US Airline Sentiment Analysis

Rishab Ghose

Table of Contents

Introduction

In the age of social media, Twitter is a powerful tool for the average user to voice their opinion about certain topics, ideas, persons, organizations, experiences, and just about anything else that someone feels the need to state publicly. It has become somewhat of a

digital “towns square” where any and everybody can say what is on their mind. It is not only the individual that has the ability to leverage the power of opinion, however, but companies themselves have the ability to take advantage of the almost limitless data that is being fed into the twitter feed on a daily basis. One immensely powerful method of utilizing the vast stores of opinion data is sentiment analysis. For use by a company, this simply refers to the process of analyzing and extracting opinions and feelings about a certain product, user experience, or company as a whole. In doing so, that company is able to make more guided and strategic decisions to cater to the need of their customers.

In this project, I will be conducting a sentiment analysis of tweets regarding several different US Airlines, and building a model to classify new tweets as positive or negative without any manual labeling. With this model, I will then be able to determine some of the most commonly used words in negative or positive tweets to determine direct reasons why customers are happy or angry with their experience with a particular airline. With this information, airlines would be able to take direct action in correcting some of the negative aspects of their services as well as ensuring they remain diligent in continuing the positive aspects of their services.

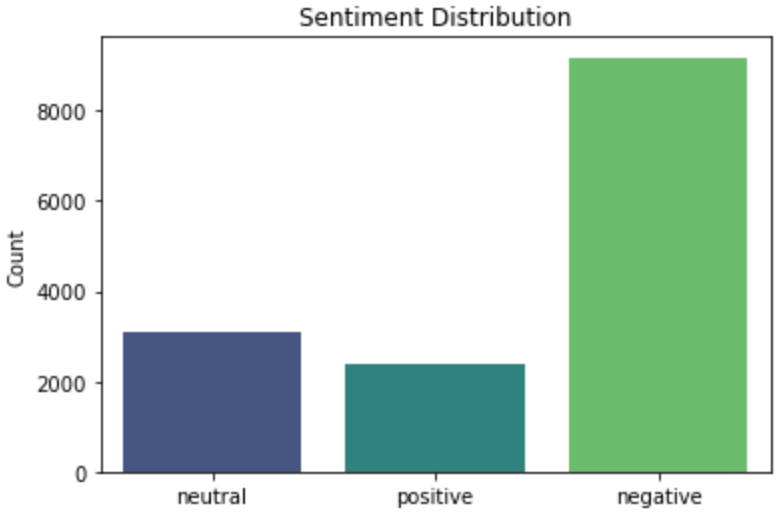
Datasource

I will be using a dataset pulled from kaggle, which is a widely-used source for public datasets. The dataset contains almost 15,000 tweets, each in its own row, in which the tweets were manually labeled as “positive”, “negative”, or “neutral”. These tweets are all tagging a certain US Airline, which was also extracted and put into its own column. For all of the tweets with negative sentiment labels, another feature is included to describe the reason for negative sentiment. The remaining features are just details about the tweet itself, including the account username, time of tweet, retweet count, etc.

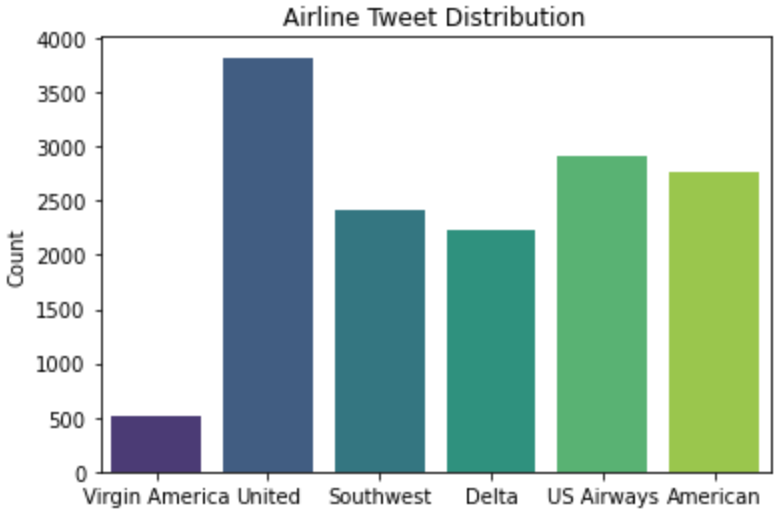
The dataset can be found here:
https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

Exploratory Data Analysis

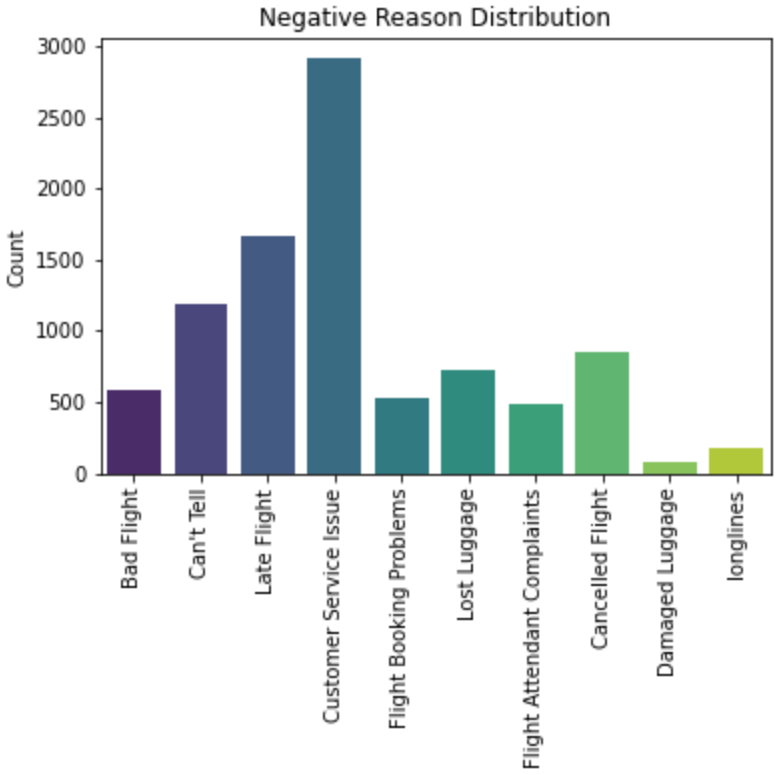
I started by first looking at the distributions for a few of the key features, including tweet sentiment, number of tweets for each airline, and the negative sentiment reasons.



The vast majority of the tweets have negative sentiments, followed by neutral and positive tweets.

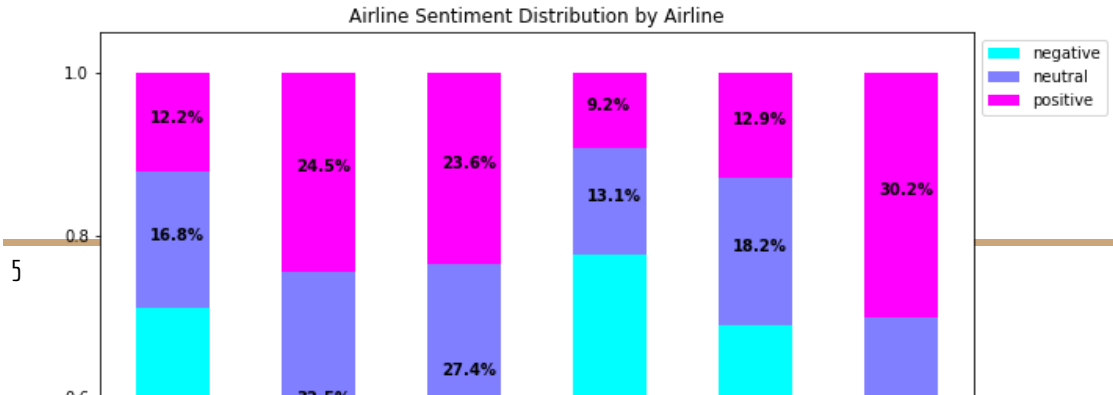


In regards to airlines, United had the most directed at them, while Virgin Atlantic had the least.



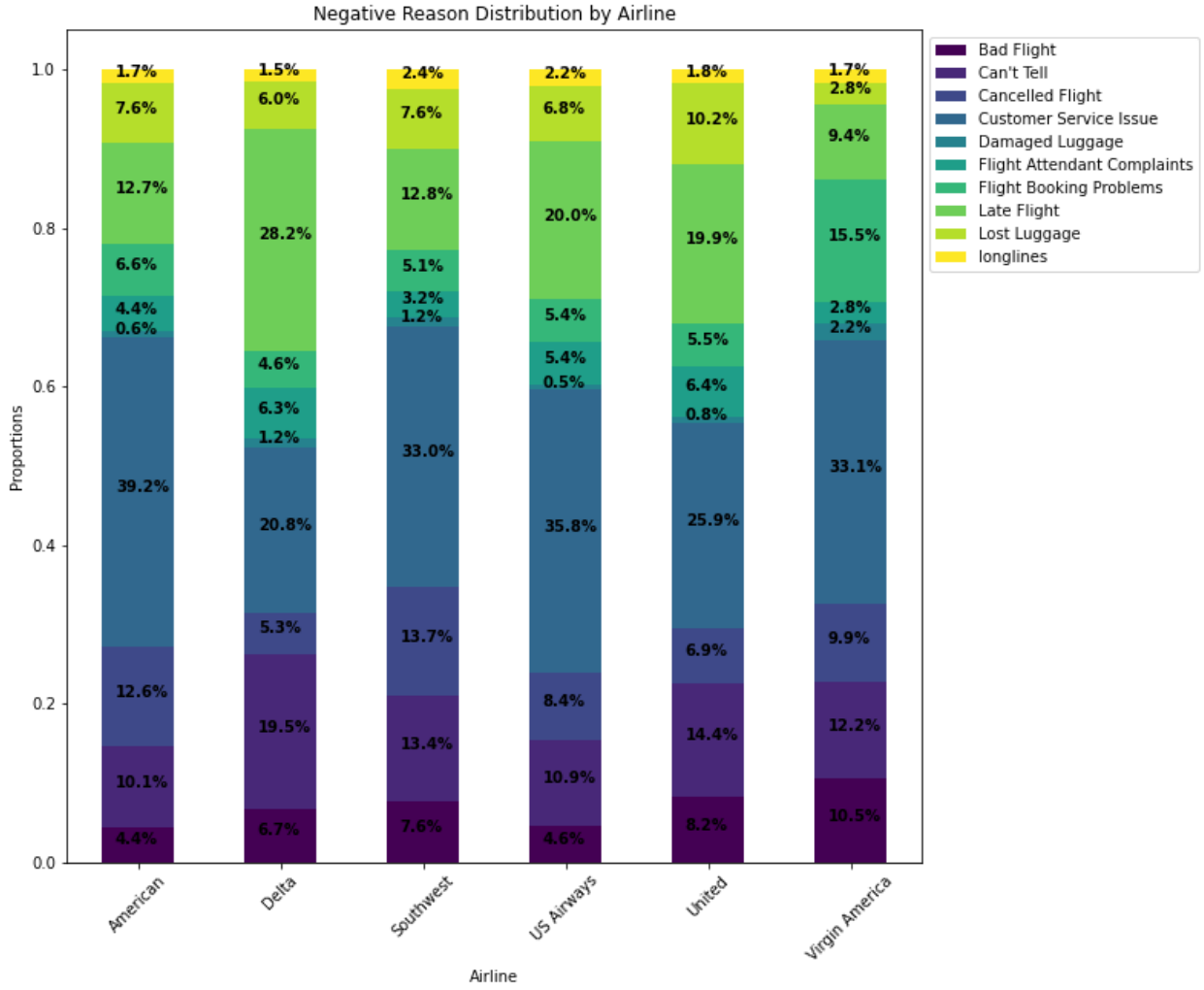
Lastly, in regards to the negative reason, we can see that most of the negative sentiments were due to customer service issues, followed by late flights. Damaged luggage and long lines contributed to the least amount of negative sentiments across all tweets.

After looking at the initial histograms, I wanted to look closer at the distributions of sentiments for each airline. As we can see below, US Airways, United, and American airlines all had over two thirds of the tweets about them having a negative sentiment. Delta and Southwest seemed to have higher proportions of positive sentiment, while the sentiment toward Virgin America was split about evenly. Remember, however, that the number of tweets at Virgin America were far less than the rest of the airlines, meaning it may not be as popular of an airline and hence not as many experiences to tweet about.



I next took a look at how the negative reasons were distributed for each airline. The below shows that Customer Service Issues made up the largest percentage of negative sentiment for all of the airlines except Delta, which had late flights as the highest. Damaged luggage made up the smallest percentage across all airlines, so this does not seem to be that common of an occurrence in the airline industry.

In ee



In the last part of the exploration, I made some word clouds for each sentiment to see what are the common words that people are tweeting when they are tweeting positive, neutral, or negative sentiments (the below word cloud images are in this order). Looking at the positive sentiments, we see words like “thank”, “great”, “love”, “awesome”, “amazing”, etc. These are all words that we would expect to see when someone is tweeting about a good

experience with an airline or flight. For neutral sentiments, we see words such as “time”, “thank”, “ticket”, “help”, “please”, etc. These are words that seem to be asking for some sort of assistance from the airlines. Lastly, negative sentiment tweets includes words like “hour”, “hold”, “cancelled”, “luggage”, “customer service”, etc. Clearly, these are words complaining about a cancelled flight, some customer service issue, or wait times. These word clouds were very helpful in understanding the type of words that made of the tweets of a particular sentiment.

Machine Learning

I will now discuss the process of building a machine learning model to classify the tweets as positive or negative sentiment. To do this, I first converted all of the text into a bag-of-words dataframe. This essentially vectorizes all of our text so that each column represents a word and each row represents a tweet. With these numerical features, I was able to fit classification models and predict our target column of positive vs negative. I tried two vectorizers and three classification models (so six combinations in total) and used ROC-AUC score as the metric to determine which model is best. The Receiver Operator Characteristic (ROC) curve is typically used to evaluate binary classification problems. It is defined as a probability curve that plots the True Positive Rate against the False Positive Rate at various threshold values to see how well the model can separate the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the classifier’s ability to distinguish between classes and is used as a summary statistic of the ROC curve. I want a higher ROC-AUC score, as this would mean the model performed better at distinguishing between the two classes across all probability thresholds.

After determining the best model to proceed with, I then further looked at the performance of this model by finding the optimal threshold. In our business case, we want a model that does very well in predicting negative tweets accurately. This is because by detecting these negative tweets, airlines can find the customers who had negative experiences and respond to them or offer them promotions as forms of damage control. They can also learn the reasons why people have negative sentiment towards them. Therefore, we would want to put more of an emphasis on maximizing precision while still maintaining a decent recall since mislabeling a negative tweet as positive is more detrimental to the use of these

models. Once I determined the optimal threshold, I used that to produce confusion matrices and classification reports.

Pre-Processing / Feature Selection

After the initial exploration, the next step was to clean the tweets to get it into a format that can be easily used in models. I started by checking if there were any leading or ending whitespace in the tweets. There were none, but if there were, then they would have to be removed. I then created a function that performed the following steps: turning all tweets to lowercase, removing words starting with "@" as these are twitter tags that will not be useful, expanding contractions, removing all non-alphabetical words, tokenizing all words, lemmatizing all words, removing stop words, and joining them back together. I applied this to each tweet to get a new column with all clean tweets. This column had about 50 tweets that were now empty after the preprocessing, so I got rid of these. I then got rid of the neutral sentiment tweets as it would be more useful for the business case to predict and classify positive vs negative tweets only. Lastly, I converted positive tweets to have a value of 1 and negative tweets to have a value of 0. This would act as our target column.

Train-test split

I split the source data set by setting aside 70% of data for training and the remaining 30% for testing (while stratifying the data for our target variables to ensure similar proportions of User vs non user in each set). The X sets included just the cleaned text column, while the y sets included only the sentiment value.

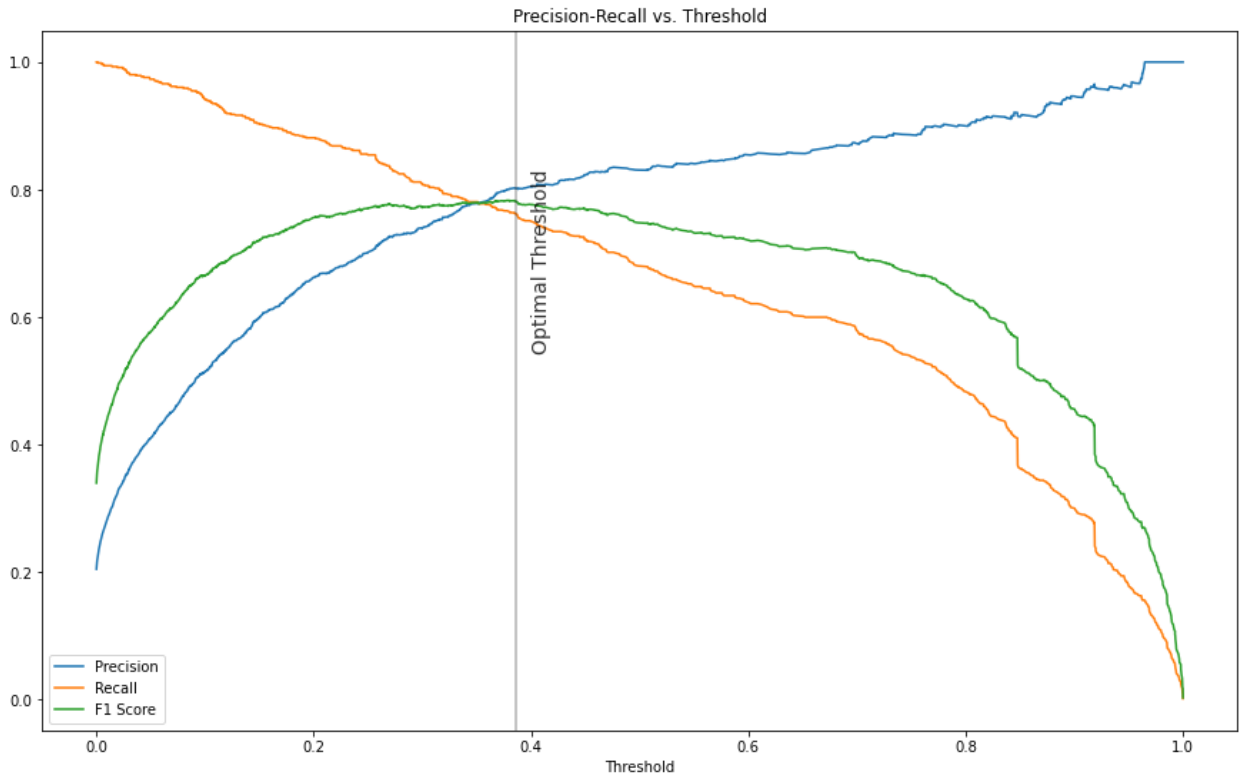
Modeling

To model the data, I first defined a pipeline for each vectorizer and classifier combination. The two vectorizers used were CountVectorizer and the TfidfVectorizer, and the three classifiers used were Logistic Regression, Multinomial Naive Bayes, and Support Vector Machine. Parameter grids were defined for each classifier and GridSearch cross validation was used to find the best parameters, in which accuracy was used as the evaluation metric. The below table displays a summary of the best parameters and ROCAUC score for each model.

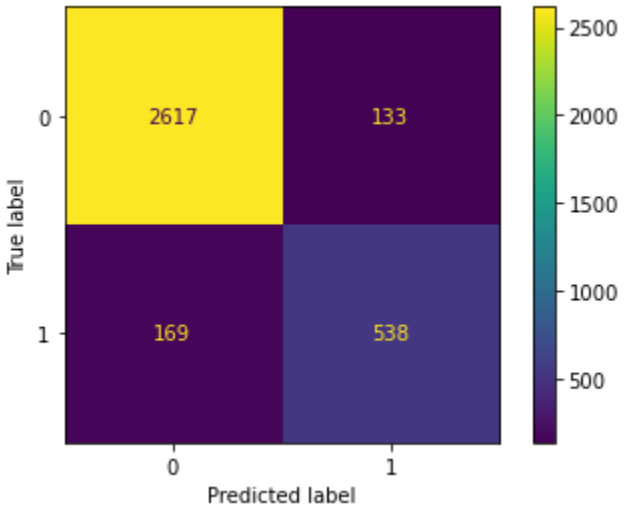
Model Performance			
Vectorizer	**Classifier**	**ROC-AUC Score**	**Best Hyperparameters**
Count Vectorizer	Logistic Regression	0.948	'C': 1, 'max_iter': 1000, 'max_df': 0.75, 'min_df': 3, 'ngram_range': (1, 1)
Count Vectorizer	Multinomial Naive Bayes	0.939	'alpha': 1, 'max_df': 0.5, 'min_df': 3, 'ngram_range': (1, 1)
Count Vectorizer	Support Vector Machine	0.785	'gamma': 'scale', 'kernel': ‘rbf’, 'max_df': 0.5, 'min_df': 3, 'ngram_range': (1, 1)
Tfidf Vectorizer	Logistic Regression	0.947	'C': 1, 'max_iter': 1000, 'max_df': 0.75, 'min_df': 3, 'ngram_range': (1, 1)
Tfidf Vectorizer	Multinomial Naive Bayes	0.937	'alpha': 0.1, 'max_df': 0.5, 'min_df': 3, 'ngram_range': (1, 1)
Tfidf Vectorizer	Support Vector Machine	0.798	'gamma': 'scale', 'kernel': 'rbf', 'max_df': 0.5, 'min_df': 3, 'ngram_range': (1, 1)

Count Vectorizer models performed quite similarly to the TfidfVectorizer models, but Count Vectorizer performed just slightly better. Of the Count Vectorizer, Logistic Regression

performed the best. Next step was to evaluate the model further and determine an optimal threshold. I plotted the precision, recall, and F1 score, and determined the optimal threshold to be 0.386, as this was where the F1 score was maximized and precision reached 0.8.



Using this optimal threshold, I then looked at the confusion matrix and classification reports to get a closer look at performance.



	precision	recall	f1-score	support
0	0.94	0.95	0.95	2750
1	0.80	0.76	0.78	707
accuracy			0.91	3457
macro avg	0.87	0.86	0.86	3457
weighted avg	0.91	0.91	0.91	3457

We can see the model does quite well in classifying negative sentiment, which is great in terms of the use case. When looking at the tweets that were misclassified as positive, they include words such as “thanks”, “best”, “great”, etc. These tweets, however, are clearly sarcastic in nature, showing that the model does not do too well in picking up this sarcasm. Some of the tweets that were misclassified as negative are typically describing a certain issue they were facing being resolved, including words such as “cancelled”, “time”, and “refund”. A clear next step in this analysis was to look at the strengths of each word, particularly seeing which words were most predictive for negative sentiment and which were most predictive for positive sentiments.

Identifying Strength of Predictive Features

To find the strongest words associated with both positive and negative sentiments, I first created a dataset where each row was exactly one text feature (this was essentially an identity matrix). I then used the trained classifier to make predictions on the matrix and sorted them by predicted probabilities. I did this overall (for all airlines), as well as individually for each airline. Below shows the results for the top 20 positive and negative words.

Good words	P(fresh | word)
thank	0.92
thanks	0.85
amazing	0.84
awesome	0.81
great	0.81
kudos	0.79
excellent	0.78
love	0.76
wonderful	0.76
thankful	0.76
Bad words	**P(fresh | word)**
paid	0.07
online	0.07
disappointed	0.06
hold	0.06
rude	0.06
delayed	0.05
hour	0.05
website	0.05
luggage	0.04
worst	0.02

The negative words are particularly helpful, as we see some major issues that the twitter community was complaining about across all airlines involved customer service issues, luggage issues, and website/phone issues.

To find the top 10 positive and negative words for each airline, I went through the same process, only splitting the data by airline prior to training of the models. In doing so, each airline would be able to get a better understanding of what reasons they are getting the most complaints for. For United Airways, the positive sentiments include a lot of words including “thanks”, “impressed”, “appreciate” - they are all celebrating a good experience and showing appreciation for good service. The negative sentiments include words like “hour”, “phone”, “speak”, and “delay” indicating lots of flight delays and issues with customer support (particularly over the phone). Next looking at US Airways, the negative

sentiments again seem to show disappointment to time related concerns. The words "reservation" and "hold" and "website" also showed up which may indicate customers are having issues with booking on the website or over the phone. See an example tweet below:
'@USAirways three hour wait and counting waiting for reservations on the phone. Are you serious!?'

For American, the negative sentiment tweets include words like "rude", "phone", and "bag" which indicates issues with customer service and friendliness of staff. We also see they may be losing or damaging bags frequently. For Southwest, we see words such as “luggage”, “wifi”, “online”, etc. which indicate issues with the in-flight wireless connection as well as luggage issues. They also seem to have cancellations and delays. Below is an example:
"@SouthwestAir if you're going to charge for wifi, do us all a solid and make sure it doesn't take the length of the flight to open a page"

Similar positive words showed up for Delta as other airlines, but negative words include “hotel”, “tv”, “member”, “online”, “point”. These could all indicate issues that customers, members in particular, are having online or when booking hotels due to cancellations/delays perhaps. This could have an impact on member churn rate which should definitely be looked into by Delta.

'@JetBlue thank you for not even coming with a solution. Great service I might say...as a TrueBlue member I am totally dissatisfied...thanks'

Lastly, the positive words for Virgin America are similar once again, with customer service showing up as well. Negative words indicate similar problems with words such as “luggage”, “seat”, “delayed”, “cancelled”, and “website” once again. Overall, the in-depth analysis of negative words for each line are incredibly helpful in determining particular reasons why customers are unhappy with the airlines, finding the unhappy customers, and figuring out how to solve these issues.

Conclusion

In conclusion, sentiment analysis is clearly an extremely effective method in determining how people are feeling towards a particular company, service, product, etc. This model can be used by each airline in specific to directly pull tweets that are tagging their profile, classify the tweets as negative or positive and ultimately determine particular reasons why

customers are unhappy. With this, they can direct more effort into correcting some of these complaints, such as providing more assistance due to flight cancellations, working to create a better online experience when booking through their website, and rewarding employees who receive praise by customers which would incentivise the customer service representatives and flight attendants to be more cordial towards customers. Airlines can also directly identify tweets with negative sentiment, reply to the disappointed customers directly (damage control, maintaining image and reputation, etc.), and offer refunds, promotions, etc. for their negative experiences. There are several further methods of analysis that I would like to do to increase the use and effectiveness of these models, such as including emotional features for each word that would give further classification power to the model. Also, determining ways to detect sarcasm in tweets would be extremely helpful in classifying more negative tweets correctly. One method to do so would be to manually label tweets as sarcastic or not sarcastic and use this as another feature when training. To truly understand how machines can find sarcasm, one would need to understand how humans themselves detect sarcasm (emotion, tone, past behavior, etc). With this, further features can be devised to make the model that much more effective. More further modeling would be to include neutral as a classification category and make this a multi-class problem, as well as create model to classify negative sentiment directly. Overall, I believe this model and ones like it can provide a major benefit to airline organizations in attracting new customers as well as influencing returning customers.