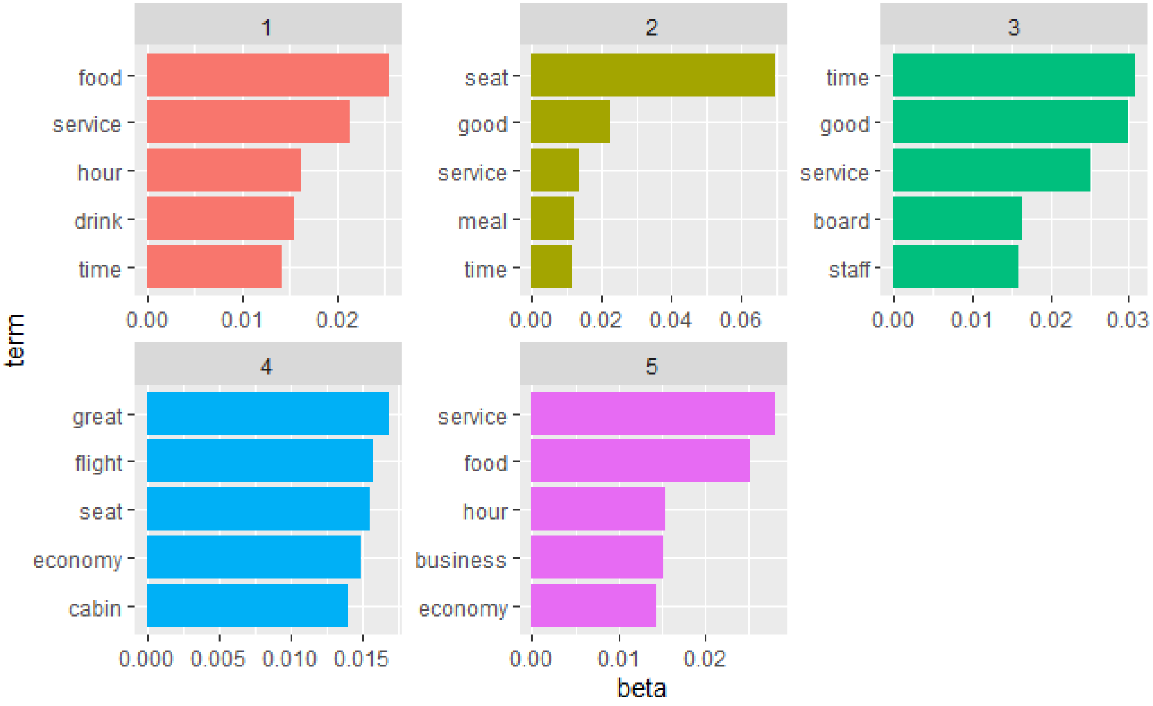


3.2.2.3. Topic Modelling using Latent Dirichlet Allocation (LDA)

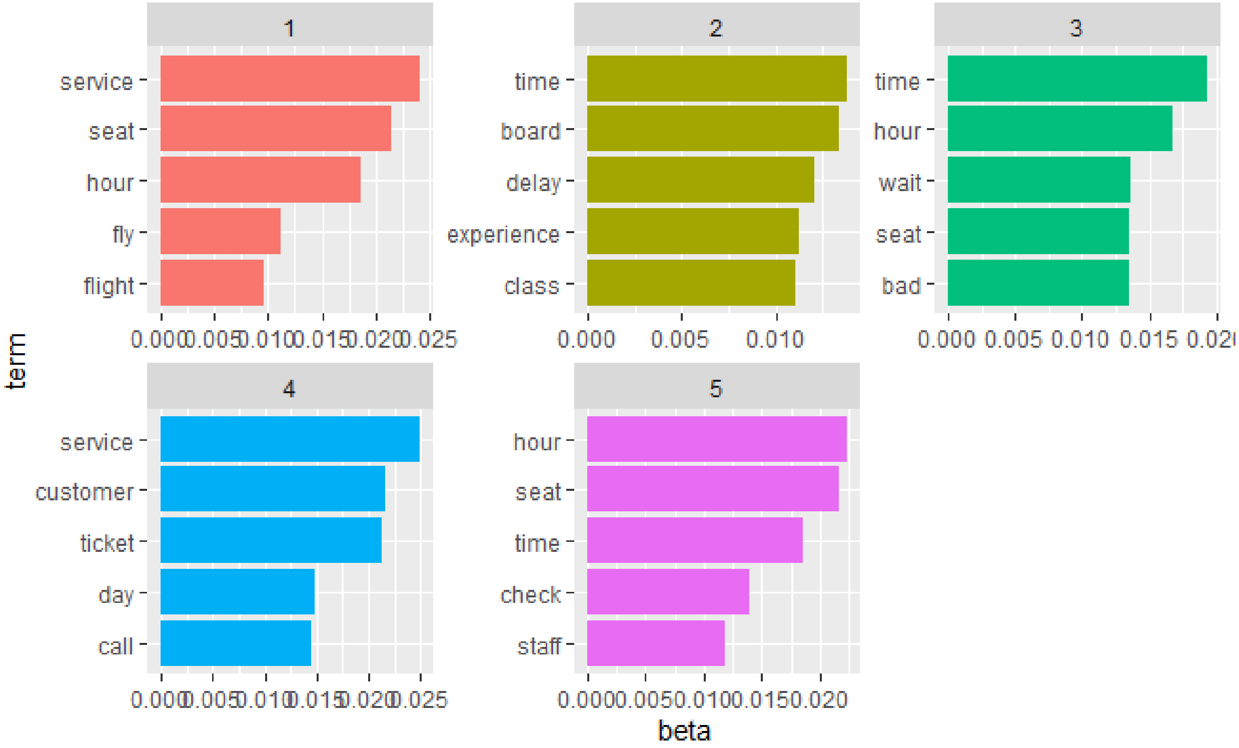
LDA is an unsupervised ML technique used for topic modeling. It discovers thematic structures in large collections of text, such as reviews, to identify abstract topics. LDA is a powerful tool in NLP, and thus, it was utilized in this project to investigate the main topics of discussion. The figure below shows the top 10 terms in the five topics identified. From these terms, it can be inferred that topic 1 is largely on service and food, topic two on value for money since there is a focus on terms that discuss pricing, topic three on time and comfort, topic four also on comfort and value for money, and topic five on comfort and service.



Partitioning the terms based on the categorized positive and negative reviews shows that the major topics that appear in positive reviews are food, service, and comfort. There is also a mention of economy and business in the last two of the top five topics, suggesting that there is indeed some satisfaction with value for money, as shown in the bar plot figures below.



As for negative reviews, as shown below, service dominates, suggesting that most passengers are not satisfied with the services they get. Time is also another key topic, suggesting a lot of delays and long waiting hours.



4. Conclusions

It is evident that the ten airlines experienced varying ratings, with Turkish Airlines having the poorest ratings while Eva Air has the best ratings. After categorizing sentiments, it is clear that a majority of reviews are positive; however, there is a substantial number of displeased travelers. In the review categorization into the three sentiments, it is clear that using NRC might be challenging as it may not accurately capture the sentiment in each word. This might be due to its reliance on individual word values. In the classification using RNN, the accuracy was quite lower than expected. This could be attributed to low computing power, as the model had limited hidden layers. Despite the ease of using GloVe in word embedding, it did not provide accurate results, making it difficult to assess its accuracy. In topic modeling, LDA proves to be crucial in identifying themes and topics of discussion. Most importantly, the findings suggest the need for more attention towards services and time among airlines as these dominate in negative reviews.

5. Recommendations

The project would have benefited from a more robust computing system, especially for RNN, where it was limited to a few layers and smaller vectors. Thus, a model that maximizes on a larger data set could improve the RNN results. Additionally, this project was done using RStudio. There are alternative programming languages, such as Python, which could ease computing, such as Keras and Tensorflow. The libraries that RNN depends on require virtual Python installations, which can be time-consuming. Furthermore, cloud computing promises to ease local computation power, which researchers should take advantage of when carrying out computationally intensive analyses such as neural networks.