# Sentiment analysis on U.S Airline Tweets

8007 Project Document
Hee Kyoung Nam
500462612

## Introduction

The digital connectivity allows users to express their thoughts, opinions, and reviews on a brand. Thus, the user posts shared on social media are forming the base of customer strategy for companies worldwide. Sentiment analysis becomes highly resourceful for companies who want to gain user insights and strategize their brands.

In this report, tweets for six US Airlines will be examined. In the airline industries, feedbacks from customers can be found on the social media platform such as Twitter. We can run a sentiment analysis on the available feedbacks to find out various factors that affect customer experience. The main objective of this project is to provide the airline industry a more comprehensive views on how traveler feels about airline services and outline customer likes, dislikes, and expectations.

## Dataset

The 'Twitter US Airline Sentiment' dataset can be found on Kaggle. Table 1 shows the information of the data frame. The data contains 14640 tweets on six US airlines (Virgin, United, American, Delta, Southwest, Airways) that are posted on twitter in February 2015. Tweets are classified in three different classes: positive, negative, and neutral sentiments and the negative tweets are further classified into several negative reasons. Note that only negative sentiment row has the negative reason column filled. Text column describes content of published tweet written by users.

| | airline_sentiment | airline | text | negativereason |
|---|---|---|---|---|
| 0 | neutral | Virgin America | @VirginAmerica What @dhepburn said. | NaN |
| 1 | positive | Virgin America | @VirginAmerica plus you've added commercials t... | NaN |
| 2 | neutral | Virgin America | @VirginAmerica I didn't today... Must mean I n... | NaN |
| 3 | negative | Virgin America | @VirginAmerica it's really aggressive to blast... | Bad Flight |
| 4 | negative | Virgin America | @VirginAmerica and it's a really big bad thing... | Can't Tell |

Figure 1: Screenshot of first 5 lines of dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   airline_sentiment  14640 non-null  object
 1   airline            14640 non-null  object
 2   text               14640 non-null  object
 3   negativereason     9178 non-null   object
dtypes: object(4)
memory usage: 457.6+ KB
```

| | airline_ sentiment | airline | text | negativereason |
|---|---|---|---|---|
| **count** | 14640 | 14640 | 14640 | 9178 |
| **unique** | 3 | 6 | 14427 | 10 |
| **top** | negative | United | @united thanks | Customer Service Issue |
| **freq** | 9178 | 3822 | 6 | 2910 |

Table 1: dataset information and description

**Exploratory Data Analysis**

Before getting into a more in-depth analysis, I have created several data visualizations to find patterns and characteristics of my dataset. The dataset contains tweets of six US airlines. I created a pie chart to observe the proportion of the tweets that each airline holds.

Figure 2 shows that 26% of total tweets are related to 'United' airline. 19.9% of tweets mention 'US Airways'. 18.8% for 'American' airlines followed by Southwest, Delta, and virgin America. According to the World Air Transport statistic (Figure 3), the greatest total kilometers flown airline in 2015 was American airline. This statistic includes US Airways as there was a merger with American Airline in 2015. Delta was the 2$^{nd}$ airline that flew the most, United was the 3$^{rd}$ and Southwest was 4$^{th}$ in the rank. The following statistic is somewhat similar rank to the number of tweets created in 2015, but United had the slightly higher number of tweets unlike other airlines.

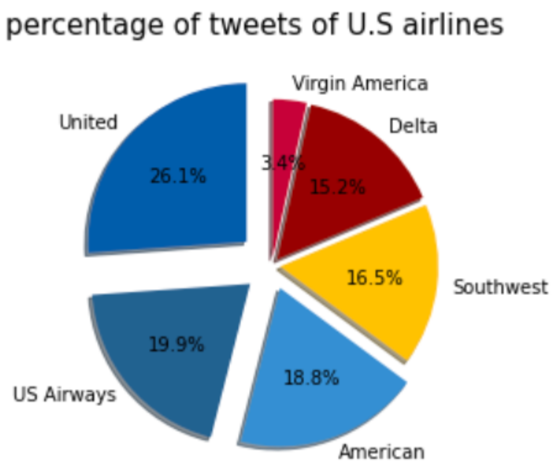



Figure 2: Proportion of tweets per airline

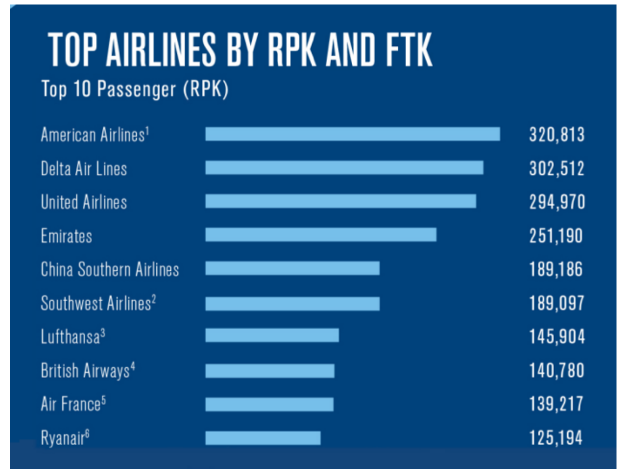



Figure 3: World Air Transport statistic in 2015

Moreover, the distribution of sentiments of the tweets is observed using a pie chart (Figure 4). It is surprising that over 62.7% of tweets are negative tweets. It shows that many tweet users are unhappy about airline services, and travelers tend to leave complaints on the twitter than compliments.

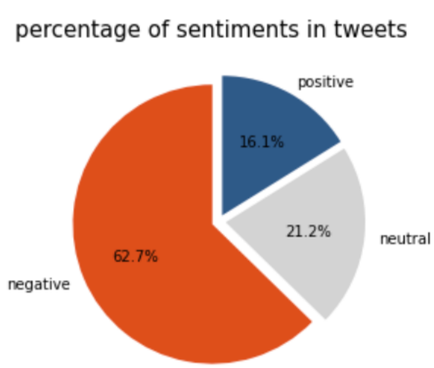



Figure 4: Distribution of sentiments in tweets

The distribution of sentiments for each airline is shown in a bar chart (Figure 5). United, US Airways and American have much higher number of negative tweets than positive and neutral tweets. Southwest, Delta, and Virgin America have relatively balanced distribution. We can summarize that many United, US Airways and American customers had unpleasant experience.

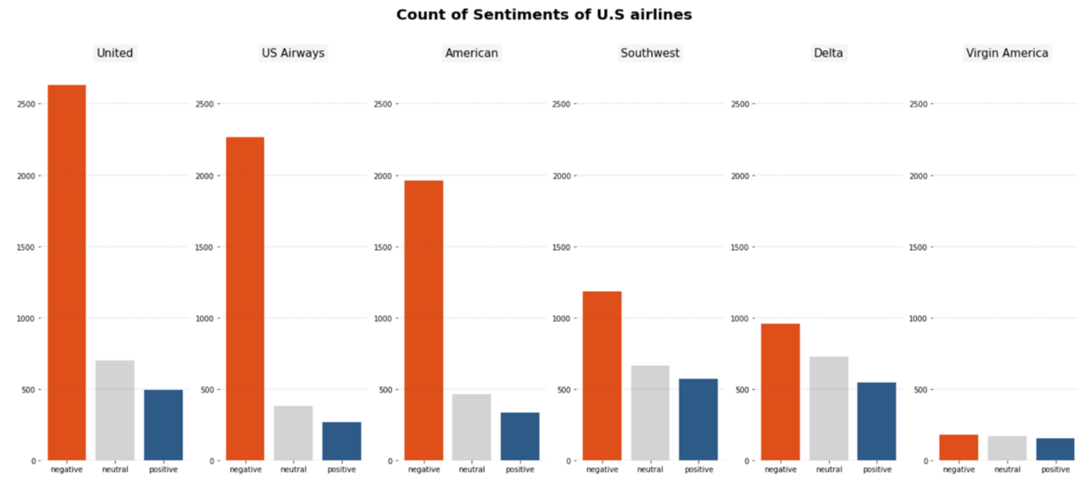
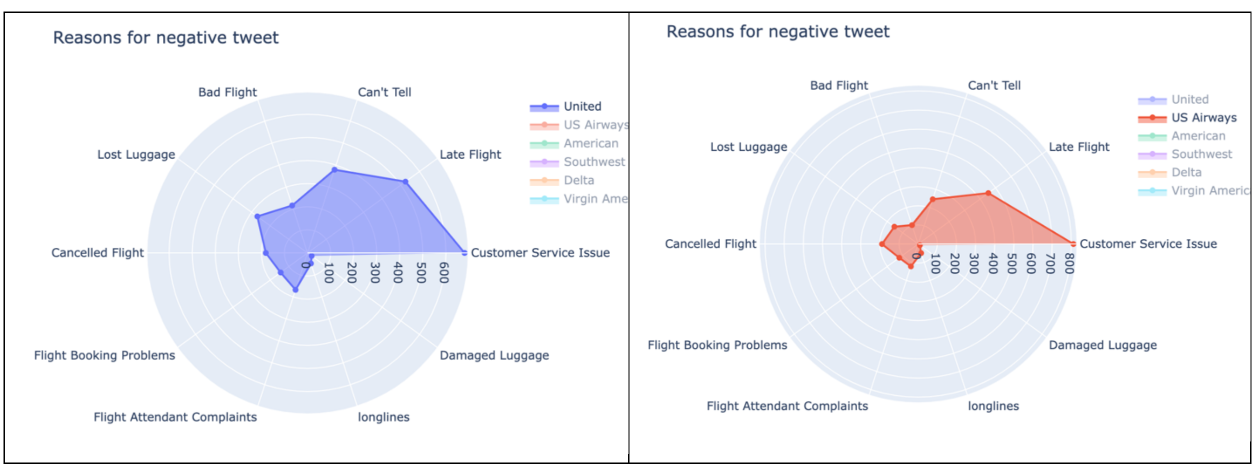


Figure 5: A Bar chart showing the distribution of sentiments for each US airline

In this dataset, reasons for negative tweets are provided. There are 10 negative reasons and I have used radar plots using Plotly (Figure 6) to find out the main reason for leaving negative tweets for each airline. Customer service issues are the main reasons for the negative sentiment to most airlines. US Airways, American, Southwest complaints are mainly about customer service. United customer complaints are about customer service, flight delay and luggage loss issues. Beside customer service problem, Delta customers also experience late flight issues. For Virgin America, flight booking problems and bad flight are the other major problems.



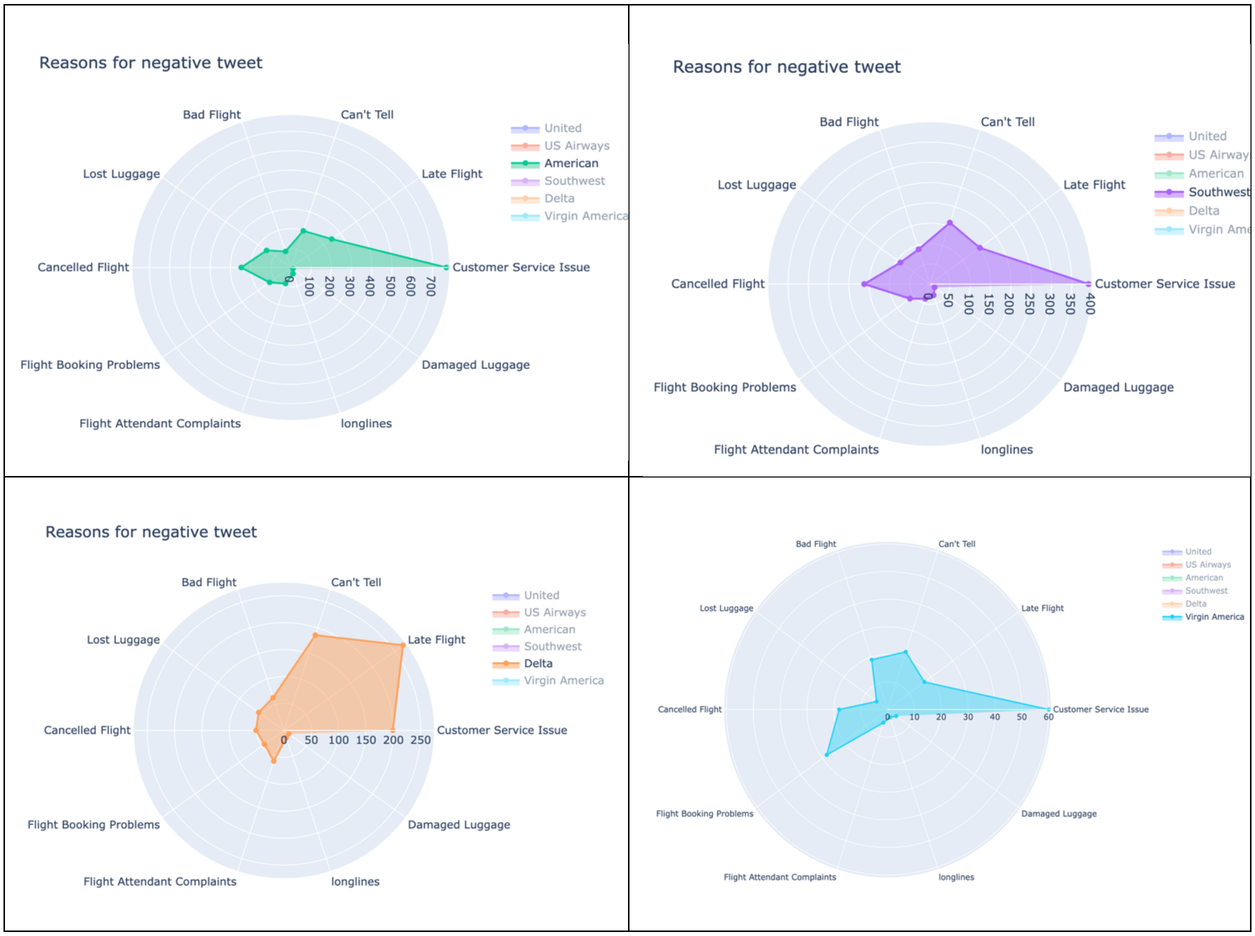



Figure 6: Radar plots for six US airlines which show reasons for negative tweets

Furthermore, I created a Word cloud (Figure 8) using Tableau to visualize most frequent words that appear in positive, negative, and neutral tweets. I have pulled 10 to 20 most frequent words from each sentiment category. The data are saved onto a excel sheet to be opened in Tableau Software (Figure 7). Duplicate words are removed. Since the number of negative tweets are much higher than the number of positive and neutral tweets, the occurrences of negative word are much greater. To solve this imbalance issue, I have divided each word occurrence by the total number of words in the respective sentiment. The count column in the excel sheet represents the proportion of the word occurrence in each sentiment.

For example, Flight with 22.73 value implies that the word 'Flight' appears in 22.73% of total negative tweets. Sentiments are labeled as {negative: -1, neutral: 0, positive: 1}.

| | A | B | C |
|---|---|---|---|
| 1 | word | count | sentiment |
| 2 | flight | 22.73181 | -1 |
| 3 | cancelled | 7.113774 | -1 |
| 4 | service | 5.761696 | -1 |
| 5 | hours | 5.06261 | -1 |
| 20 | please | 9.660297 | 0 |
| 21 | need | 8.704883 | 0 |
| 22 | tomorrow | 5.785563 | 0 |
| 23 | fleek | 5.679406 | 0 |
| 24 | know | 5.573248 | 0 |
| 26 | thanks | 18.57708 | 1 |
| 27 | thank | 13.83399 | 1 |
| 28 | great | 7.175433 | 1 |
| 29 | love | 4.864701 | 1 |
| 30 | much | 3.314077 | 1 |

Figure 7: Screenshot of excel sheet containing word, count, and sentiment information

The size of word in the word cloud is scaled based on the 'count' column value. The color is red for the negatives, gray for the neutral and blue for the positive words. Frequent negative words are flight, delayed, customer, service, time, phone, late, call, cancelled, bag and more. I see that negative words are related to customer issue, flight delay and cancellation. Frequent positive words are thank, great, awesome, airline, crew, best and more. Frequent neutral words are please, tomorrow, need, fleet, know.

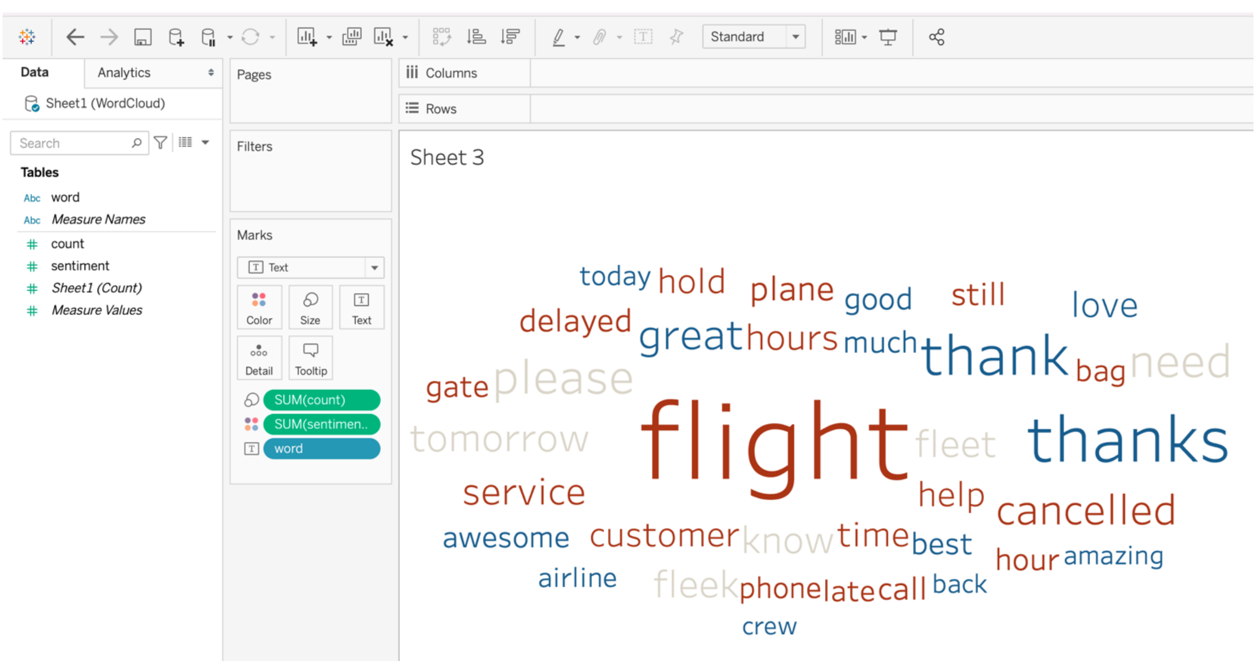
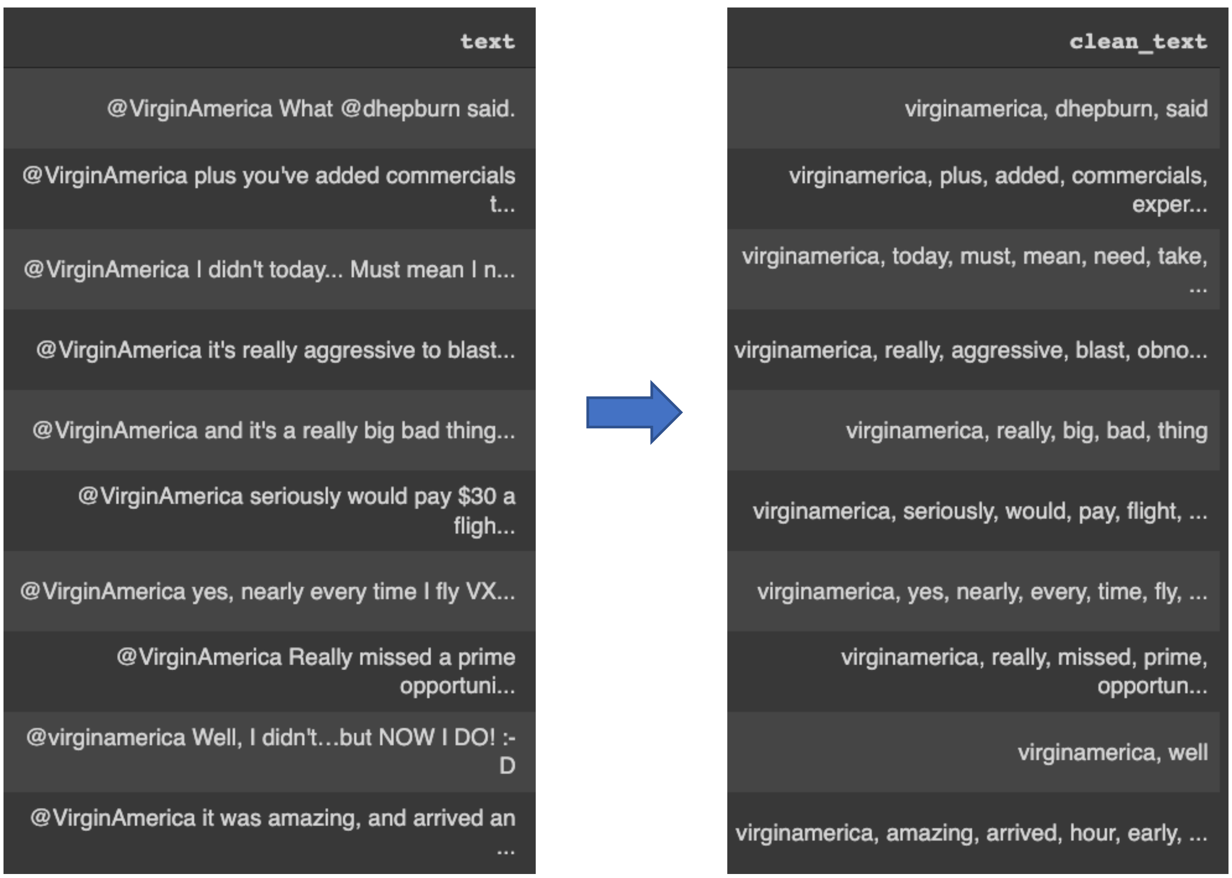


Figure 8: Screenshot of positive, negative, and neutral words containing Word Cloud which is created using Tableau software

**Text Pre-processing**

Tweets are long strings. We need to break the raw text data into words, and they are called tokens. Raw strings and pre-processed tokens are shown in figure 9. Tokenization is a necessary step before fitting the context into the machine learning models. In the process of converting raw string into words, I have filtered out unnecessary characters like punctuation, contractions and stopwords which are common words that does not hold much information to the text. Examples are “the”, “is”, and “and”.



| text | clean_text |
| --- | --- |
| @VirginAmerica What @dhepburn said. | virginamerica, dhepburn, said |
| @VirginAmerica plus you've added commercials t... | virginamerica, plus, added, commercials, exper... |
| @VirginAmerica I didn't today... Must mean I n... | virginamerica, today, must, mean, need, take, ... |
| @VirginAmerica it's really aggressive to blast... | virginamerica, really, aggressive, blast, obno... |
| @VirginAmerica and it's a really big bad thing... | virginamerica, really, big, bad, thing |
| @VirginAmerica seriously would pay $30 a fligh... | virginamerica, seriously, would, pay, flight, ... |
| @VirginAmerica yes, nearly every time I fly VX... | virginamerica, yes, nearly, every, time, fly, ... |
| @VirginAmerica Really missed a prime opportuni... | virginamerica, really, missed, prime, opportun... |
| @virginamerica Well, I didn't…but NOW I DO! :-D | virginamerica, well |
| @VirginAmerica it was amazing, and arrived an ... | virginamerica, amazing, arrived, hour, early, ... |

Figure 9: Conversion of raw string tweet text to tokens

The final step of text pre-processing is encoding text data. Encoding is a process that convert tokens into a vector representation so it can be consumed by machine learning algorithm. Encoding also helps to preserve the context and relationship between words. I used two popular methods to accomplish text vectorization: CountVectorizer and Tf-Idf vectorizer. The encoded vector is returned with a length of the entire vocabulary in each tweet and each index contains an integer value which is the number of times each word appeared in the tweet. On the other hand, TfidfVectorizer returns a float vector because it not only takes the frequency of words into account but also calculate inverse document frequency which gives more weights to the words that appear frequently across tweets.

**Methods and results**

- **Sentiment analysis**

To identify what are the customer likes and dislikes about airline service, a model that classify positive and negative words is built. The positive and negative words classified using the following model are used to examine key factors that affect customer experience and customer needs. Neutral sentiment has not much meaning in this project so data in the neutral category are added to the negative dataset. All the positive tweet data are labeled as 1. Negative and neutral text data are labeled as 0. This becomes a binary classification problem.

The pre-processed text data is split into training and test set. I have set 80% of data as my training set and 20% for the test set. There are four machine learning models that are used for this project: 1) Logistic regression 2) Naïve Bayes 3) Support vector machine 4) XGboost. Each model is tested with count vectorizer and Tf-idf vectorizer. Grid Search is applied to models with two-fold cross validation to find the best parameters for better positive and negative word classification.

For Logistic regression model, C and solver parameters are tuned. C parameter controls the penalty strength and lbfgs and liblinear are selected for the solver parameter. Grid search results show that the best parameters are C with the value of 1 and solver with lbfgs. For Naïve Bayes model, MultinomialNB classifier is used as it is suitable for discrete feature classification. Hyperparameters chosen for support vector machine are C, kernel, and gamma parameters. Grid search results show that C with 0.1, kernel with linear and gamma with scale are the best parameters. For XGboost model, n_estimators and learning rate hyperparameters are tuned and best parameters for XGboost are learning_rate with 0.1 and n_estimator = 100.

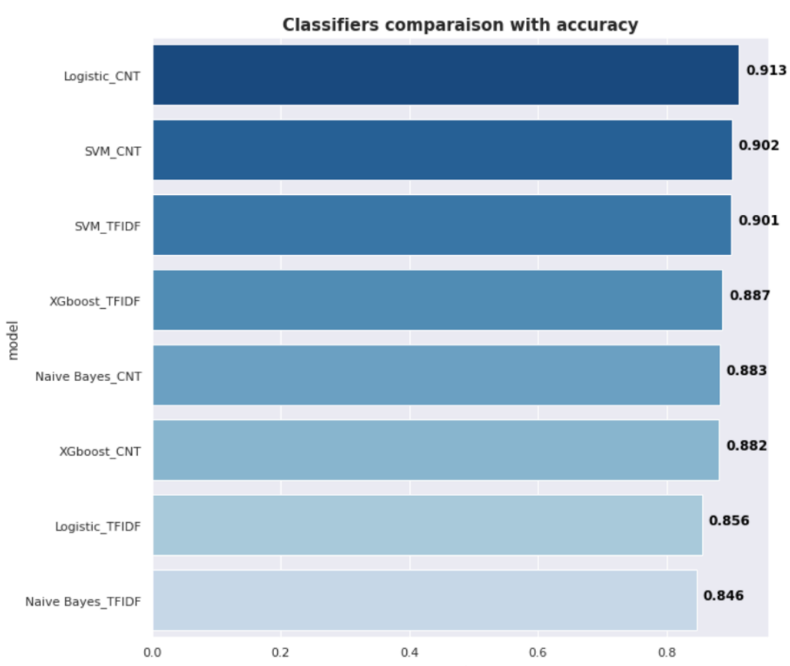



Figure 10: Accuracy of classifiers

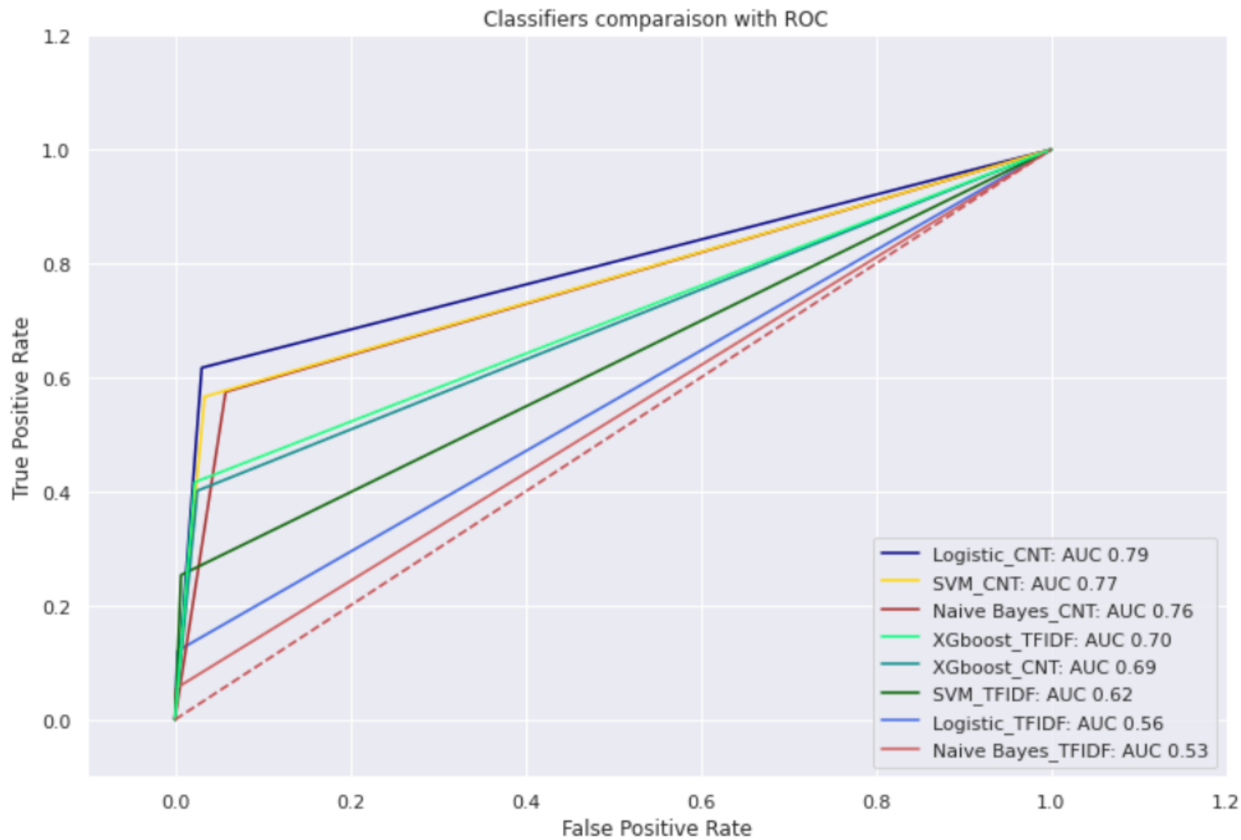



Figure 11: Classifier comparison with ROC curve

Results from Figures 10 and 11 show that **Logistic Regression classifier with count vectorizer** works the best in classifying positive and negative words from tweets. Using logistic regression model, I have drawn top 20 positive words and negative words from tweets. The words are displayed with its coefficient in bar charts. This visualization is generated using Tableau (Figure 13 and 14). The lists of positive and negative words are similar to the words seen in the previous Word Cloud (figure 8). This shows that the model classify sentiment well. Positive words are {Thank, awesome, excellent, great, amazing, wonderful}. Negative words are {worst, rude, hours, nothing, running, paid, hold}. Positive words in tweets are mostly positive expressions but it is hard to investigate meaningful insights from a single expression. On the other hand, most of negative words are more meaningful than positive words. Most of the negative words address customer service, flight, fare rate, and policy issue.

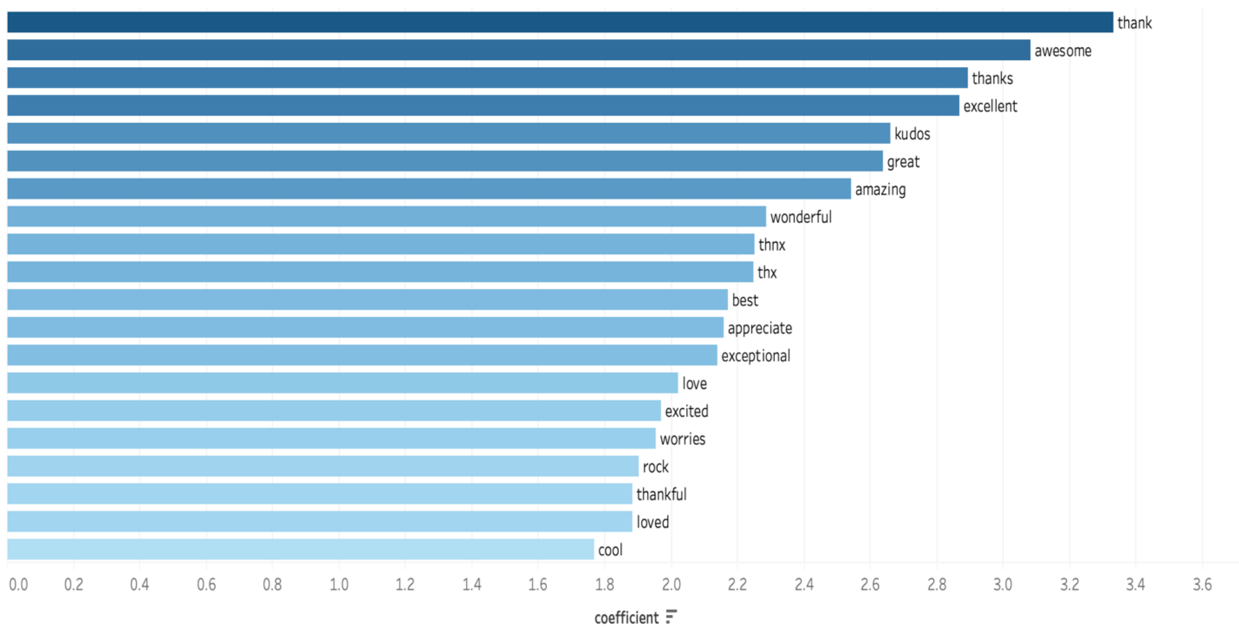



Figure 13: Top 20 positive words classified by logistic regression model with count vectorizer

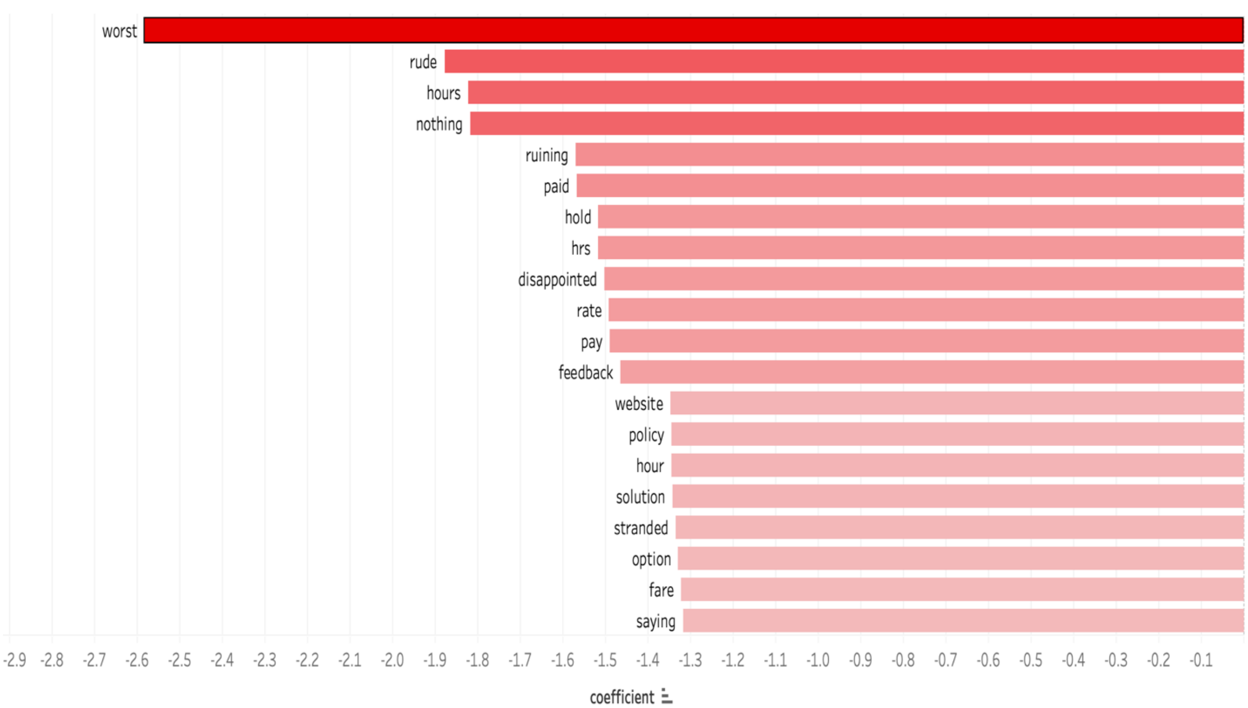
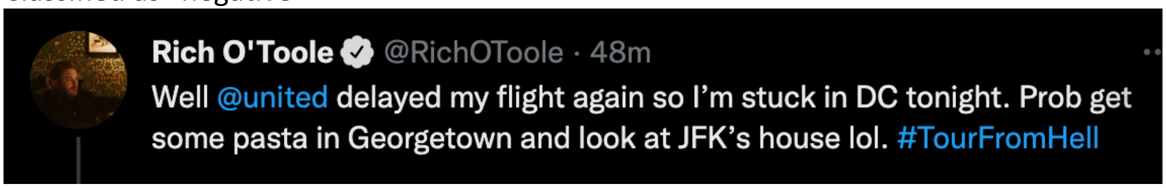



Figure 14: Top 20 negative words classified by logistic regression model with count vectorizer

I have also tested the model with the recent tweets about united airline.

Classified as "negative"



Classified as "negative"

Classified as "positive

Classified as "positive"

- **Keyword analysis**

To further understand keywords hidden behind tweets, the most similar words for top 10 positive and negative words are examined using Doc2Vec model. While Word2Vec computes a feature vector for every word in the corpus, Doc2Vec computes a feature vector for every document in the corpus. Doc2vec model is based on Word2Vec, with only adding another vector (paragraph ID) to the input. Using genism library's Doc2Vec.most_similar feature, most similar words of top 10 positive and negative words are examined. This method computes cosine similarity between the weight vector of input and the vector of each word in tweets. The result of a sample run for doc2vec.most_similar("rude") is shown in the Table 2. I applied this most_similar feature to the top 10 positive and negative words that are retrieved from the previous logistic regression model and plot them using TSNE plot (Figure15). T_SNE is a non-linear technique primarily used for data exploration and visualizing high-dimensional data.

| **Similar words** | **similarity** |
|---|---|
| unhelpful | 0.80477 |
| helpful | 0.79054 |
| employees | 0.75781 |
| pleasant | 0.75437 |
| tisk | 0.74888 |
| staff | 0.71573 |
| desk | 0.70736 |
| unprofessional | 0.70372 |
| jacquie | 0.70342 |
| named | 0.69359 |

Table 2: Similar words of "rude" obtained using doc2vec model

Finding a set of similar words can be more helpful in finding hidden meanings in the data than a single word. For example, the word "**rude**" which is one of the negative words obtained from logistic model have similar words such as "unhelpful", "staff", "employees", "desk" and "unprofessional". From this set of words, we can interpret that airline customers feel negative emotions when airline employees are rude and being unprofessional. Also, the negative word '**hold'** has similar words such as "phone", "disconnected", "hung", "hang" and "cut". These words can imply that customers are unhappy when they cannot reach customer service line quickly and when phone gets disconnected while on the phone. If airlines enhance their customer-facing service policy and add more employees and call-back services in the help-line department, it will reduce the number of complaints and improve customer satisfaction rate.

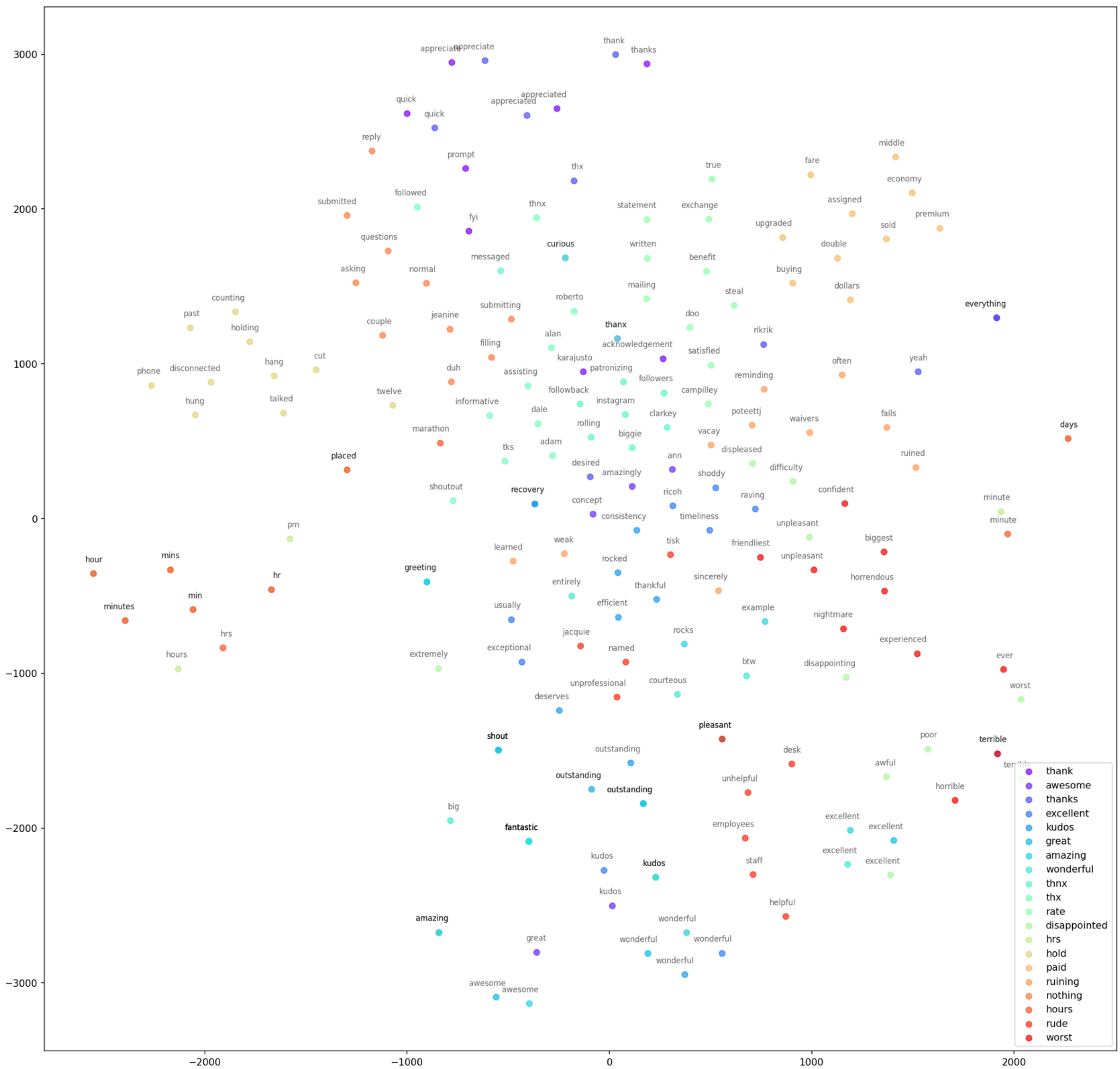



Figure 15: t-SNE plot of similar words for top 10 positive and negative words

**Conclusion**

While I was able to run in-depth analysis on negative tweets, it was difficult to find meaningful insights from positive tweets due to a relatively small number of positive tweets. If more positive tweets could have been provided, I would have been able to find factors that customer liked about airline. Among various machine learning algorithms, logistic regression performed the best in classifying sentiments from tweets. Using the model, sentiments of any future tweets can be identified with the accuracy value over 91%. Classified tweets can further be used to analyse customer experience and rooms for improvement.

**Reference:**

"Summary of passenger and freight traffic". World Air Transport Statistics. IATA. 2016. Archived from the original on 2016-07-05. Retrieved 2016-07-06. https://web.archive.org/web/20160705112926/http://www.iata.org/docx/WATS_2016-infographic.pdf

Scott, H. H. (2020, May 31). Visualizing Word2Vec Embeddings with tSNE. *Scottergories.* https://hedges.belmont.edu/scottergories/jupyter/2020/05/31/Visualizing-Word2Vec-Word-Embeddings-Using-tSNE.html.

**Github code link:**
https://github.com/katehee/data_viz_project