Predicting sentiment from tweets about airlines

Springboard - Capstone 2

Jenny Rhee

Background

- Airlines in the United States handle approximately 10 million flights containing one billion passengers per year.
- It is important for airlines to understand how they are performing with customer service to prevent losing consumers to competing airlines.
- 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.

Data

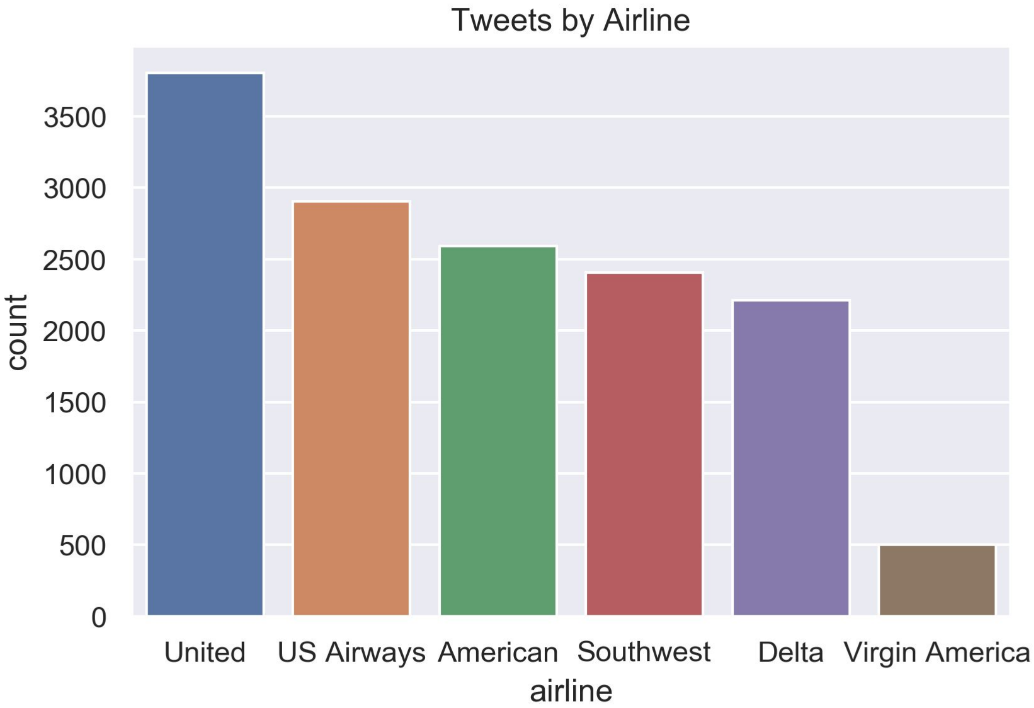
- The data comes from Figure Eight, who provides high quality training data to solve various machine learning problems.
- The relevant columns in this particular data set are airline sentiment, negative reason, airline, Twitter handle, retweet count, tweet content, tweet timestamp, and user timezone.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

Data Cleaning

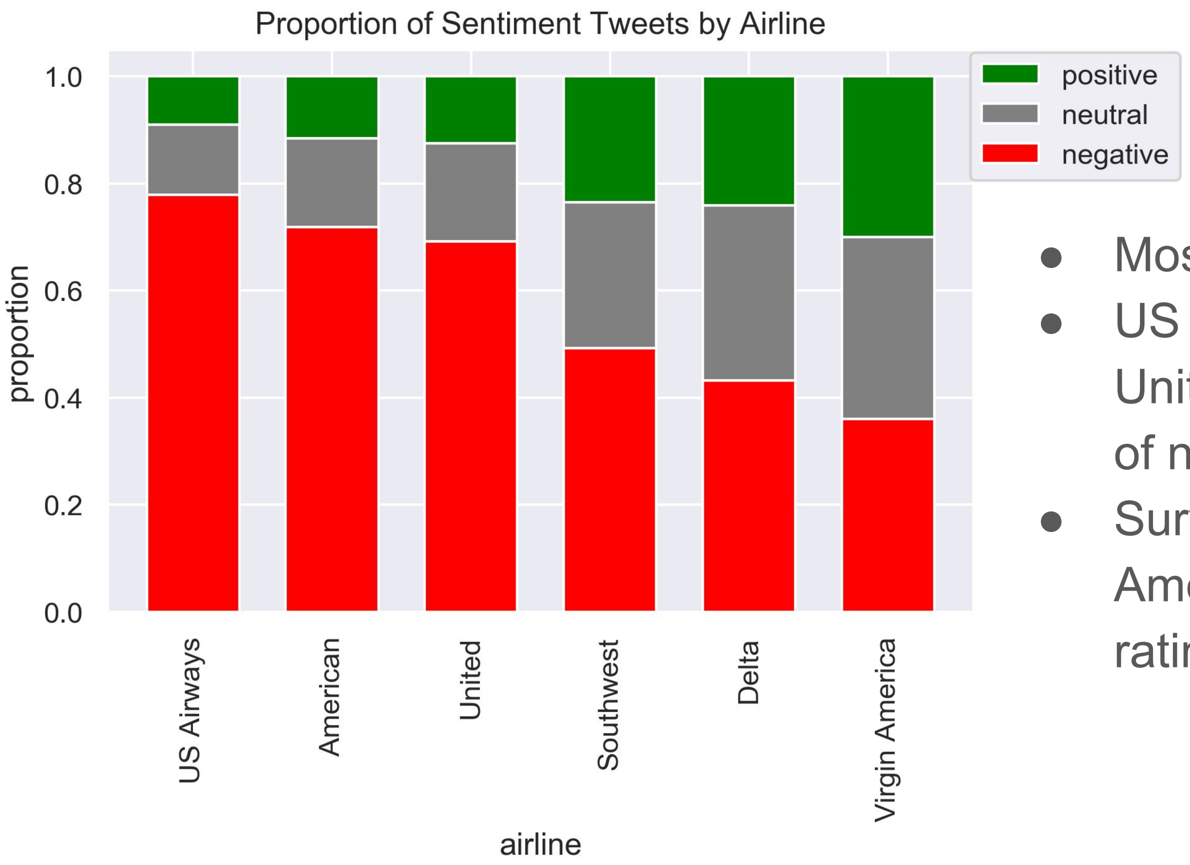
- Duplicate tweets (i.e., retweets) were removed.
- Removed URLs, mentions, punctuation, numbers, whitespace, and stop words
- Sentences were broken up into tokens and stemmed
- Tokens vectorized

Feature Engineering

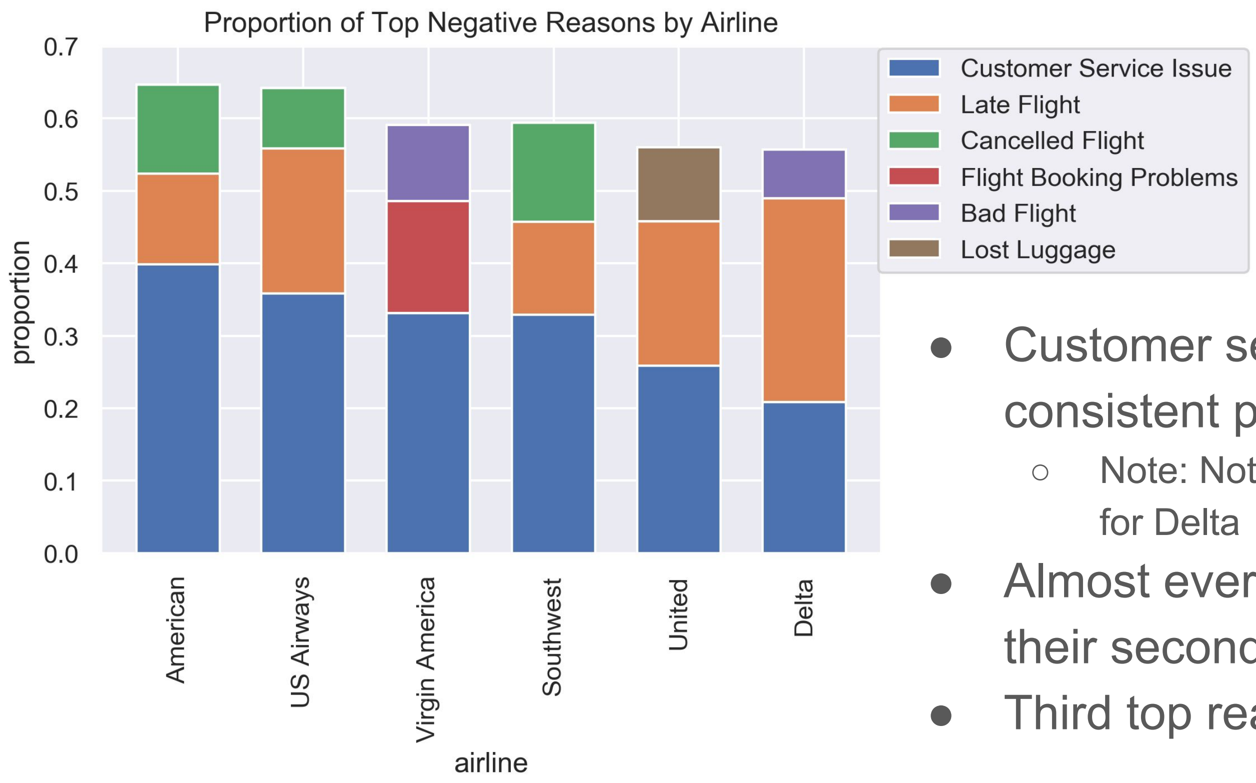
- Features from original, uncleaned tweets:
    - Total character count
    - Number of capital letters
    - Number of words
    - Capital letter to character count ratio
    - Number of happy emoticons
    - Number of sad emoticons
    - Number of exclamation marks
    - Number of question marks



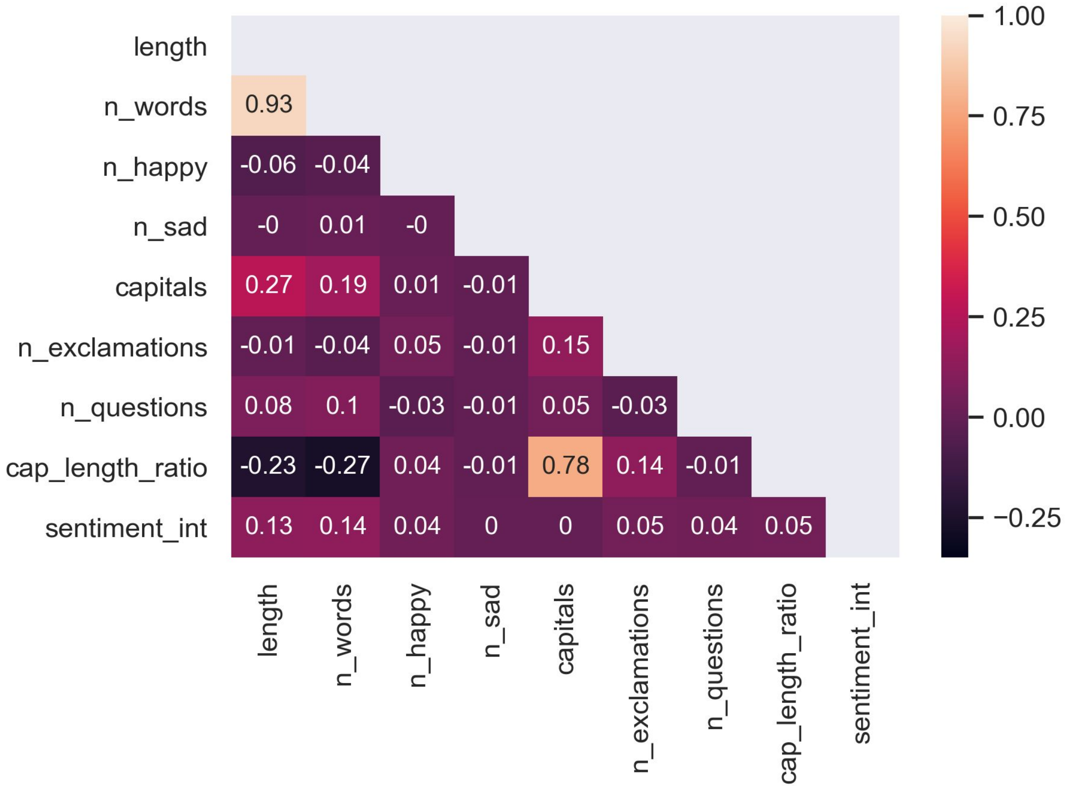
- Considering the top airlines by passengers carried, American Airlines, Delta, Southwest, and United were ranked top 4 in North America.
- Most tweets mentioned United (though they're the 4th busiest)
- Virgin America mentioned the least



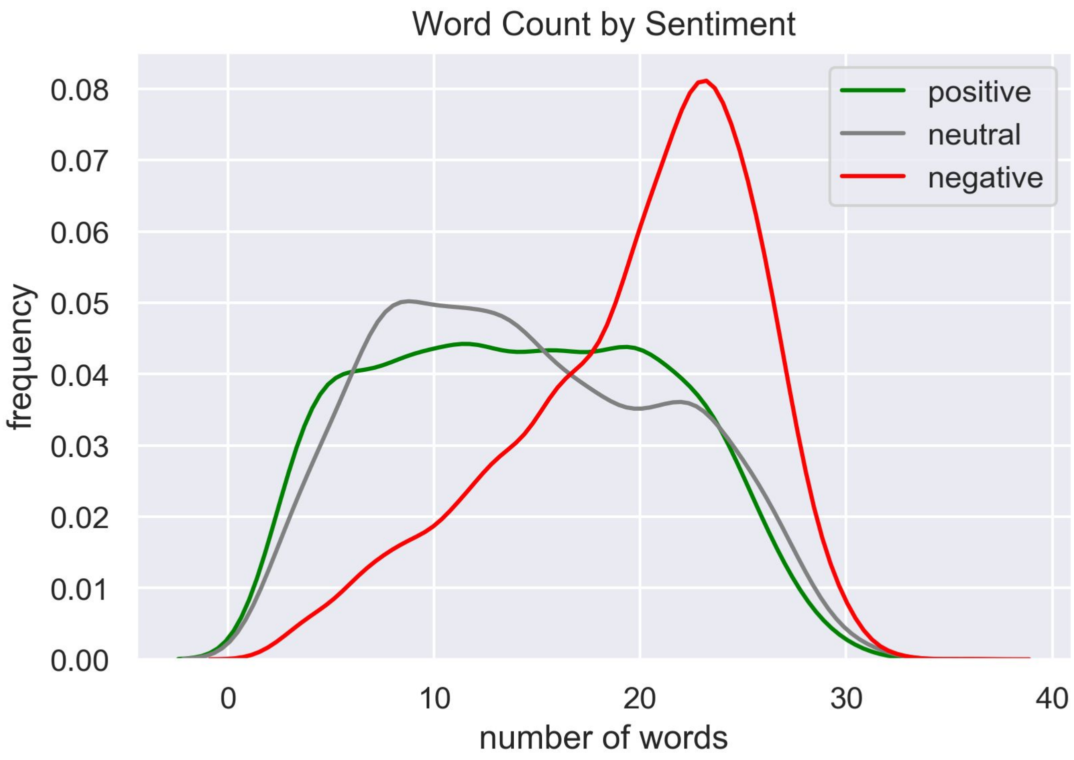
- Most tweets were negative
- US Airways, American, and United had the highest proportion of negative tweets.
- Survey from 2018: United and American were among the lowest ratings.



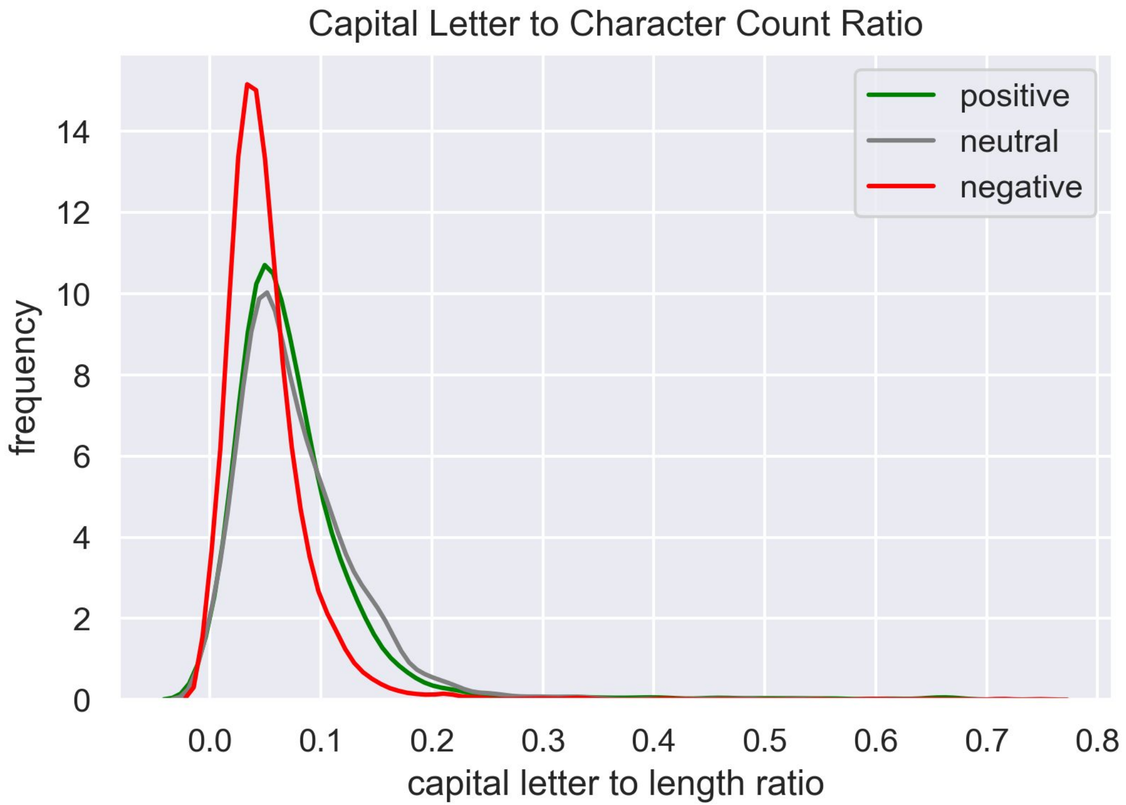
- Customer service issues were a consistent problem among all airlines.
    - Note: Not the most common negative reason for Delta
- Almost every airline had late flights as their second highest complaint.
- Third top reason varied airline-to-airline.



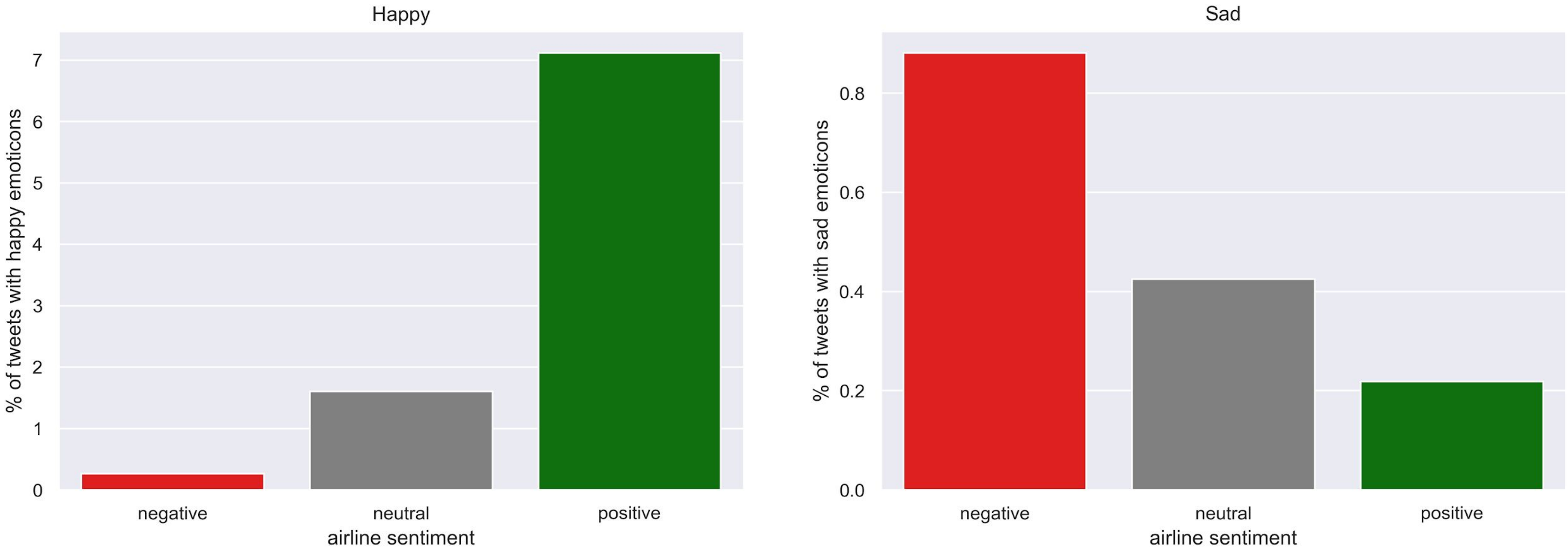
- Most of the predictive power from vectorized vocabulary (not shown)
- Interesting to see a small correlation between sentiment and length/number of words



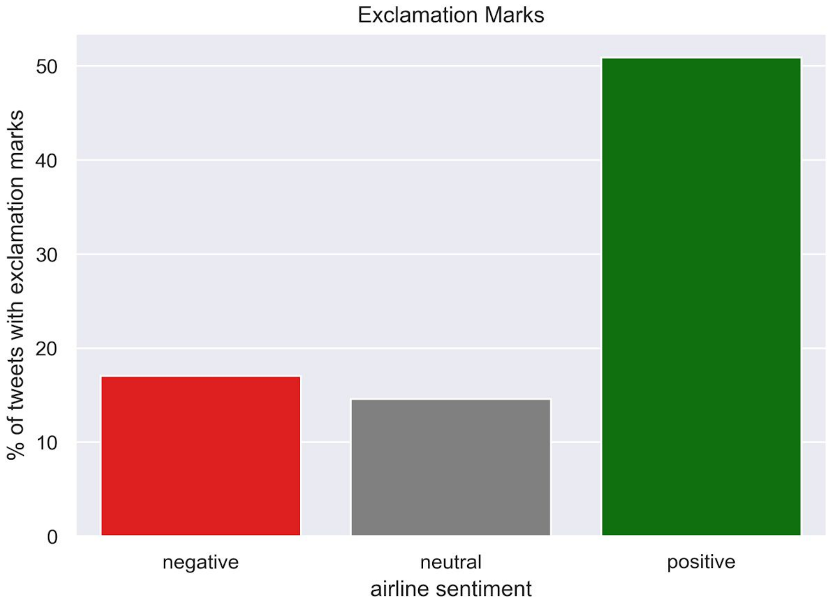
- Positive and neutral tweets had a wide distribution of word counts.
- Interesting to see distinct difference between negative and positive/neutral
- Not surprising: People have more to say when they're upset.

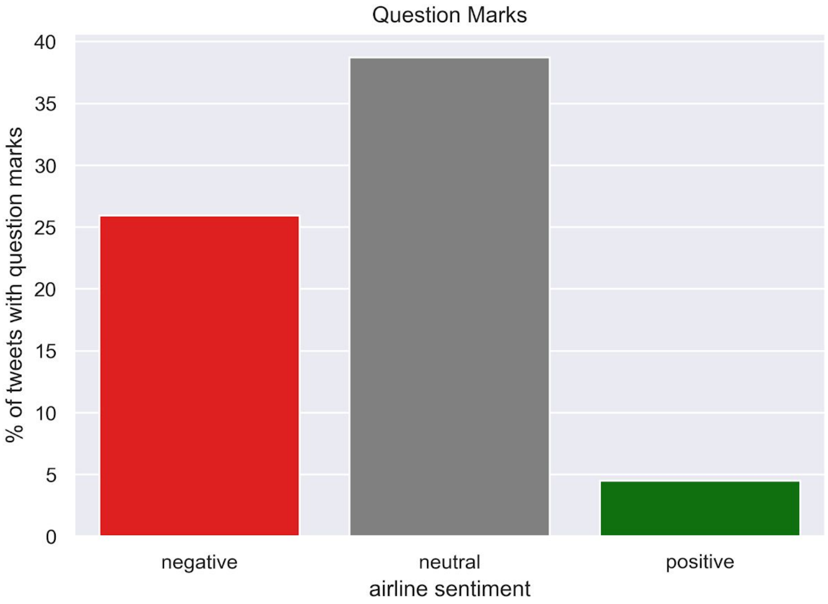


- Positive and neutral tweets had higher average capital letter to character count ratios.
- Positive tweets: All caps when excited
- Flight industry has a lot of abbreviations (e.g., airlines, flight numbers, etc).
- Neutral tweets: Information seeking



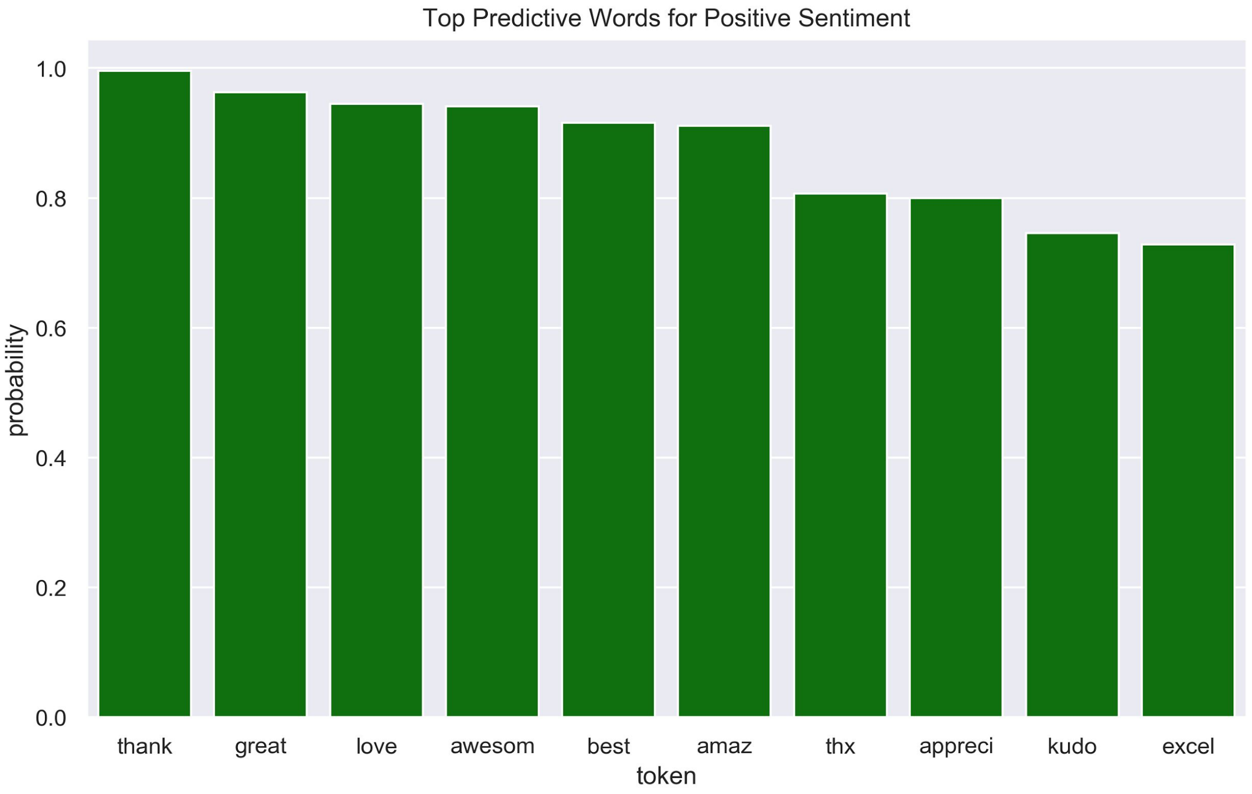
- Highest percentage of tweets with happy emoticons: positive
- Highest percentage of tweets with sad emoticons: negative





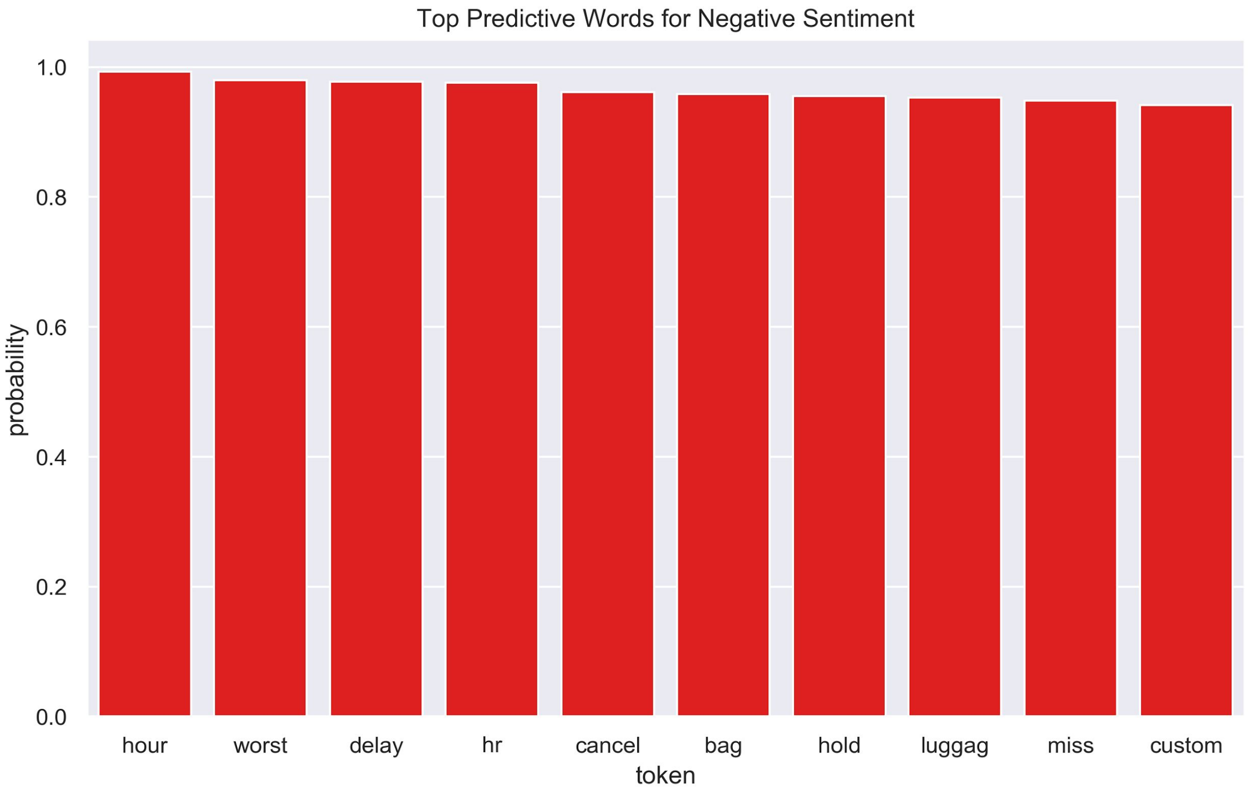
- Over 50% of positive tweets had exclamation marks; less than 20% of negative and neutral tweets had them
- Exclamation marks can be associated with happiness or anger, but it looks like in this case, they were primarily used in happy tweets.
- Neutral had the highest percentage of tweets with question marks (information seeking)

Positive Word Cloud



- Thank, great, love, awesom, and amaz were the top 5 most predictive tokens for positive sentiment.

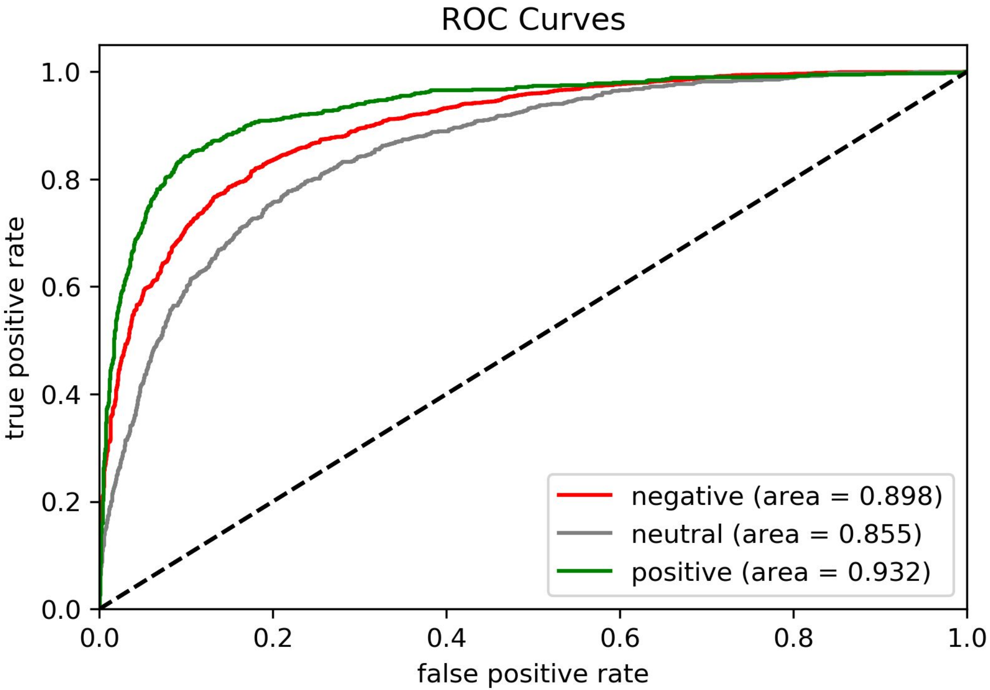
Negative Word Cloud



- Hour, delay, worst, hold, and cancel were the top 5 most predictive tokens for negative sentiment.

Sentiment Model Overview

- Modeling task: predict positive, neutral, or negative sentiment from tweets about airlines
- Comparison of two algorithms - Naive Bayes and random forest
  - Naive Bayes mean accuracy = 0.679
  - Random forest mean accuracy = 0.763
- Comparison of two vectorizers - bag-of-words and tf-idf
  - Bag-of-words mean accuracy = 0.757
  - Tf-idf mean accuracy = 0.763
- Hyperparameter of both the vectorizer and classifier using randomized search



- Hyperparameter tuning:
  - Tf-idf:
    - ngram_range = (1, 2)
    - max_df = 0.6
  - Random forest:
    - n_estimators = 1500
    - min_samples_split = 0.001
    - max_features = log2
    - class_weight = balanced_subsample
- Mean accuracy score = 0.783
- Overall ROC-AUC score = 0.889

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

	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

Recommendation

- Random forest performed the best with mean accuracy score = 0.763 vs. Naive Bayes' mean accuracy score = 0.679
- Final model with hyperparameter tuning mean accuracy score = 0.783 and ROC-AUC = 0.889
- Implement a random forest classification model and use 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.

Conclusion & Future Direction

- Tweets directed towards airlines are generally negative.
- Unsurprisingly, the frequency of negative tweets per airline resembles the results from previous surveys.
- Most issues are related to customer service or late flights, which airlines may not have much control over.
- However, examining customer service tweets will be useful for airlines to pinpoint what is going wrong.
- Next step: Create a web application that airlines can use to monitor sentiment about themselves from recent tweets or about their competition.