Airline Tweet Sentiment Analysis

Business Problem

In an industry as competitive as air-travel, reputation and loyalty are critical for keeping and drawing new customers (Heathcote, 2024). Social media offers businesses a new avenue into authentic customer opinions. To inform business strategy ahead, the goal of this project will be to create a model that can process tweets from airline customers. The analysis will seek to answer the following.

Can a model process and predict customer sentiment from tweets as humans would?
What aspect of the flying experience was most common in negative sentiment?
What airline do tweets favor the most?
What airline do tweets favor the least?

Background/History

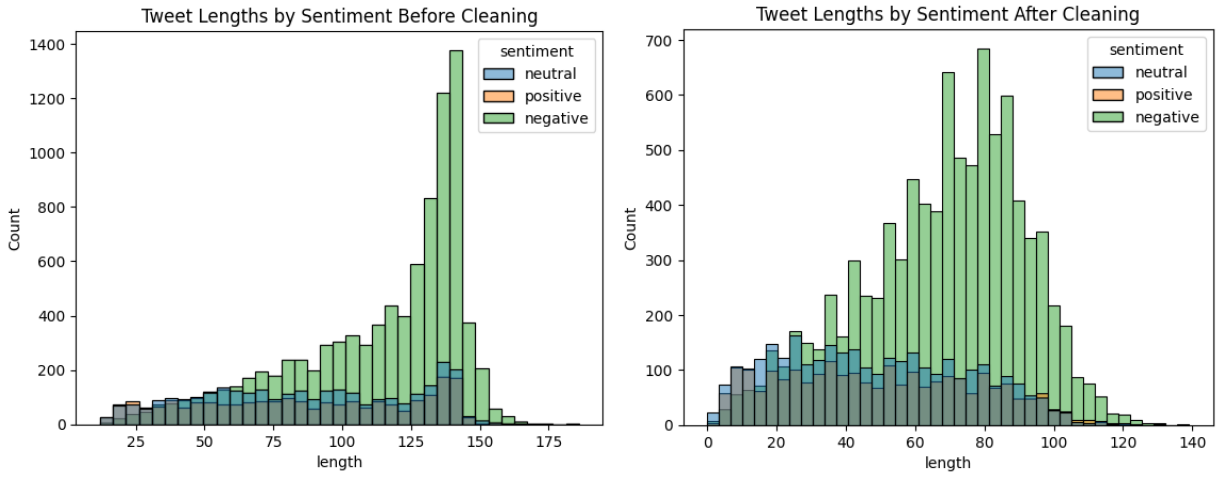
In the past business would have to rely on focus-groups, polls, and surveys to gain insights into public opinion or customers. With social media platforms like Twitter, attitudes are accessed readily for anyone who wants to see. Sentiment analysis on social media posts provides an unfiltered view of customer perceptions and experiences. Information gained can offer businesses information they need to evolve, compete, attract, and retain customers. Here an analysis of airline related tweets will be processed for sentiment analysis. A model will be built to classify opinions as positive or negative. This model can then be applied in the future to gain insights from social media posts as the airlines evolve.

Data Explanation (Data Prep/Data Dictionary/etc)

The dataset used for this project was sourced from kaggle.com containing 15 features, including tweets, location, creation time, time zone and sentiment labels and more. The data, which was originally sourced from CrowdFlower, was labeled for sentiment by a human panel along with a consensus score. These labels were used to train and test my model which can then be used in the future for analyses as the brand evolves or incidents occur. The set appears to have over 14k tweets related to 5 US Airlines.

To prepare this data for sentiment analysis, tweet texts were manipulated to remove user handles, URLs, special characters, and stop words using the NLTK library. Class counts were visualized along with counts for the airlines in which tweets were directed. Visualization was also used to examine tweet lengths prior to modeling. Negative tweets were readily seen as having a much higher average than neutral or positive labeled tweets, peaking at 1400 characters.

Figure 1



After examining lengths, tweets were filtered to contain over 3 characters, leaving 14,607 rows of data. Once these were removed tweets were tokenized and a porter stemmer was applied. My final step in preparing the data before splitting was to drop the "neutral class" of tweets, leaving 11,530 rows.

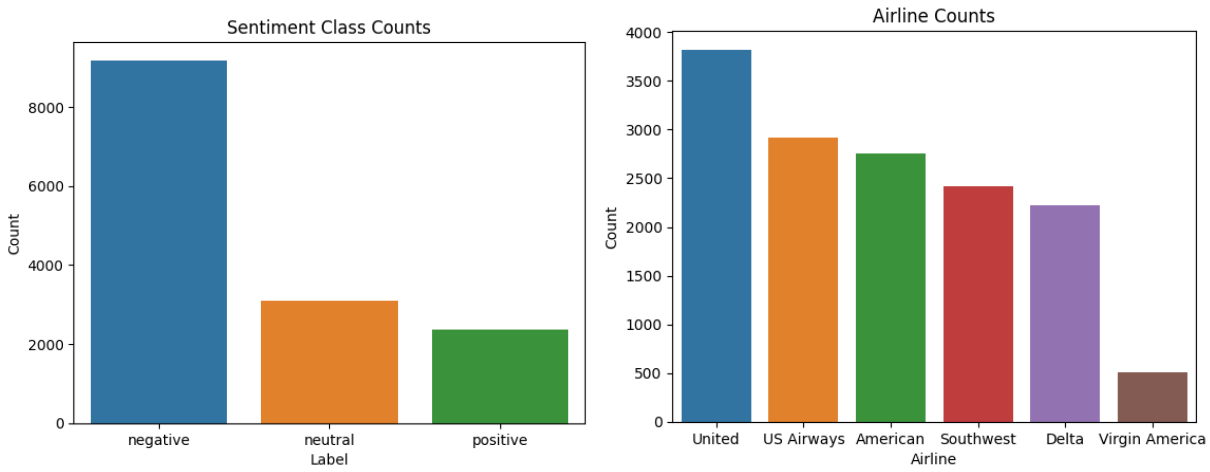
Methods

My dataset was split at a ratio of 80:20 for training and testing. I used the test dataset to calculate null accuracy score at 79.5%. I ran separate models using TDIF vectorization and count vectorization to transform the test data. I also examined accuracy across a logistic regression, a gradient boosting classifier, and a random forests classifier. Model accuracy and accuracy on test set was used to evaluate which model to use. A confusion matrix (see Appendix Figure 1) was used to visualize results followed by a classification report which included scores across target class for precision, recall, and F1.

Analysis

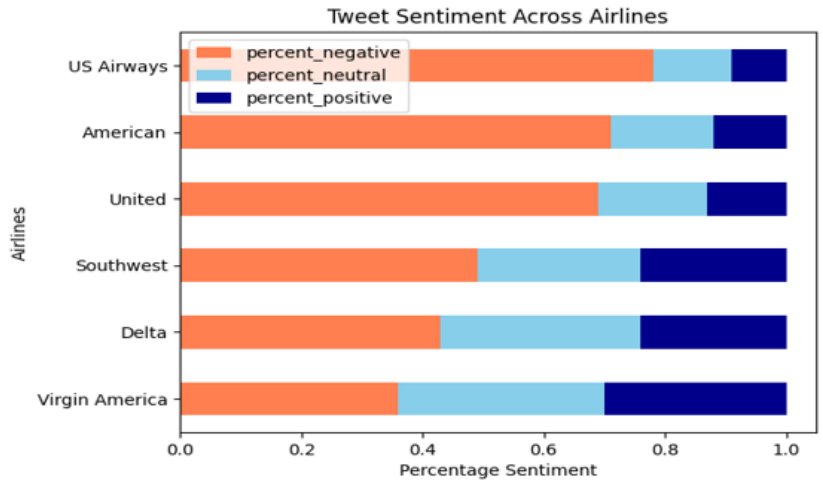
My initial analysis found that the tweets in the data set were imbalanced. With the majority labeled "negative" at 9,178, "neutral" at 3,099, and "positive" at 2,363. Visualization also showed that the 5 US airlines included in the tweets were well represented with 4/5 having over 2k related tweets. The lowest tweet count was Virgin America airlines, at around 500 tweets.

Figure 2



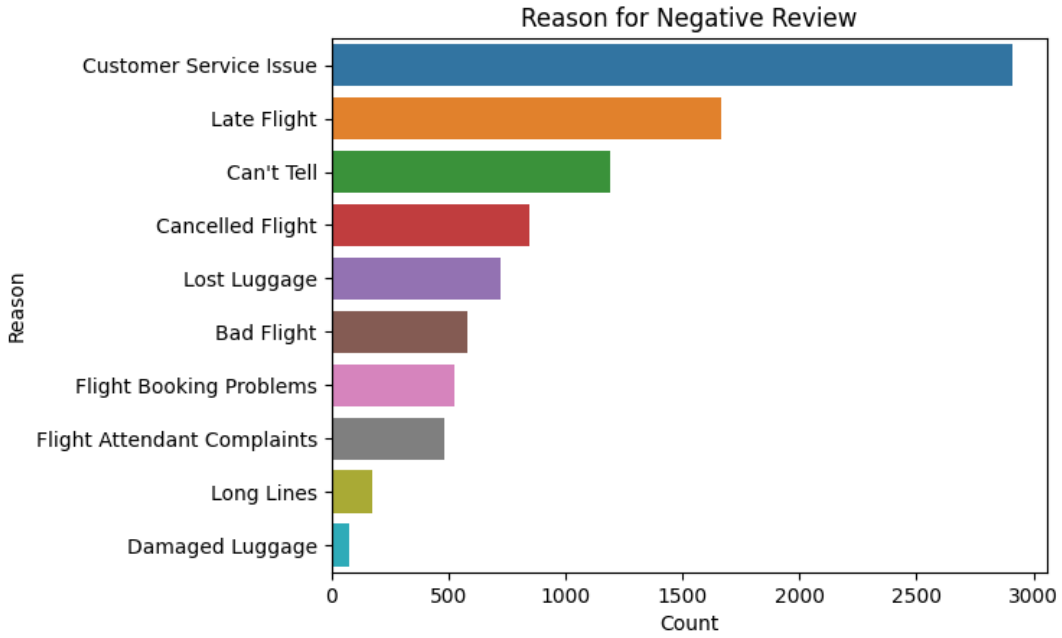
To compare airlines, a percentage of positive and negative reviews was calculated from each airline review total. Bar charts were used to visualize the results which showed that Virgin America had the highest percentage of positive reviews at 30%. The next highest percentage was Southwest Airlines at 24%, which also had a much higher sample count. Negative tweets had higher percentages in general, and US Airways had the highest 78%, while the lowest was Virgin America at 36%. A complete table of all review type counts and percentages across airlines can be found in appendix Table 1.

Figure 3



Visualizing the reasons for negative reviews was relatively easy for this dataset as these were provided by the dataset. The visualization below shows that the most common complaints related to airlines overall were related to customer service followed by late flights.

Figure 4



When comparing vectorization methods, count vectorization performed better than TDIF vectorization. After running all three models I chose, the random forest classifier performed that best with model accuracy of 99%, and prediction accuracy 91% on test data. A classification report was run on the forest classifier which showed in better detail how the model performed across target classes.

Figure 5

	precision	recall	f1-score	support
0	0.92	0.96	0.94	1834
1	0.83	0.69	0.75	472
accuracy			0.91	2306
macro avg	0.88	0.83	0.85	2306
weighted avg	0.90	0.91	0.90	2306

We can see that the model performs well overall, with high precision, recall, and F1-score for class 0 (negative class). However, for class 1 (positive class), while the precision is relatively high, the recall and F1-score are lower, suggesting that the model may have difficulty correctly identifying instances of positive tweets. This could indicate the class imbalance is weakening the training process.

Conclusion

Of all the US airlines included in this analysis, Virgin American appeared to have the strongest performance with customers. They had the highest percentage of positive reviews and the lowest percentage of negative reviews. If we exclude this airline due to its low sample size in the dataset, the next top performing airline would be Delta. Delta had the next highest percentage of positive reviews and smallest ratio to negative to positive reviews. It should be noted, however, that Delta also had the next smallest sample size after Virgin America. This may suggest an overall trend, where the larger the customer base, the larger proportion of negative reviews.

For sentiment analysis, modeling to classify tweets went very well. My random forest classifier performed the best at 91% accuracy, which outperformed the null accuracy at 79.5%. The model

was particularly good at identifying negative tweets which is valuable information for a company. With a more balanced sample, it is likely this accuracy would improve.

Assumptions

One major assumption of this model is that the labels are accurate. The data source claims the data was labeled by a team of people where the majority classification won, and the individualized votes were expressed as a confidence score for each label. No evidence seen from the data suggests these sentiment labels were not appropriate.

Limitations & Challenges

One of the primary hurdles of sentiment analysis is the inherent ambiguity and subjectivity of human language. Textual data often contains sarcasm, irony, slang, and context-dependent expressions that can confound sentiment analysis algorithms. As TF-IDF relies on word frequencies to capture document features, it can struggle to interpret the nuances of language accurately. Additionally, sentiment analysis requires understanding not only individual words but also their context within sentences and documents. TF-IDF does not capture semantic relationships between words or consider the syntactic structure of sentences which limits what can be captured. This project is also limited by class imbalance for positive tweets in the training data. While the results may suggest it is simply harder to predict positive sentiment, it is likely that with more data the results would improve.

Future Uses/Additional Applications

This model would easily be used for analysis of other data sources, such as Facebook posts, Google reviews and other hosts for public opinion data. It could also work with a company's internal data sources of customer opinions from open-ended surveys, focus groups, and customer service feedback. Meanwhile, insights from exploratory analysis can readily inform marketing strategies, operational changes, and branding. Companies looking to compete can identify areas of niche or characteristics that widely appeal to customers based on their own strategy and needs.

Recommendations

Getting a larger sample of positive tweets would likely improve the model's accuracy. It may also be improved with a pre-trained model that specializes in tweet text analysis such as BERT or RoBERTa. Another valuable step for this project would be to include modeling to identify and extract reasons for negative and positive reviews from tweets. Most importantly, a care remodeling to include neutral tweets will offer more usability, as new tweets will only be forced into positive or negative categories.

Implementation Plan

With a well-trained model, a company can continue to pull tweets to gather insights daily, or as events happen. With automated reporting to periodically gather tweets, a pipeline can be created to clean and preprocess them for modeling. This would render a consistent reporting method that would be relatively easy to adjust over time. Using dashboarding, a company could

gather insights in a time-sensitive fashion while making them easily accessible to various departments and stakeholders.

Ethical Assessment

Like any real-world data, the potential for bias when training a model is a concern. Different sub-groups within a customer base may have different concerns and priorities. While a large sample is likely to be more representative, language and word choices vary greatly across American sub-cultures. Bias in NLP modeling has been found for gender and race (Kiritchenko & Mohammad, 2018). Given the potential use of this data, to inform areas of improvement or branding, I would recommend the overall ethical risks of this analysis are low in terms of potential harm.

References

Barreto, S., Moura, R., Carvalho, J., Paes, A., & Plastino, A. (2022). Sentiment analysis in tweets: an assessment study from classical to modern word representation models. *Data Mining and Knowledge Discovery*, *37*(1), 318–380. https://doi.org/10.1007/s10618-022-00853-0

Gu, C., & Kurov, A. (2020). Informational role of social media: Evidence from Twitter sentiment. *Journal of Banking & Finance*, *121*, 105969. https://doi.org/10.1016/j.jbankfin.2020.105969

Heathcote, A. (2024, January 30). Digital marketing in aviation: The importance of marketing for airlines. *Adido Digital*. https://www.adido-digital.co.uk/blog/digital-marketing-in-the-airline-aviation-industry/

Kiritchenko, S., & Mohammad, S. M. (2018). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1805.04508

Stembridge, G. (2023, November 20). *Consumer trends to watch out for in 2024*. Qualtrics. https://www.qualtrics.com/blog/global-consumer-trends/

Topic: Passenger airlines in the U.S. (2024, February 27). Statista. https://www.statista.com/topics/5575/passenger-airlines-in-the-us/#topicOverview

What is customer sentiment and how do you measure it? (2023, June 16). Qualtrics. https://www.qualtrics.com/experience-management/customer/customer-sentiment/

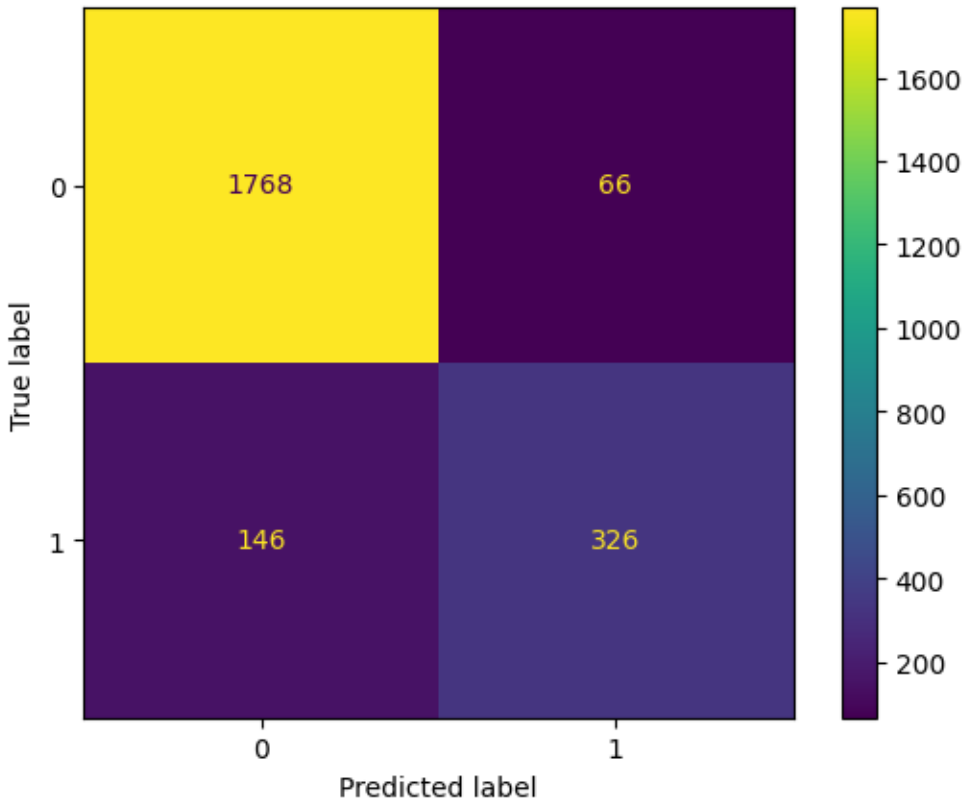
What is sentiment analysis and how can users leverage it? (2023, September 28). Qualtrics. https://www.qualtrics.com/experience-management/research/sentiment-analysis/

Appendix

Table 1

	airline	sentiment_negative	sentiment_neutral	sentiment_positive	percent_positive	percent_negative	total_reviews
5	Virgin America	181	171	152	0.30	0.36	504
1	Delta	955	723	544	0.24	0.43	2222
2	Southwest	1186	664	570	0.24	0.49	2420
4	United	2633	697	492	0.13	0.69	3822
0	American	1960	463	336	0.12	0.71	2759
3	US Airways	2263	381	269	0.09	0.78	2913

Figure 6



Questions & Answers

1. Why not include neutral tweets in analysis?

   Leaving out neutrally labeled tweets allowed me to focus this model to focus only on more extreme sentiments in the data. For continued modeling on unlabeled data, inclusion would be preferred, for the purposes of this project, it was a way to simplify access to more valuable information.

2. What does a precision score tell us?

   This score captures the proportion of correctly predicted positive values out of all the predicted positive values. In this case, it represents the correctly predicted positive sentiment out of all the predicted positive sentiment.

3. What does a recall score mean?

   As opposed to the precision score, a recall score gives the proportion of correctly predicted positive values out of all the actual positive values.

4. What is an F1 score?

   The F-1 Score ranges from 0-1 and represents a balanced combination score of precision and recall combined. Zero score represents the worst performance of both measures, and 1 representing the best performance of both.

5. What is the difference between TF-IDF vectorization and count vectorization?

   TF-IDF vectorization means term frequency – inverse document frequency. Like count vectorization, it converts words to numerical values. These values represent importance in TFIDF vectorization, where in count vectorization, these values represent frequency of a term throughout the collection terms in the data. In TFIDF vectorization, It not only counts the instances of a term within a text, but uses these counts to distinguish what terms are so highly shared between groups that they offer less value. By that same mechanism, less shared terms, or unique terms between groups are given more weight.

6. How does the model used in this analysis differ from a BERT model?

   This project model takes in input in the form of list of words, and uses the words, and their term frequency to predict labels. A BERT model accepts input of various sections of

a sentence to allow the model a wider context of words and how they relate to each other.

7. Why are there so many more negative tweets than positive for all airlines?

   There is a widely known phenomena in world of gathering customer opinions where people, in general, appear more inclined to go out of their way to express negative experiences that rather than positive ones. However, we can see from this analysis, that some airlines tweets do not reflect this disproportion.

8. What is tokenization?

   Tokenization is a process of splitting/separating words in a sentence. This turns a typical sentence, into a list of comma-separated words, along with their punctuation and any other characters.

9. What is Porter Stemmer?

   Stemming is shortening of words to their root. This allows words that only differ in prefix or suffix to be grouped, or reduced to the same root word. For example take, taken, taking would all stem down to 'take'.

10. Why drop only rows with less than 4 characters? Why not more?

    When reviewing the results of filtered comments at 3, 4, and 5 character lengths. Useful, distinct, words appeared at 4 characters that did not appear at 3. So this became my threshold for filtering length minimum.