Predicting sentiment from tweets about US airlines

Jenny Rhee

1. Introduction

Airlines in the United States handle approximately 10 million flights containing one billion passengers per year (Source). It is important for airlines to understand how they are performing with customer service to prevent losing consumers to competing airlines. However, the proportion of people who actually submit customer feedback is generally small with respect to the size of the actual customer base. With the growing ubiquity of social media in recent years, specifically Twitter, it has become a valuable source of data for companies to receive more frequent customer feedback.

Building a model that can predict sentiment from tweets about airlines is beneficial for various airlines to understand how they are doing overall with customer service, as well as pinpoint areas of strength and weakness. This is especially important for struggling airlines that could benefit from a tool that helps them understand their customer service faults in order to explicitly try to improve them. The goal of this project is to train a model to predict sentiment and develop a web application that can be used to monitor the current sentiment of various airlines day-to-day and display representative tweets.

2. Data

The data comes from Figure Eight, who provides high quality training data to solve various machine learning problems, and can be found here. The relevant columns in this particular data set are airline sentiment, negative reason, airline, Twitter handle, retweet count, tweet content, tweet timestamp, and user timezone. Once the model is trained, Twitter data will come from the API to determine the current sentiment of different airlines.

2.1. Data Wrangling Summary

Text data can be very messy, especially when it comes from a social media platform such as Twitter. It generally includes typos and characters or words that are not beneficial to analyzing the data. It is also important to find and filter any non-English tweets that can negatively impact the training of a model, but this particular data set contained tweets that are all in English. First, duplicate tweets (i.e., retweets) were removed. Each tweet was processed to remove URLs, mentions, punctuation, numbers, whitespace, and stop words (i.e., words that are not meaningful to the context of a sentence). Finally, the sentences were broken up into tokens and stemmed (i.e., reduces inflectional forms to its word stem).

Machine learning models can only take numerical data as inputs so the final cleaned data must be converted from text into numbers. There are several approaches to this, but the method used for this particular project was term frequency-inverse document frequency (tf-idf), which essentially relates how often a word appears in a document with respect to how often a word appears in the entire corpus. Additional features were created from the original tweets: the total character count, the number of capital letters, the capital letter to character count ratio, the number of words, the number of happy emoticons, the number of sad emoticons, the number of exclamation marks, and the number of question marks.

3. Exploratory Analysis Summary

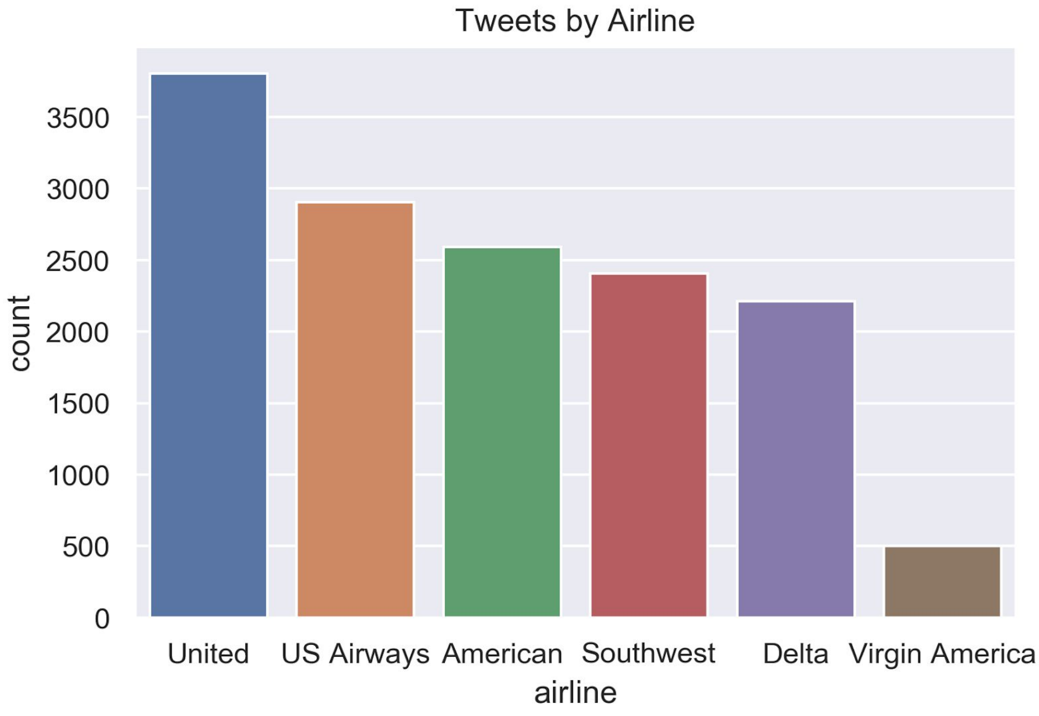


Figure 1. Tweet counts by airline.

Most of the tweets in this data set were directed towards United, whereas the least amount of tweets were directed towards Virgin America (Figure 1). Considering the top airlines by passengers carried, American Airlines, Delta, Southwest, and United were ranked top 4 in North America (Source). Additionally, American Airlines and US Airways merged in 2015, which was why US Airways does not appear in the previously mentioned list (Source). Interestingly, United had the most number of tweets although they were the 4th busiest airline by passengers carried, while American Airlines, Delta, and Southwest had similar numbers of tweets to passengers carried ratios.

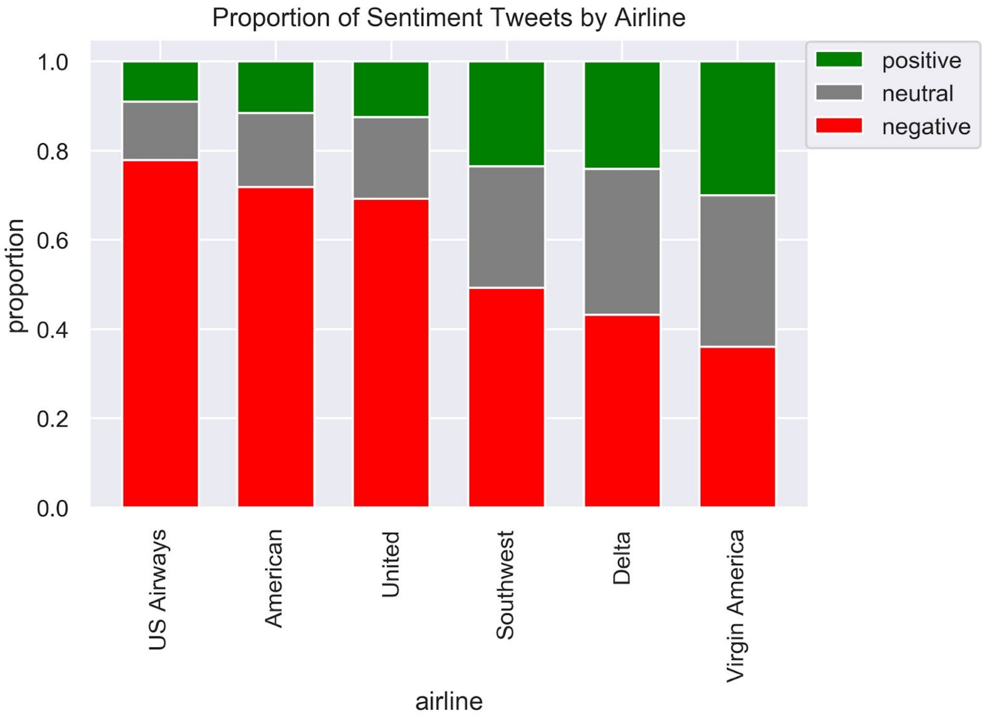


Figure 2. Sentiment percentages by airline.

Unsurprisingly, the general trend appeared to be that most people will tweet about an airline for negative reasons (Figure 2). The airlines that received the most negative criticism in this data set were US Airways, American Airlines, and United. This aligned with this Consumer Reports survey from 2018, where Virgin America and Southwest were among the highest ratings, and United and American Airlines (again, merged with US Airways before this survey) were among the lowest ratings. Around 70-80% of tweets about American Airlines, US Airways, and United were negative, but United had 31% more tweets than US Airways and 47% more tweets than American Airlines. Overall, positive and neutral tweets were much less common, which could potentially lead to some challenges when training a model from imbalanced classes.

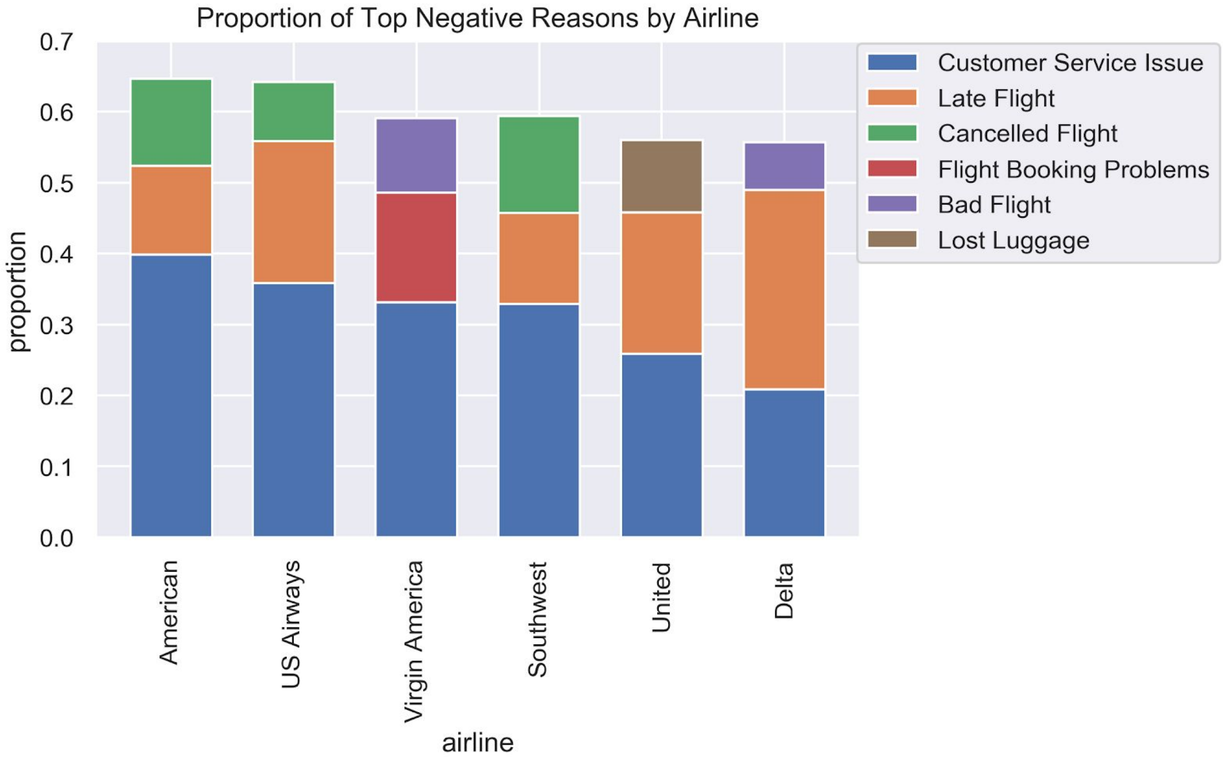


Figure 3. Top 3 negative reasons for each airline. Tweets, where the negative reason could not be identified, were excluded.

For all airlines, customer service issues were a consistent problem (Figure 3). However, it should be noted that it was not the most frequent negative reason for Delta, which was late flights. Almost every airline (excluding Virgin America) had their second highest amount of complaints due to late flights. America, Southwest, and US Airways also had common complaints regarding cancelled flights. Virgin America and Delta differed from the other airlines because their third top reason was about bad flights. One example for this category was, “@VirginAmerica I flew from NYC to SFO last week and couldn't fully sit in my seat due to two large gentleman on either side of me. HELP!” A second example was, “@JetBlue someone should screen what you play on your flights.” Other notable differences were that United also suffers from lost luggage complaints, and Virgin America customers experienced flight booking problems.

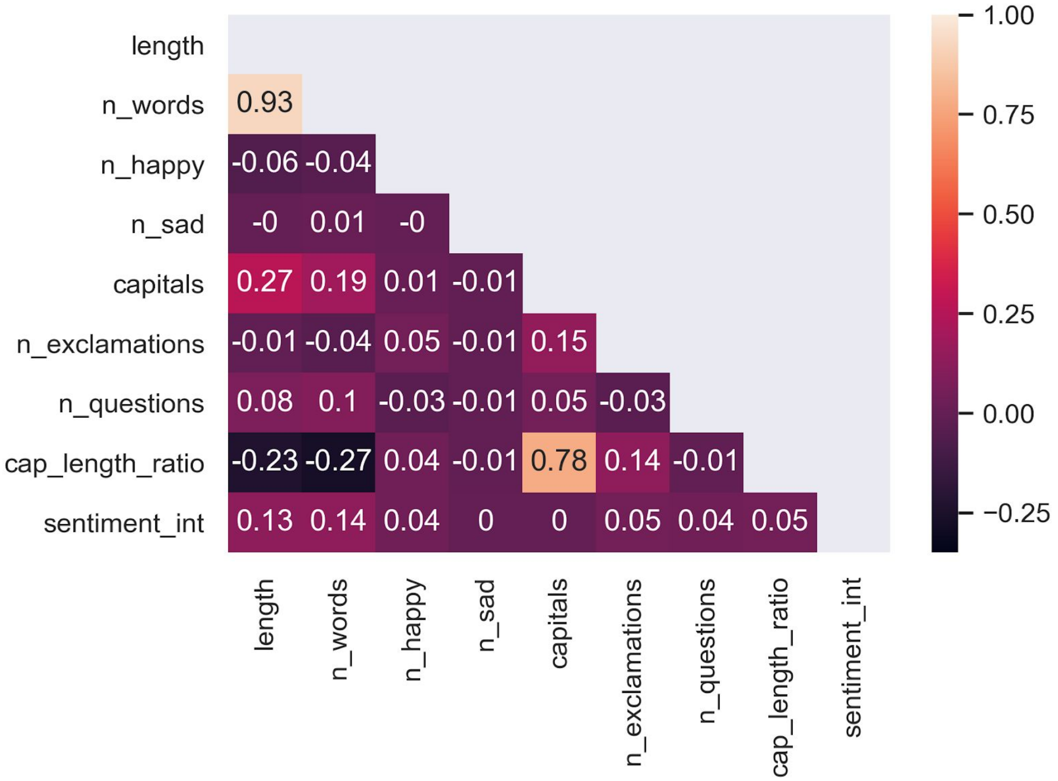


Figure 4. Correlation matrix.

The correlation values for sentiment_int vs. the other features were calculated using the correlation ratio, which ranges from 0 to 1 (Figure 11). Most of the predictive power of the model will likely come from the vectorized vocabulary, but it was interesting to see a small correlation between sentiment and length/number of words. As we'll see below, this is mostly due to a relationship between number of words and negative sentiment.

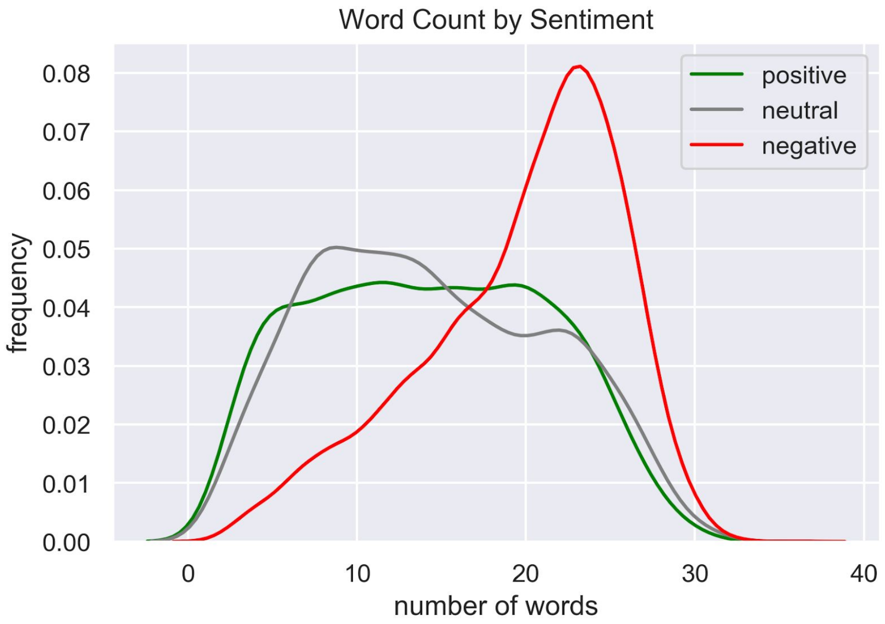


Figure 5. Word count distributions for each sentiment

Positive and neutral tweets had a wide distribution of word counts, but negative tweets were consistently longer (Figure 5). It was interesting to see such a distinct difference in word counts between negative tweets and other tweets, but it wasn't very surprising; people generally have more to say when they're upset.

ANOVA was then used to test whether the differences in mean number of words between sentiment were statistically significant and Tukey's tests to determine statistical significance of differences between individual pairs. On average, negative tweets were 5.24 words longer than neutral tweets and 5.42 words longer than positive tweets.

One-way ANOVA Test:

F-value = 1190.3

p-value < 0.001

Group 1	**Group 2**	**Mean Difference**	**Lower**	**Upper**
negative	neutral	-5.24	-5.55	-4.93
negative	positive	-5.42	-5.77	-5.08

Table 1. Tukey's test results for number of words (significant only).

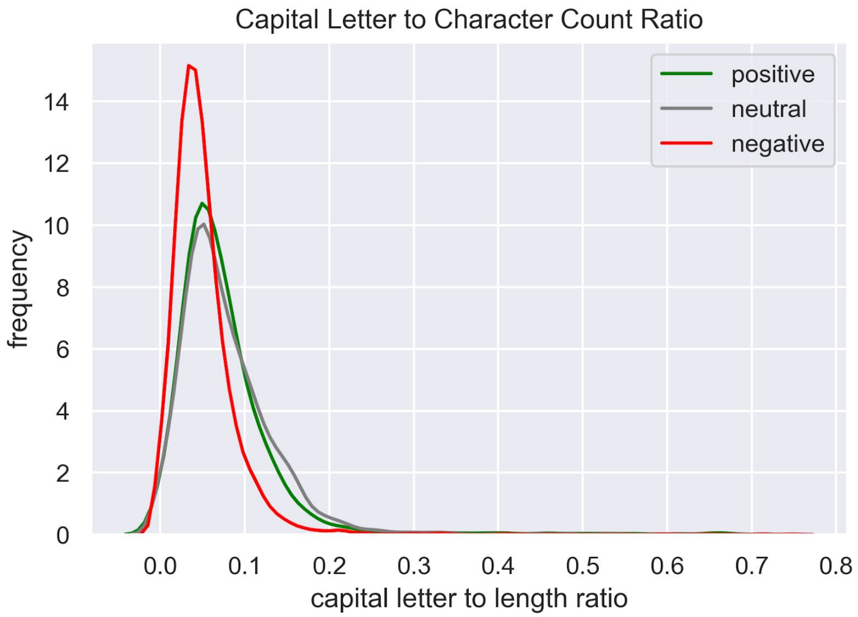


Figure 6. Capital letter to character count ratio distributions for each sentiment.

On the other hand, positive and neutral tweets had a higher average capital letter to character count ratio (Figure 6). The largest average difference was between negative and neutral tweets (Table 2). Neutral tweets had a 0.026 higher capital letter to character count ratio than negative tweets, and positive tweets had a 0.023 higher ratio than negative tweets. For positive tweets, some people tend to type in all caps when they're excited. An example of a positive tweet with a higher ratio was, "@SouthwestAir @love_dragonss LAUREN OMG BEST AIRLINE EVER." Additionally, there are a lot of abbreviations in this industry (i.e. flight numbers, airports), and neutral tweets may contain a lot of these abbreviations when seeking information. An example of a neutral tweet with a higher ratio was, "@united PTY to PIT via IAH." There were some negative tweets that had a high frequency of capital letters when a customer was exceptionally angry, such as, "@united @getmeontop 7 WEEKS Late FlightR AND I STILL HAVE NOT RECEIVED MY MILES FROM THE MileagePlus Gift Card $150 STARBUCKS CARD I HANDED OVER!!!", but these seem to happen less often.

One-way ANOVA Test:

F-value = 375.2

p-value < 0.001

Group 1	**Group 2**	**Mean Difference**	**Lower**	**Upper**
negative	neutral	0.026	0.023	0.028
negative	positive	0.023	0.020	0.026

Table 2. Tukey's test results for capital letter to character count ratios (significant only).

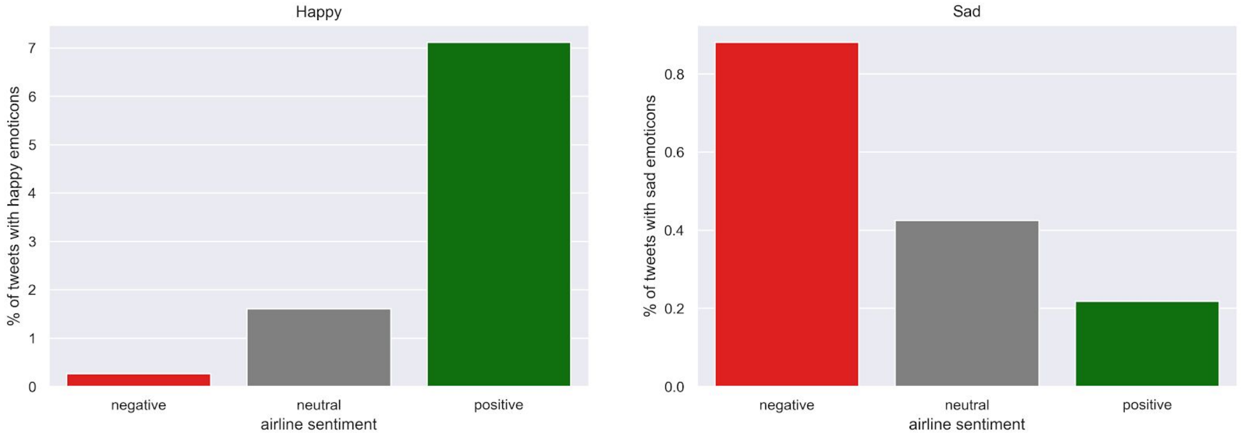


Figure 7. Percentage of tweets for each sentiment that contains happy emoticons and sad emoticons. Note the differences in the y-axis for each graph.

As expected, a higher percentage of positive tweets contained happy emoticons, and a higher percentage of negative tweets contained sad emoticons (Figure 7). There were significant differences in the number of happy emoticons between each sentiment (Table 3). The largest difference was between negative and positive tweets, where positive tweets had 0.071 more happy emoticons on average.

One-way ANOVA Test (Happy):

F-value = 277.4

p-value < 0.001

Group 1	**Group 2**	**Mean Difference**	**Lower**	**Upper**
negative	neutral	0.013	0.007	0.020
negative	positive	0.071	0.064	0.078
neutral	positive	0.057	0.049	0.066

Table 3. Tukey's test results for number of happy emoticons

There were significant differences in the number of sad emoticons between negative and neutral tweets, as well as negative and positive tweets (Table 4). The largest difference was between negative and positive tweets, where negative tweets had 0.007 more sad emoticons on average.

One-way ANOVA Test (Sad):

F-value = 277.4

p-value < 0.001

Group 1	Group 2	Mean Difference	Lower	Upper
negative	neutral	-0.005	-0.009	-0.001
negative	positive	-0.007	-0.012	-0.002

Table 4. Tukey's test results for number of sad emoticons (significant only).

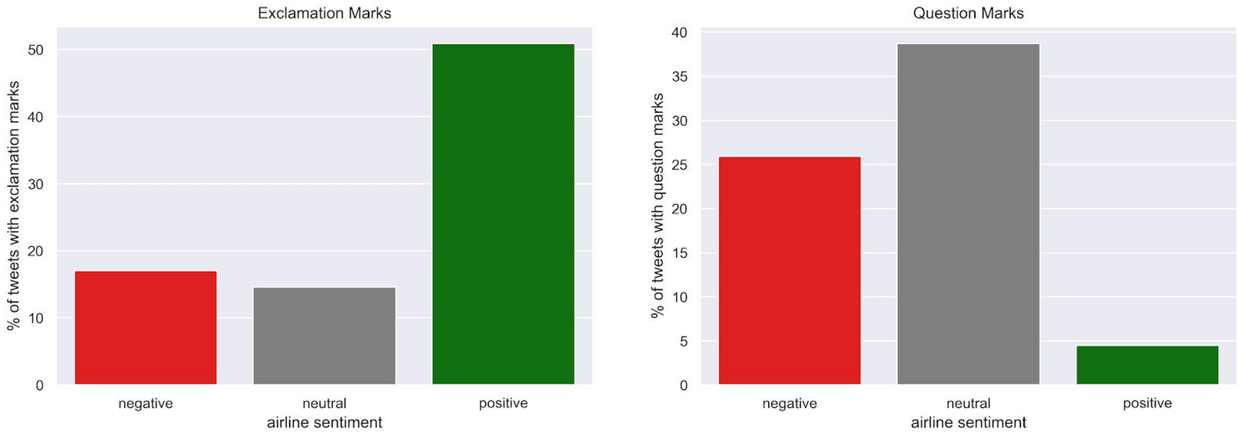


Figure 8. Percentage of tweets for each sentiment that contains exclamation marks and question marks. Note the differences in the y-axis for each graph.

Notably, over 50% of positive tweets contained exclamation marks, whereas less than 20% of negative and neutral tweets contained them (Figure 8). There were statistically significant differences between every sentiment (Table 5). Positive tweets, on average, had 0.557 more exclamation marks than negative tweets and 0.627 more exclamation marks than neutral tweets. Generally, exclamation marks can be associated with happiness or intense anger, but it looked like in this case, they were used primarily in happy tweets.

One-way ANOVA Test (Exclamations):

F-value = 401.3
p-value < 0.001

Group 1	Group 2	Mean Difference	Lower	Upper
negative	neutral	-0.070	-0.114	-0.026
negative	positive	0.557	0.508	0.606
neutral	positive	0.627	0.569	0.685

Table 5. Tukey's test results for number of exclamation marks.

Close to 40% of neutral tweets contained question marks, compared to 26% and almost 5% of negative and positive tweets, respectively. The average neutral tweet had 0.106 more question marks than negative tweets and 0.4 more question marks than positive tweets (Table 6). This made sense since it indicated that neutral tweets may contain a fair amount of questions, which tend to be neutral with the goal of obtaining new information. It was also interesting to see that negative tweets had an average of 0.295 less question marks than positive tweets.

One-way ANOVA Test (Questions):

F-value = 264.5
p-value < 0.001

Group 1	Group 2	Mean Difference	Lower	Upper
negative	neutral	0.106	0.074	0.138
negative	positive	-0.295	-0.330	-0.259
neutral	positive	-0.400	-0.443	-0.358

Table 6. Tukey's test results for number of question marks.

3.1 Most Predictive Words By Sentiment

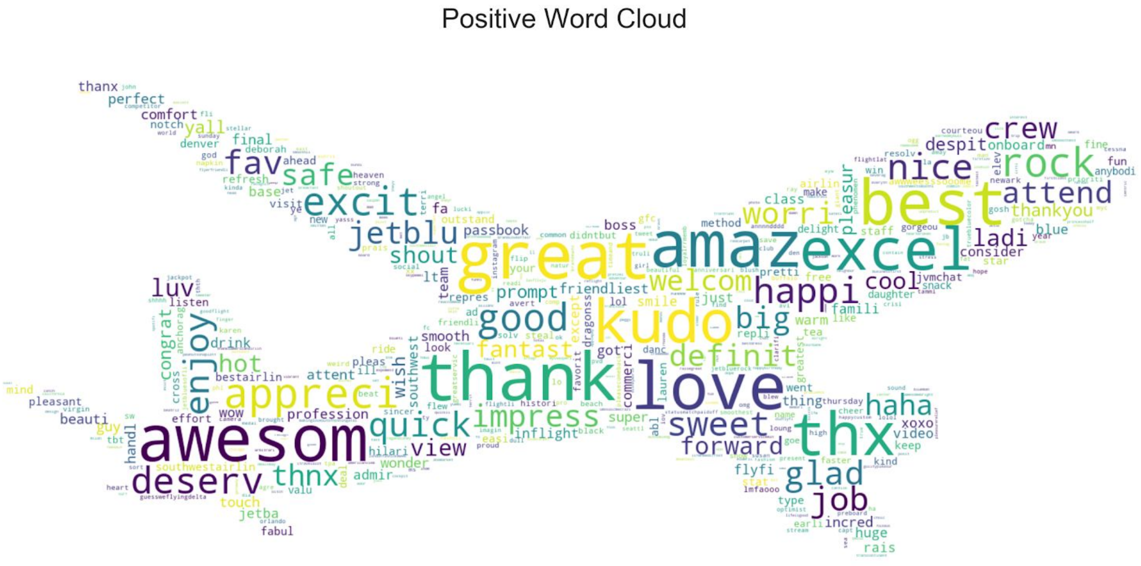


Figure 9. Word cloud from positive tweets based on predicted probabilities.

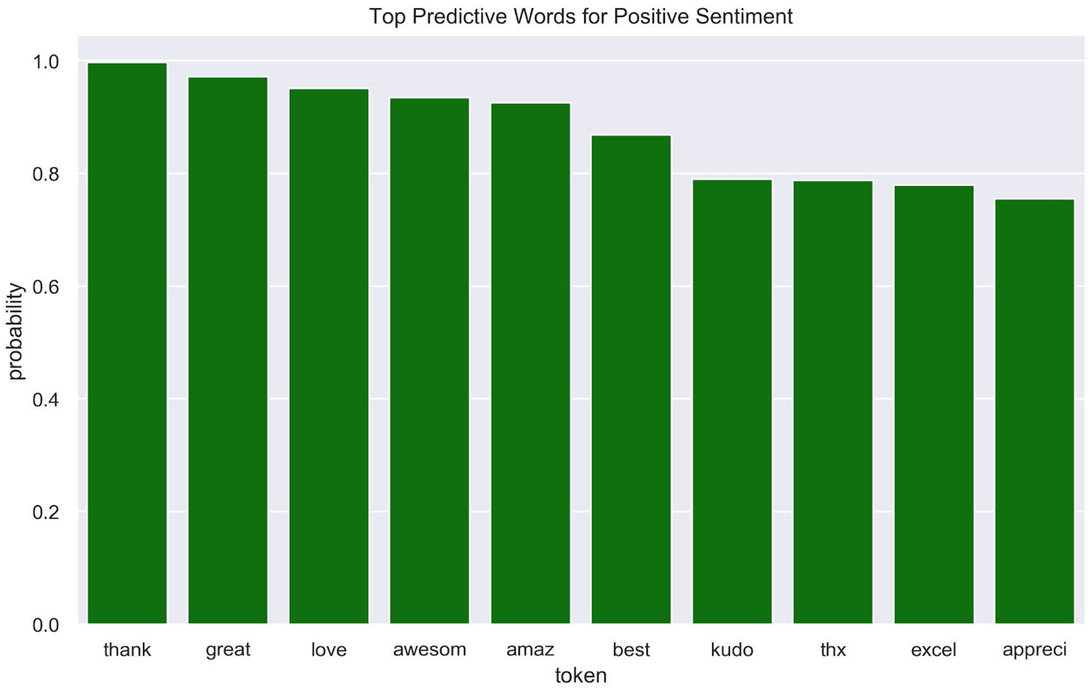


Figure 10. Most predictive words for positive sentiment.

A word cloud with all words and probabilities for positive sentiment can be seen in Figure 9. Thank, great, love, awesom, and amaz were the top 5 most predictive tokens for positive sentiment (Figure 10).

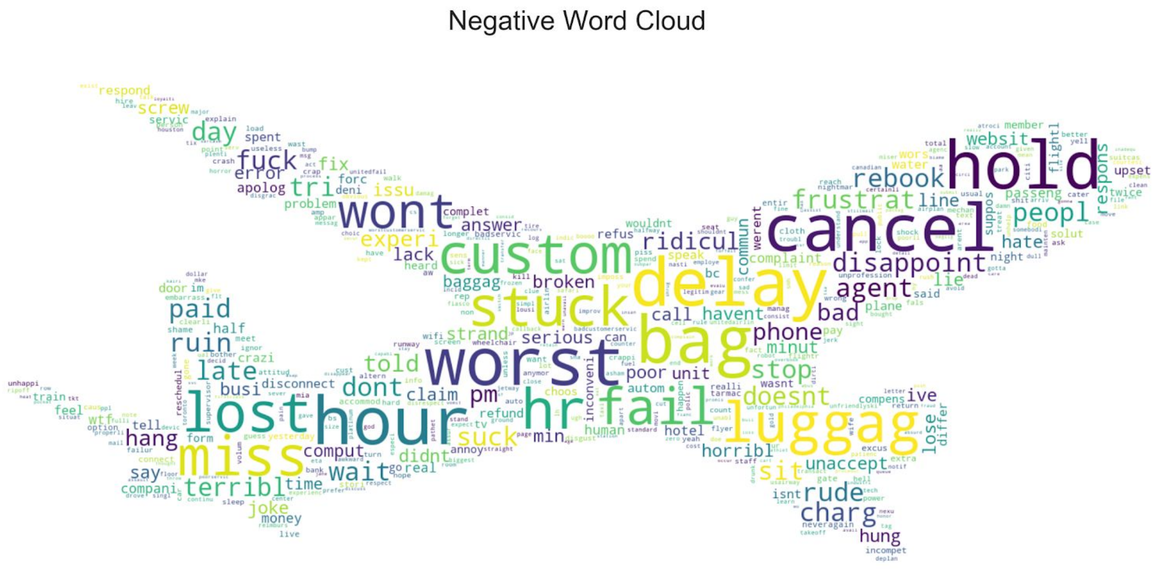


Figure 11. Word cloud from negative tweets based on predicted probabilities.

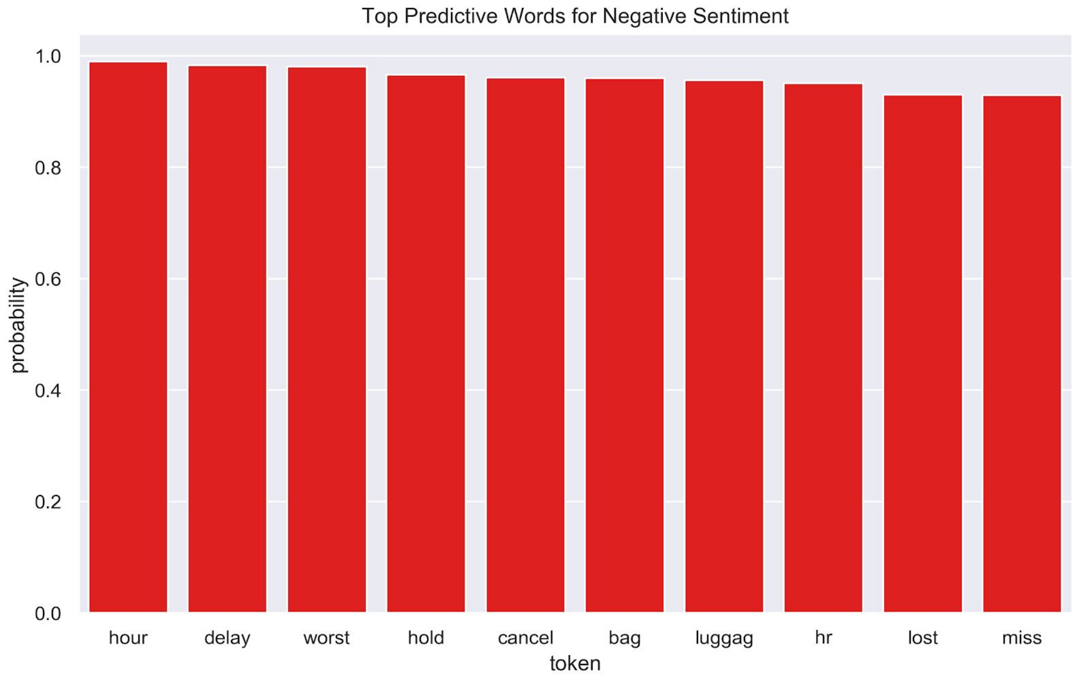


Figure 12. Most predictive words for negative sentiment by airline.

A word cloud with all words and probabilities for negative sentiment can be found in Figure 11. Hour, delay, worst, hold, and cancel were the top 5 most predictive tokens for negative sentiment (Figure 12).

4. Sentiment Model

This modeling task was to predict positive, neutral, or negative sentiment from tweets about airlines, which is a multiclass classification (supervised learning) problem. The majority of the features were tokens based on the text, but there were also numerical features based on word length, character count, number of sad/happy emoticons, etc.

Two classification algorithms were used to compare performance with tf-idf vectorizer: Naive Bayes and random forest. Naive Bayes performed significantly worse than random forest; the mean accuracy score for Naive Bayes was 0.679 versus 0.763 for random forest. The bag-of-words model was then compared to tf-idf, which had a mean accuracy score of 0.757 vs. 0.763 for tf-idf.

4.1 Final Model

The hyperparameters for the final model were optimized using randomized search. The best performing parameters for tf-idf were ngram range = (1, 2) and max_df = 0.6. The best performing parameters for random forest were n_estimators = 1500, min_samples_split = 0.001, max_features = log2, and class_weight = balanced_subsample. The final mean accuracy score was 0.783, and the ROC-AUC score was 0.889.

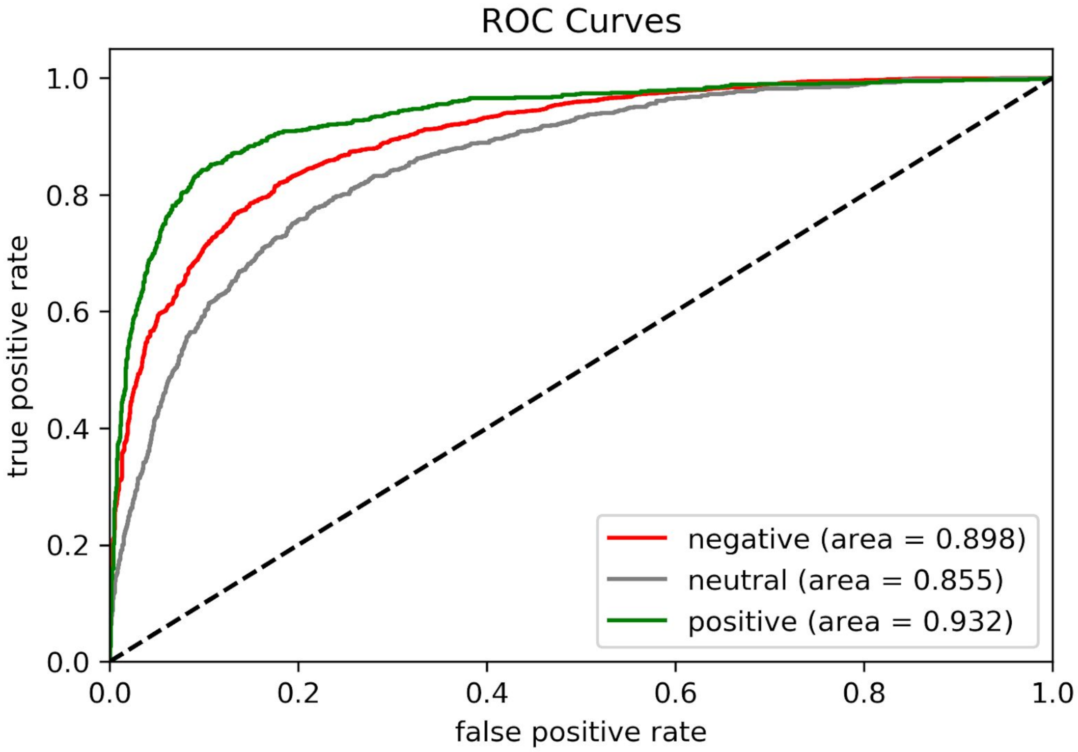
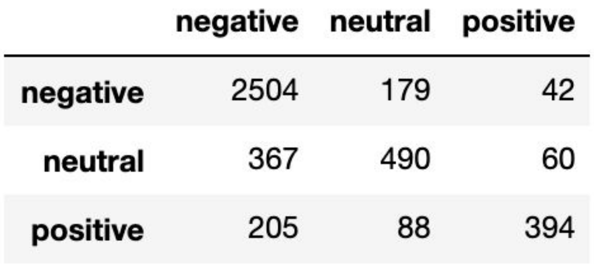


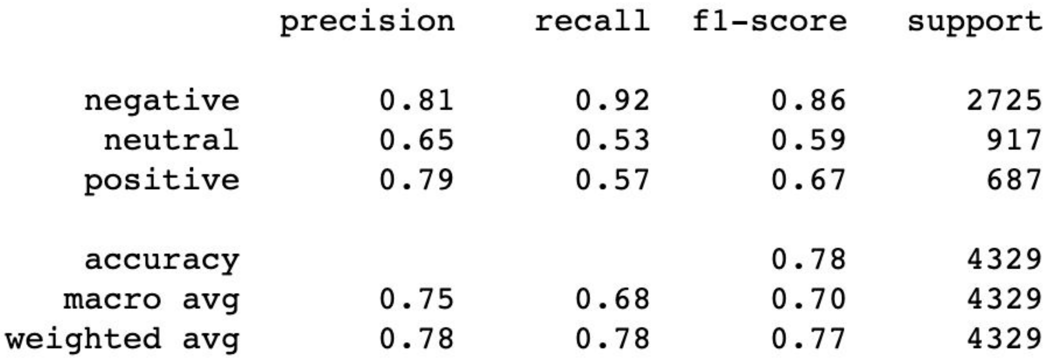
Figure 13. ROC curves for each sentiment class

The ROC curves can be found in Figure 13; positive sentiment had the highest AUC (0.932) while neutral had the lowest (0.855). In other words, the model is better at distinguishing positive and negative sentiment, which is ideal because those are generally the most interesting and important tweets. As we've seen, neutral tweets tend to be more about information gathering, but the purpose of this model is to understand the positive or negative performance of airlines.



	negative	neutral	positive
negative	2504	179	42
neutral	367	490	60
positive	205	88	394

Table 2. Confusion matrix



```
              precision    recall  f1-score   support

    negative       0.81      0.92      0.86      2725
     neutral       0.65      0.53      0.59       917
    positive       0.79      0.57      0.67       687

    accuracy                           0.78      4329
   macro avg       0.75      0.68      0.70      4329
weighted avg       0.78      0.78      0.77      4329
```

Table 3. Classification report

The confusion matrix from the final random forest model can be found in Table 1, and the classification report can be found in Table 2.

5. Recommendation

The random forest performed the best with mean accuracy score = 0.763 vs. Naive Bayes' mean accuracy score = 0.679. The final mean accuracy score for the random forest after hyperparameter tuning was 0.783, and the ROC-AUC score was 0.889. I would recommend implementing a random forest classification model and using this to monitor each respective airline's overall sentiment. The tweets that are classified for each sentiment can be used to see what the airline is doing correctly and what they are doing poorly.

6. Conclusion & Future Direction

Sentiment analysis about tweets directed towards airlines demonstrated that the opinions people share on Twitter are generally negative. Unsurprisingly, the frequency of negative tweets per airline resembles the results from past surveys. Most issues relate to customer service or late flights, which airlines may not always have control over. However, examining negative tweets about customer service will be useful for airlines to pinpoint what is going wrong. The next steps for this model is to create a web application that airlines can use to interact with the model to view recent tweets and compare sentiment about themselves, as well as their competition.