Customer Experience Indicator: Twitter Sentiment Analysis

Capstone Project on Tweets about US Airlines

Background:

In 2021, 192 million daily active users are on Twitter and over 330 million users log onto the site at least once per month. Social Media has become an increasingly popular venue for customers to share their negative (or positive) customer experiences and link this to the company via their corporate social media accounts. The potential for negative publicity if a customer's tweet goes viral is a real risk for a company's reputation.

Sentiment Analysis is a Natural Language Processing (NLP) technique which derives the attitude of the writer on a particular topic from a written or spoken source. It is becoming a popular tool for companies to better understand the real-time feelings of their customers about a topic, product, or experience from social media posts. Using this mined information allows for service recovery efforts as well as opportunities for improvement and targeted marketing for customers.

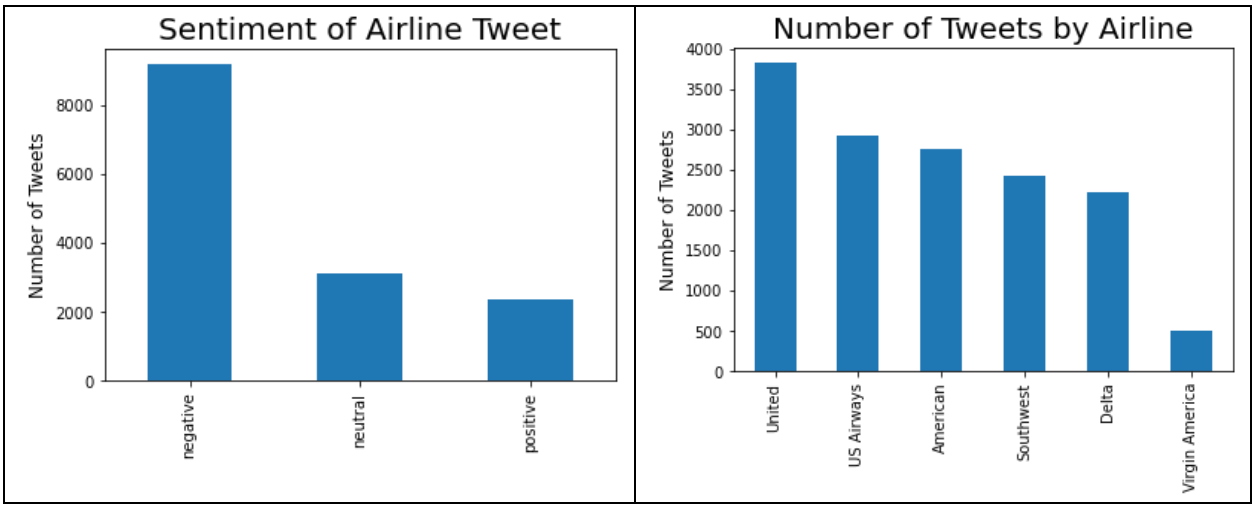
Problem Statement

Airlines operate in a challenging environment where each customer's experience is influenced by factors that are within and out of the airline's control: weather, plane cleanliness, space for overhead baggage, changing regulations, cancellations, gate changes, and efficiency and empathy of customer service to name a few. This is made even more difficult because they have employees operating all over the world so creating and maintaining a consistent experience requires constant vigilance, keeping a pulse on what is happening both globally as well as locally.

To better understand customer's experience in real-time this project aims to predict the sentiment of tweets that reference a particular airline. If there is a high success rate it will allow airlines to have a real-time view into customer experience, identify the specific factors related to a negative or positive experience and allow them to continually improve their operations.

Data Overview:

The data for this project was scraped from Twitter in February of 2015. Individual contributors were then asked to review each tweet and classify it as positive, negative, or neutral. If the tweet was identified as negative, they were asked to categorize into a negative reason code. The dataset contains about 14.5K individual tweets referencing 6 major US airlines.



Data Features

The data features can be split into two different categories: features related to the tweet and tweet sentiment categorization features. To allow for reproducibility on new datasets, none of the tweet categorization features were used in the model. They did aid data exploration and analyzing predictions.

Tweet Information

Below is a summary of the information included about each tweet. The location data was of inconsistent quality and so was not included in the model but in the future could be valuable to understanding where the biggest opportunities are.

	text	airline	tweet_coord	tweet_created	tweet_location	user_timezone
count	14640	14640	1019	14640	9907	9820
unique	14427	6	832	14247	3081	85
top	@united thanks	United	[0.0, 0.0]	2015-02-24 09:54:34 -0800	Boston, MA	Eastern Time (US & Canada)
freq	6	3822	164	5	157	3744

Percent of Missing Information by Feature

```
tweet_id          0.000000
retweet_count     0.000000
text              0.000000
airline           0.000000
tweet_coord       0.930396
tweet_created     0.000000
tweet_location    0.323292
user_timezone     0.329235
```

Tweet Sentiment Categorization

Airline sentiment was the predicted variable. There were three different categories: negative, neutral, or positive. The sentiment and reason for each negative rating was coded by individual contributors and the confidence they had for their rating is included as well.

	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence
count	14640	14640.000000	9178	10522.000000
unique	3	NaN	10	NaN
top	negative	NaN	Customer Service Issue	NaN
freq	9178	NaN	2910	NaN
mean	NaN	0.900169	NaN	0.638298
std	NaN	0.162830	NaN	0.330440
min	NaN	0.335000	NaN	0.000000
25%	NaN	0.692300	NaN	0.360600
50%	NaN	1.000000	NaN	0.670600
75%	NaN	1.000000	NaN	1.000000
max	NaN	1.000000	NaN	1.000000

Percent of Missing Information by Feature

```
airline_sentiment              0.000000
airline_sentiment_confidence   0.000000
negativereason                 0.373087
negativereason_confidence      0.281284
```

Data Wrangling

Sentiment analysis requires Natural Language Processing (NLP) techniques to format text data for a machine learning model to interpret. Tweets require a unique approach because they use informal language and heavily rely on emojis and hashtags to convey emotion. The overall approach in this project was to reduce tweets into individual tokens (words, emojis, hashtags) and then identify the most frequently used tokens. Once the model has this information it can make associations between tokens and the sentiment that they convey.

Tweet Cleaning

The first step was to convert the tweets to lower case to ensure that regardless of the capitalization of a word, it is categorized as the same word. Next, all "@" mentions were removed from the tweets. In certain situations, these could be valuable to a model but the vast majority either indicated the airline (information already included in the dataset) or were not repetitious enough to be included in the final model dataset. Ae conscious decision was made to remove all hyperlinks because they were masked and therefore did not provide meaningful information about what they linked to. Finally, all digits were removed from the text.

Tokenization & Dimensionality Reduction

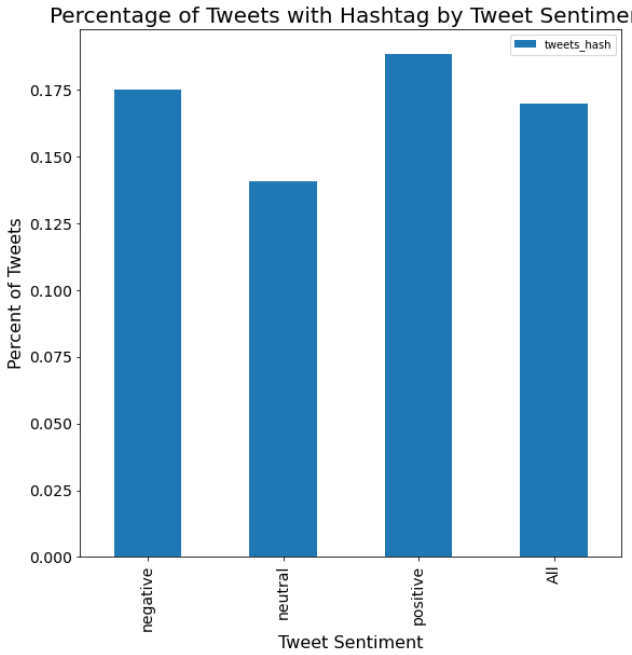
Tokenization is the process that allows each individual token to be separated in the text string. Multiple tokenizers exist. The Tweet Tokenizer from the Natural Language Toolkit (NLTK) was chosen because it has a built-in tool that allows it to maintain emojis and hashtags which are crucial when working with a Twitter dataset. Once the tokenization was complete standalone punctuation was removed as were

common words, also known as stopwords to declutter the corpus. Upon completion of the first pass of tokenization and tweet cleaning there were almost 12,500 unique tokens from the dataset.

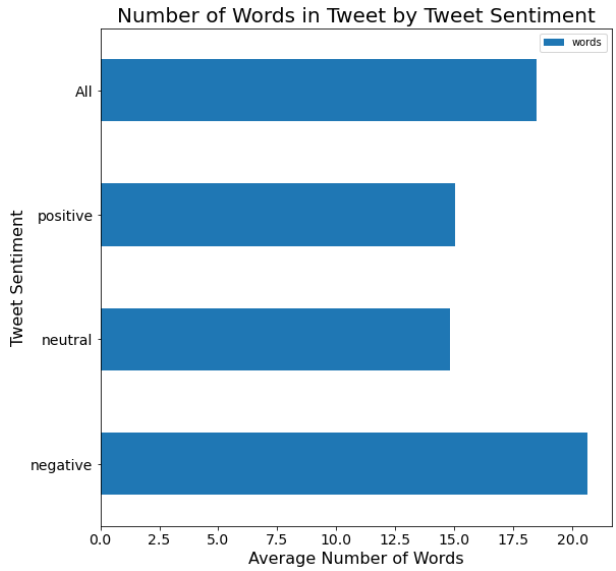
A technique to further reduce the number of words without losing the associated sentiment of them is to use stemming or lemmatization. Stemming is rule-based approach that truncates words that all come from the same root word. For example – “play”, “player”, “played”, “plays” and “playing” would all become – “play”. Lemmatization results in a similar outcome but is based on a dictionary rather than rules and is a more conservative approach. Both approaches were tested and included in the model dataset.

Additional Features

A new feature was created that counted the number of hashtags included in a tweet. During the exploratory phase it was identified that positive tweets tended to include hashtags more frequently.

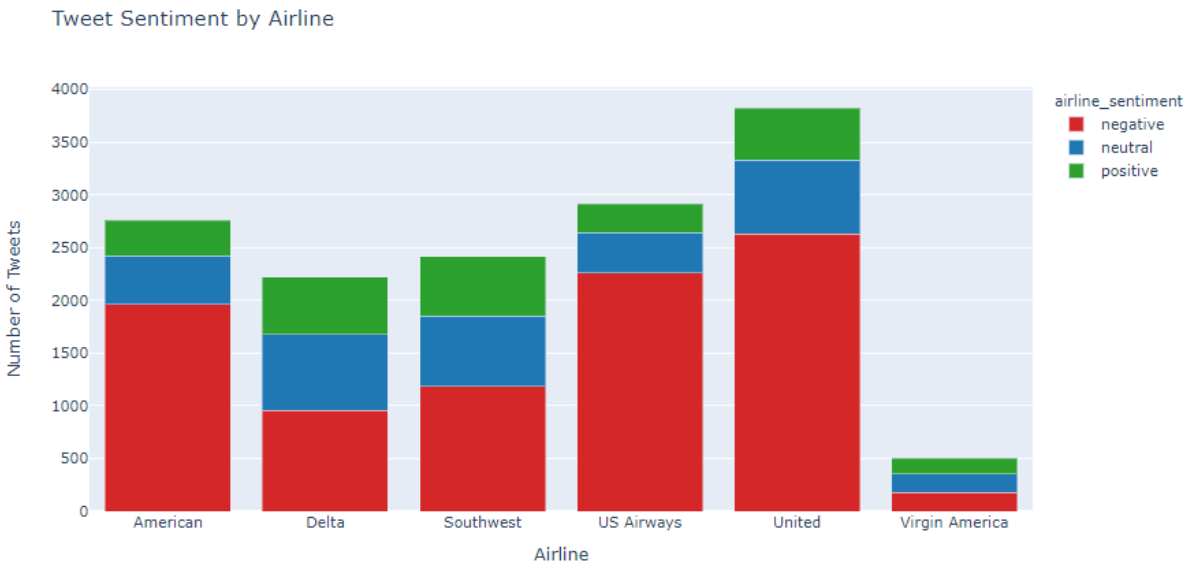


Similarly, a new feature was added to the dataset which is a count of the tokens included in the tweet. Through EDA it was identified that negative tweets tended to contain more tokens than positive or neutral ones.

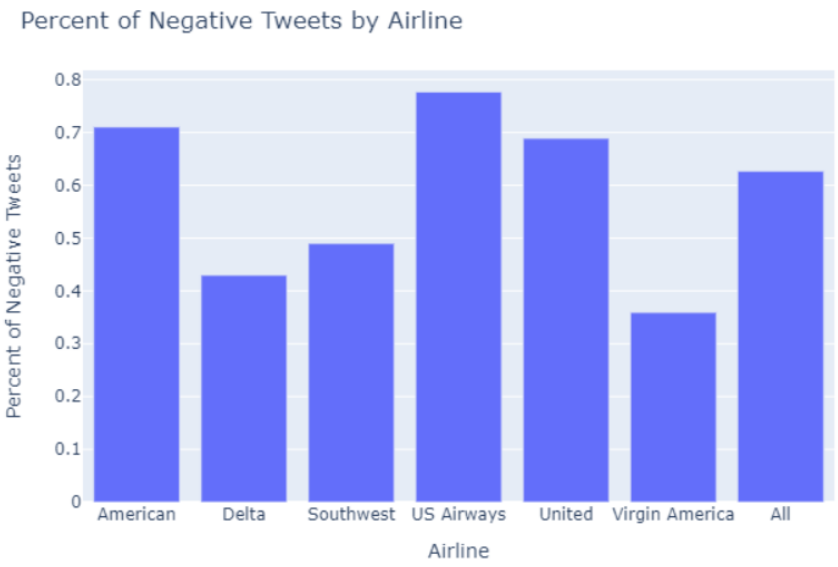


Exploratory Data Analysis

The purpose of exploratory data analysis is to gain a better understanding of the overall dataset and capture data patterns by visualizing the data. To better understand the makeup of the dataset each airline's tweets were broken up by sentiment.

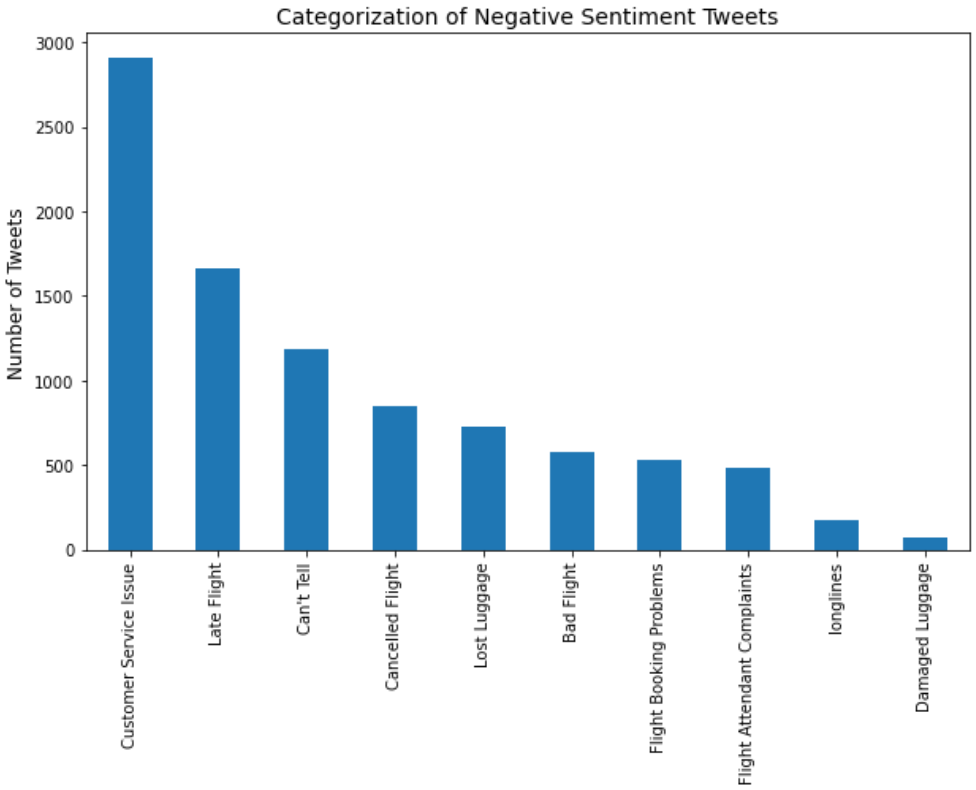


The number of tweets by airline was variable in the dataset as well as the makeup of those tweets. US Airways, American and United had the most tweets but also had a higher percentage of negative tweets than the other airlines.

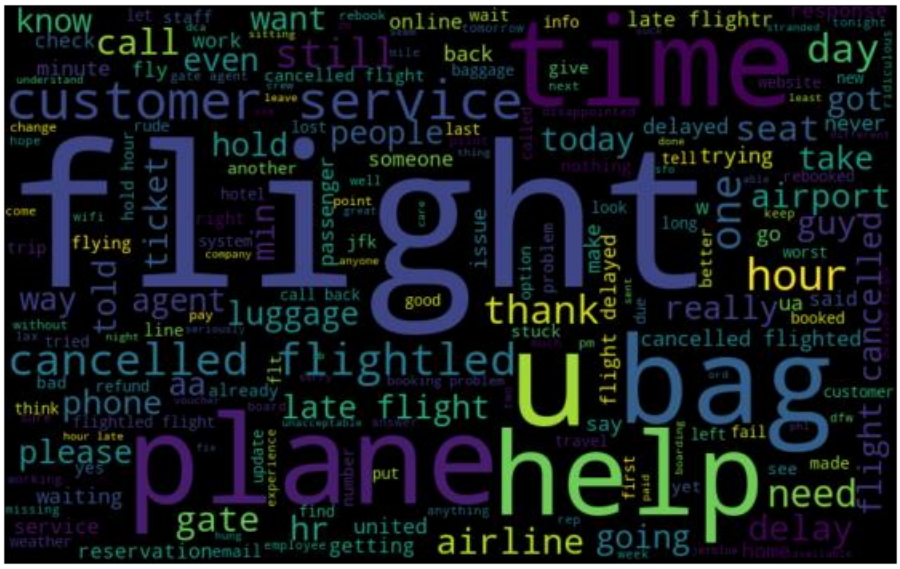


Negative Tweets

To create this dataset individual contributors tagged the sentiment of each tweet as well as categorized any negative tweets. Below is a summary of the categorization for the negative tweets. While this information was not used in the model, it does provide context to the reasons that customers were upset, and the themes of what words will be present in their tweets.



Word clouds are helpful a tool to visualize the data when using a text-based dataset. The corpus with word frequency was subdivided to represent only the words contained in the text with each sentiment. Below is the resulting word cloud for the negative tweets.

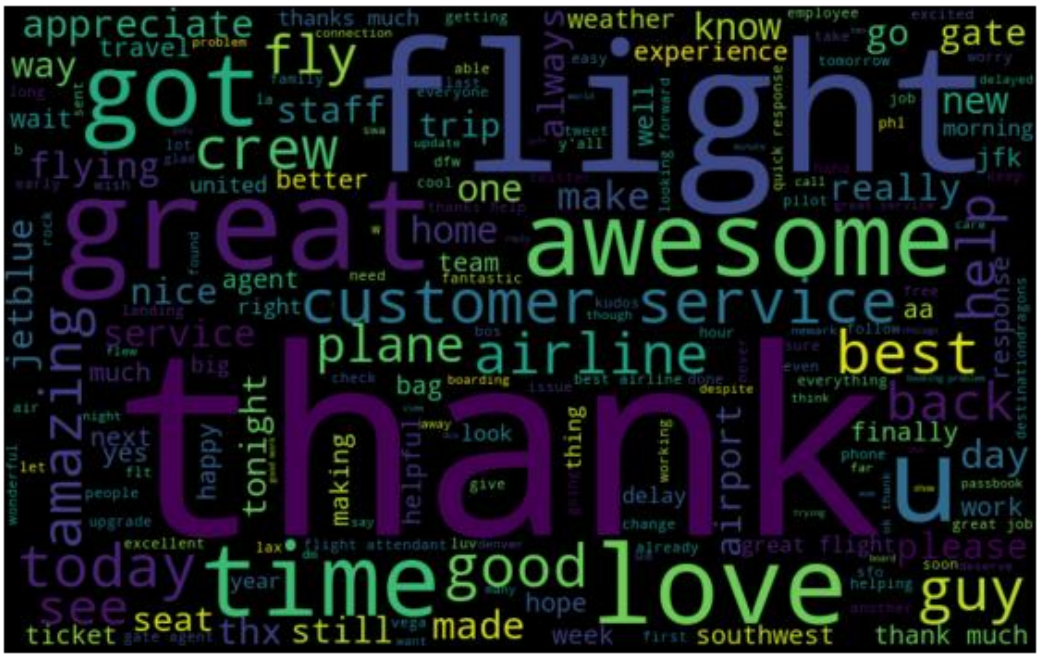


The most common word was flight which was high on the list for all three sentiments, but the next ten more common words provide more insight. They included “cancelled” (cancelled flights), “bag” (lost baggage), “help” (requesting help/customer service) and “hour” (referencing time waiting).

	Word	Frequency
0	flight	3335
1	hour	1104
2	get	1016
3	cancelled	918
4	customer	773
5	service	765
6	u	757
7	...	741
8	time	737
9	bag	662
10	hold	613

Positive Tweets

Similarly with the positive tweets a word cloud was constructed to look at what words were used most frequently.



The result was quite different from the negative tweet word cloud. Words were included like “thanks” (gratitude), “great”, “good” and even the 😊 emoji made it in the top 10.

	Word	Frequency
0	thanks	607
1	thank	453
2	flight	435
3	great	236
4	service	163
5	u	137
6	love	132
7	customer	123
8	guy	122
9	get	120
10	:)	109

Further analysis found an even more surprising trend, that for words showing up in at least 10 tweets there was a surprisingly small number (36) of unique words that appeared in positive tweets that never showed up in negative tweets. (The reverse of this found almost 900 unique words in negative tweets) These included words like “excellent”, “warm”, “welcome”, “enjoy” as well as a whole series of positive emojis. This information was tied into the model dataset to provide the model with more information regarding positive tweets.

Modeling

The airline's goal for developing this model is to have an indication on how well they are performing. The model should be able to accurately distinguish the general sentiment of an individual tweet based on its content. Ultimately having the ability to aggregate the general sentiment of all tweets over time to understand trends.

Model Evaluation

The models considered will be evaluated based on their overall accuracy or f1 score. This considers both precision and recall across all three classes. One additional challenge for model is that the dataset is imbalanced with almost triple the number of negative tweets than either neutral or positive tweets. To ensure that the model is not solely focused on the negative tweets, the f1-score will also be examined individually for the neutral and positive tweets to evaluate how well each model handles the minority classes.

Model Approach

The approach for this project was to use Bag of Words (BOW). Bag of Words is a basic model for Natural Language Processing that focuses on the frequency of words rather than other context dependent factors like grammar and word order. After tokenizing the text, a matrix of tokens is created that identifies, using a 1 or 0, if that token is included in the text (tweet).

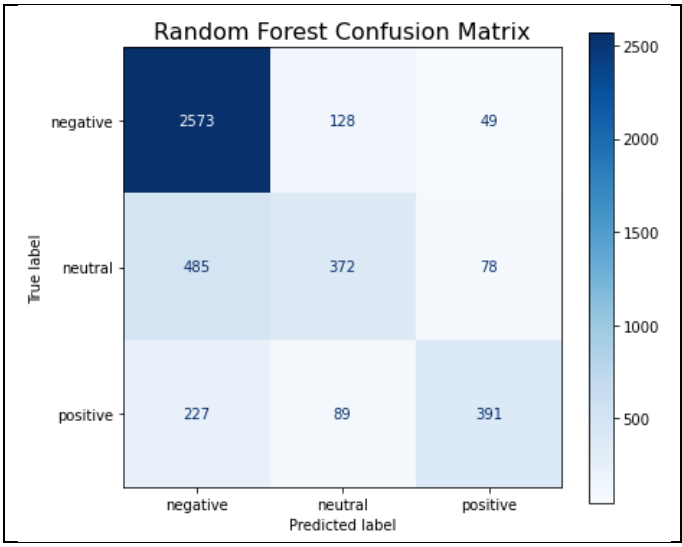
In this case the Bag of Words matrix was limited to tokens that appeared at least 10 times in the dataset. This was done to reduce dimensionality and because model associations are weak when words appear very infrequently. In addition to the BOW matrix, the number of words (scaled using Standard Scaling), number of hashtags (scaled using Standard Scaling), and the associated airline were all included as features for the model. The model dataset was then divided into a 70-30, train-test, split.

Model Comparison & Results

A baseline model was considered for this dataset using the concept of weighted guessing which assumes that your guesses are proportionally aligned with the true frequency of each class in your model. Using this methodology our baseline accuracy would be ~46%.

A Random Forest model with 100 estimators was implemented that performed significantly better than the baseline. The overall f1-score for the model was 76% with better performance on the negative tweets (85%) than either the neutral (49%) or positive (64%) tweets.

Random Forest Model

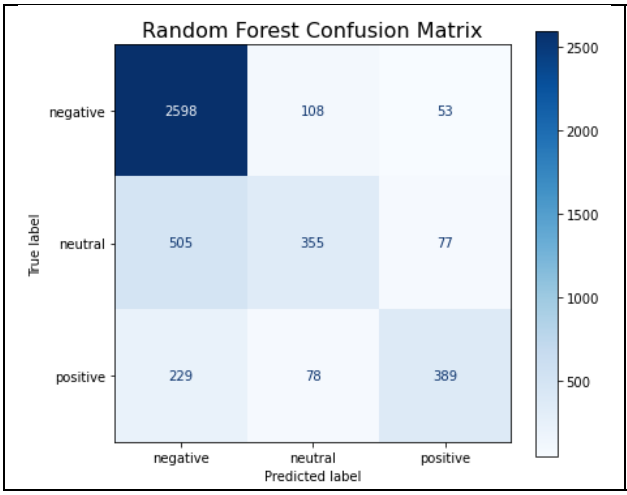


	precision	recall	f1-score	support
negative	0.78	0.94	0.85	2750
neutral	0.63	0.40	0.49	935
positive	0.75	0.55	0.64	707
accuracy			0.76	4392
macro avg	0.72	0.63	0.66	4392
weighted avg	0.75	0.76	0.74	4392

An additional NLP technique, SentiWordNet, was implemented to hypothetically increase the model's predictive power. This technique leverages a "sentiment dictionary" that assigns a positivity and negativity score to particular words. This allows you to calculate an overall sentiment score for a tweet. Each individual token was assessed for a positivity and negativity score and then these results were summed for the tokens in an individual tweet to derive a sentiment score. This sentiment score was then included in the model as a new feature.

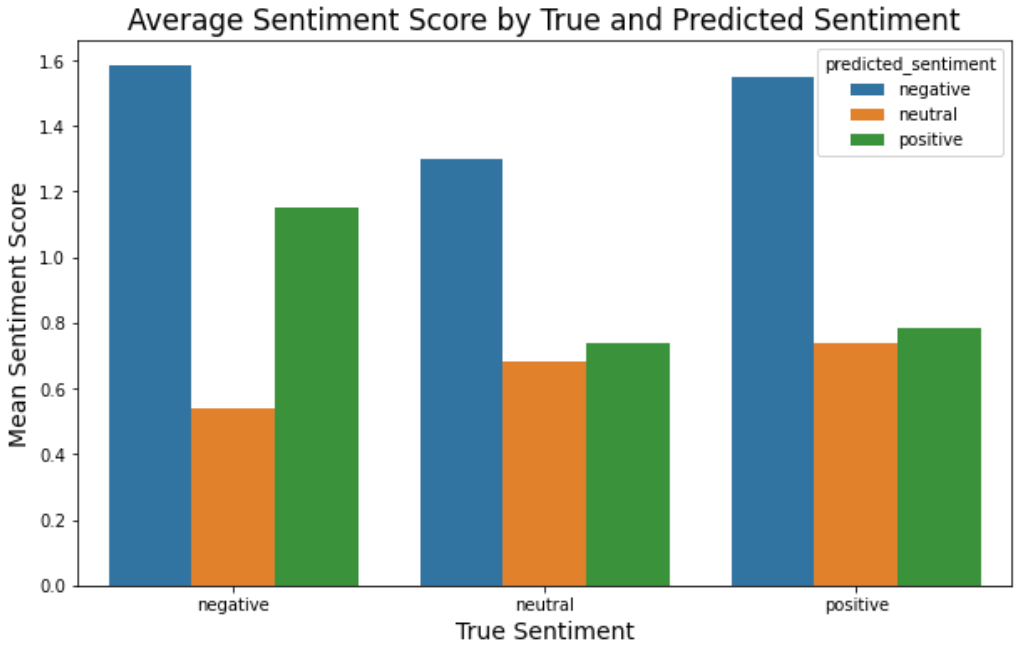
The same Random Forest model was used with this additional new feature included in the model dataset. The results were nearly identical to the original Random Forest model. The overall f1-score was 76% and there were no statistically significant changes in the f1-score for the individual categories.

Random Forest Model with Sentiment Score



	precision	recall	f1-score	support
negative	0.78	0.94	0.85	2759
neutral	0.66	0.38	0.48	937
positive	0.75	0.56	0.64	696
accuracy			0.76	4392
macro avg	0.73	0.63	0.66	4392
weighted avg	0.75	0.76	0.74	4392

Upon reviewing the tweets that the model misclassified the sentiment score did not correlate as expected with the true label. As can be seen below the model predicted "negative" for tweets with the highest sentiment scores (negative tweets should have negative or lower scores). Digging deeper into the individual misclassifications, many of the tweets were ambiguous, relied on sarcasm or started with an expectation of what should have happened that would inadvertently increase the positivity score without understanding the rest of the context. In addition, tweets rely heavily on hashtags and emojis to convey emotion. Since this information would not be included in the sentiment dictionary the model missed out on key signals.



Based on these learnings the addition of the sentiment score did not add much value to this project and seems like it would be more helpful with traditionally structured text rather than tweets. Overall, the model is still able to perform 30% higher than the baseline model. This level of performance will address the problem we set out to solve by allowing airlines to identify problem areas and focus resources intentionally.

Recommendations

To operationalize this model and determine the benefit of real-time sentiment data the airline should implement a dashboard that runs the model nightly on any new relevant tweets. The dashboard could provide the airline's customer service team and leadership with the following:

1. Trended data over time to understand if the customer's sentiment about the airline is getting better or worse.
    a. Additional data should be included around location and time of day of individual tweets so they can be correlated with performance issues like delays, over filled flights, maintenance issues ect. This additional data will allow the airlines to correlate which events have the largest impact on their customer's feelings about them.

2. Trending issues based on key words or hashtags in the tweets.
    a. This model would benefit from frequently being recalibrated for it to learn as new issues arise like COVID or recalls on certain aircrafts.
3. Overall volume of mentions, retweets, etc. to understand how they brand is doing and if negative or positive issues associated with the airline are going viral.

Future Work

Depending on the uptake of this information as being instrumental to running their business the airline could invest into improving the performance of the model by:

1. **Implementing TF-IDF** into the model pipeline. TF-IDF is an NLP approach to weigh down frequent terms like “the” while scaling up rarer ones that appear more frequently.
2. **Using N-grams** to “chunk” the text which allows the model to better understand the context of words rather than the simplification of word frequency used by Bag of Words.
3. **Test VADER** as a substitute for SentiWordNet for which is a rule-based sentiment analysis tool and dictionary that is designed specifically for use with social media text.