Athena Zhang
Jun 8 · 8 min read · Listen

Save

Airline Customer Sentiment Analysis about COVID-19

Global business has been heavily affected by COVID-19 in the past three years, especially the traveling industries and services. As we enter post-COVID time now and the number of traveling passengers starting to increase, we want to examine COVID-19's influences on airline industries by analyzing customer sentiments for airline companies.

and can be find here: https://github.com/AthenaZhang/NLP-Airline-Sentiment-Analysis.git

Objectives

Our primary goal is to understand the features that impact passenger satisfaction by analyzing Skytrax customer reviews, especially for some aspects related COVID-19. Then develop a model that can be adopted by the airline companies, which could potentially help them to identify areas for improvement in their pandemic safety protocols.

An example of a customer review from Skytrax during COVID pandemic:

2/10

"They are not enforcing masks"

Elizabeth Aaberg (United States) 23rd November 2020

✅ *Trip Verified* | I am appalled at United's decision to have the plane at full capacity during this pandemic. Cases are rising astronomically in the US. They are not enforcing masks and man sat coughing in the seat in front of me the whole flight with his mask down below his nose and they did nothing. This company really cares only about the money it makes and not about its employees or it's passengers. I have been a United card holder and member for years. I will be switching airlines moving forward. The staff was nice which was their only saving grace this flight.

Type Of Traveller	Solo Leisure
Seat Type	Economy Class
Route	Chicago to Portland
Date Flown	November 2020
Seat Comfort	★☆☆☆☆
Cabin Staff Service	★★★★★
Ground Service	★☆☆☆☆
Value For Money	★☆☆☆☆
Recommended	✖

Data Descriptions

To obtain our data, we first selected 5 airlines that we are interested in, which are Alaska Airlines, American Airlines, Delta Airlines, Southwest Airlines and United Airlines. Then

```
for page in US_airline_pages:
    driver.get(page)
    time.sleep(10)
    try:
        print("="*40)   #Shows in terminal when a new airline is being scraped
        print("Scraping " + page)

# Find total number of reviews for the airline
# Turn value into a float
# Each page defaults to showing 10 reviews, so take the ceiling of the total number of reviews divided by 10
# to get the number of pages of reviews for the airline
        review_count = driver.find_element_by_xpath('//div[@class = "rating-totals"]//span[@itemprop = "reviewCount"]').text
        review_count = float(review_count)
        n = int(ceil(review_count/10))

        # Iterate through all the pages of reviews for the airline in question
        index = 1
        while index <= n:
            driver.get(page + "page/" + str(index) +'/')
            time.sleep(5)

            try:
                print("Scraping Page number " + str(index)) ——»# Shows in terminal when a new page of reviews is being scrap
                index = index + 1

               #Find all the reviews:
                reviews = driver.find_elements_by_xpath('//article[@itemprop = "review"]')
                for review in reviews:

                    # Initialize an empty dictionary for each review
                    review_dict = {}
```

web scraping code

We scraped 17 features from the website, they are different categories that filled by customers in their reviews, some of them are categorical variables like ‘airline’ and ‘cabin’, some are numeric variables like ‘Seat Comfort’ and ‘cabin_service’ of rating scale 1–10, and the ‘recommend’ column is a binary variable that shows the customer recommended or not.

```
Scraping http://www.airlinequality.com/
Scraping Page number 1
Scraping Page number 2
Scraping Page number 3
Scraping Page number 4
Scraping Page number 5
Scraping Page number 6
Scraping Page number 7
Scraping Page number 8
Scraping Page number 9
Scraping Page number 10
Scraping Page number 11
Scraping Page number 12
Scraping Page number 13
Scraping Page number 14
Scraping Page number 15
```

```
review_dict['airline'] = airline
review_dict['overall'] = overall
review_dict['author'] = author
review_dict['review_date'] = review_date
review_dict['customer_review'] = customer_review
review_dict['aircraft'] = aircraft
review_dict['traveller_type'] = traveller_type
review_dict['cabin'] = cabin
review_dict['route'] = route
review_dict['date_flown'] = date_flown
review_dict['seat_comfort'] = seat_comfort
review_dict['cabin_service'] = cabin_service
review_dict['food_bev'] = food_bev
review_dict['entertainment'] = entertainment
review_dict['ground_service'] = ground_service
review_dict['value_for_money'] = value_for_money
```

Then we created 2 features by ourselves, 'COVID-related' and 'sentiment'. Like shown in the screenshots below, the variable 'COVID-related' is to identify whether the review is related to COVID, we identified that by seeing if the review contains keywords like "COVID|covid|mask|masks|pandemic). And the binary variable 'sentiment' is to identify whether the review is positive or neutral/negative. If the overall rating is over 7 then it's positive and marked as 1, otherwise it's neutral or negative and marked as 0.

```
import numpy as np
df['COVID_related'] = df['reviews_only'].str.contains('COVID|covid|mask|masks|pandemic')
df['COVID_related'] = np.where(df['COVID_related'] == True, 1, 0)
```

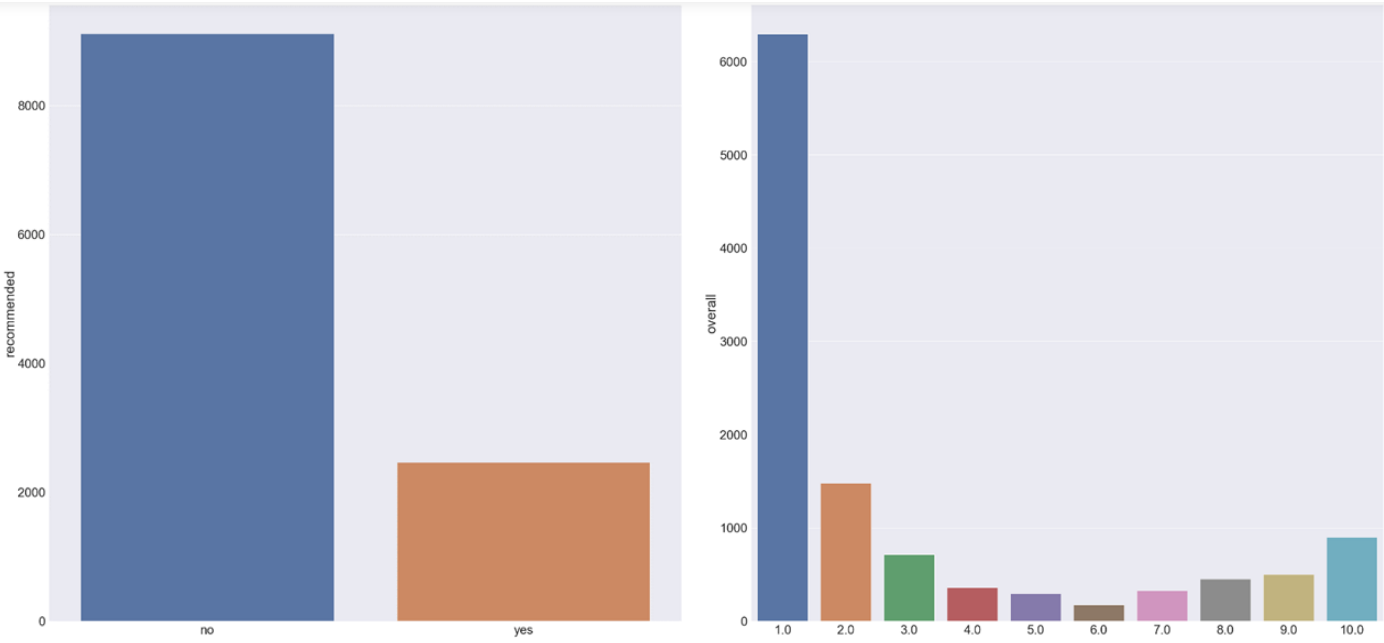
```
df['sentiment'] = np.where(df['overall'] > 7, 1, 0)
df.head()
```

Then we split the data to a training set (80%) and a test set (20%) for our modeling.

Methodology

EDA (Exploratory Data Analysis)

Generally, we found there are much more negative reviews than positive ones. For example, from the below (left) chart of 'recommended' column, we can see there are a lot of 'no'(more than 9000) and less 'yes' responses (more than 2000). Also, looking at the 'overall' scores below (right)that given by customers, we found that the average overall score is very low and 1.0 is the most common one.

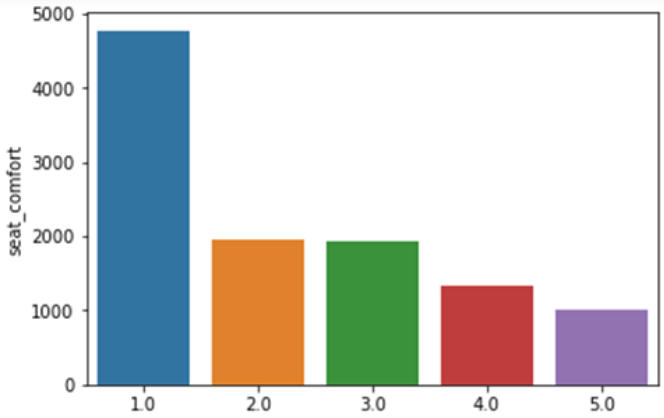


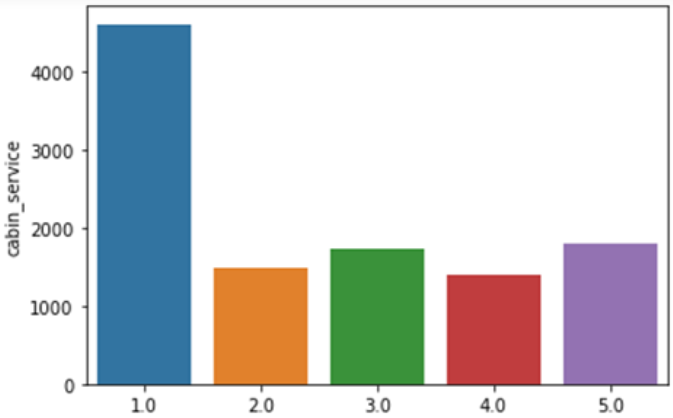
'Recommend' and 'Overall'

As for the traveler type and cabin classes, we found that Solo Leisure is the most common type followed by Couple Leisure, and Economy Class is the major cabin in all customer reviews.

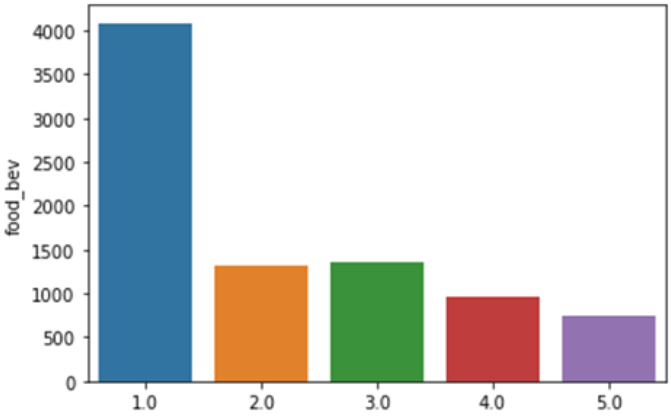
'Traveler type' and 'Cabin'

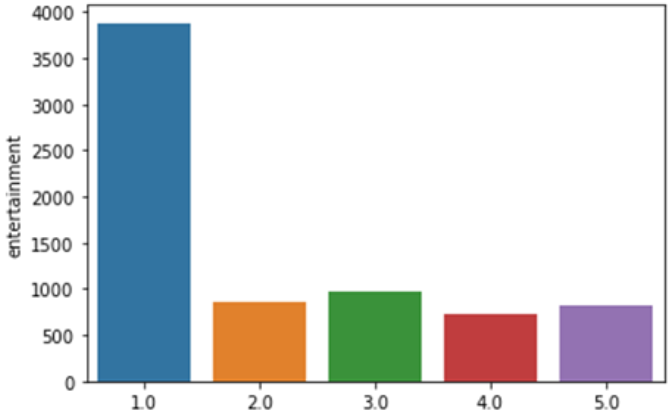
We also analyzed several rating columns like seat comfort, cabin service, food/beverage, entertainment, ground service and value for money. Generally, they all have very low average score around 2, which shown most of customers are not satisfied with theses services.



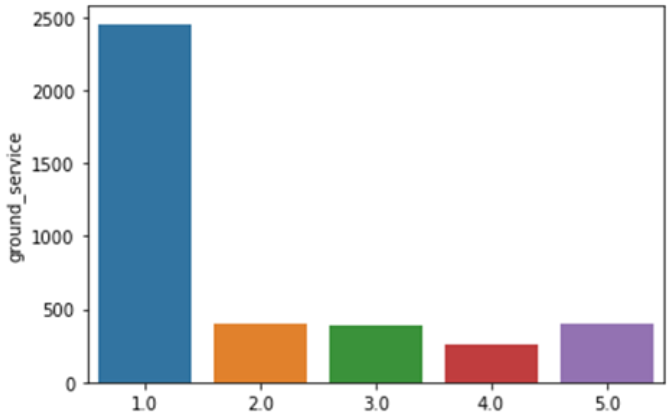


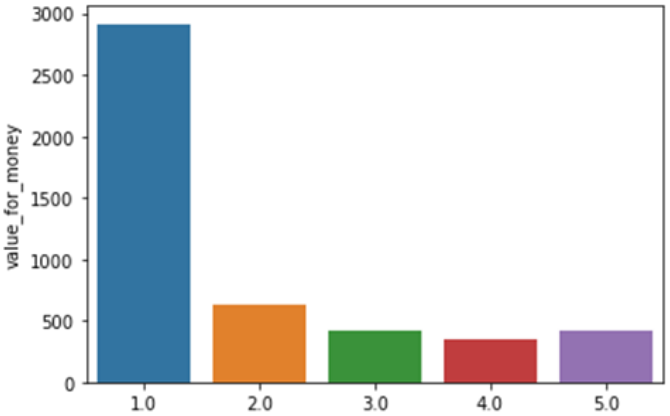
'Seat comfort' and 'Cabin service'





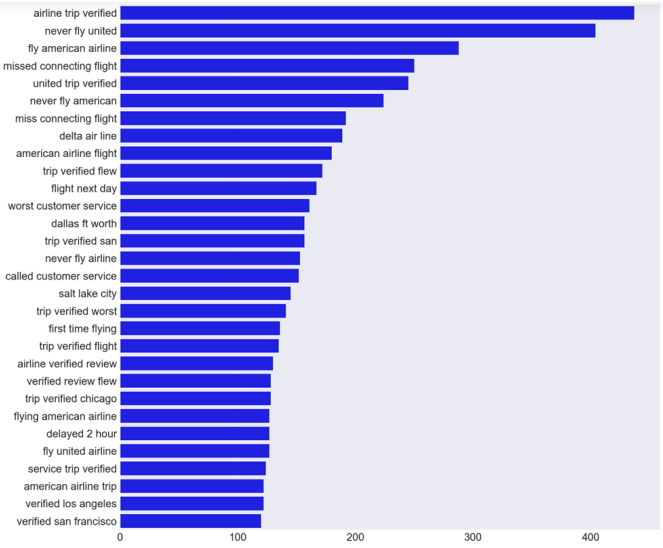
'Food/Bev' and 'entertainment'

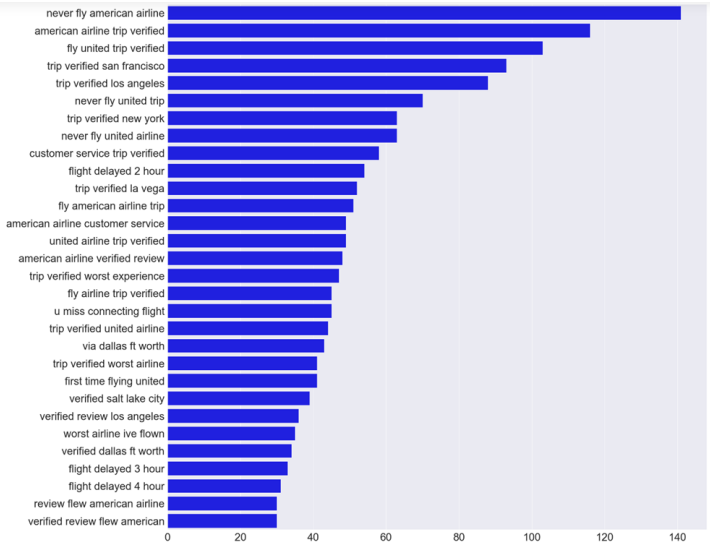




'Ground service' and 'Value for money'

We then tried N-gram analyses to find the most common topics, generally, most of the top topics are all negative sentiments like "never fly American airline" and "never fly united":





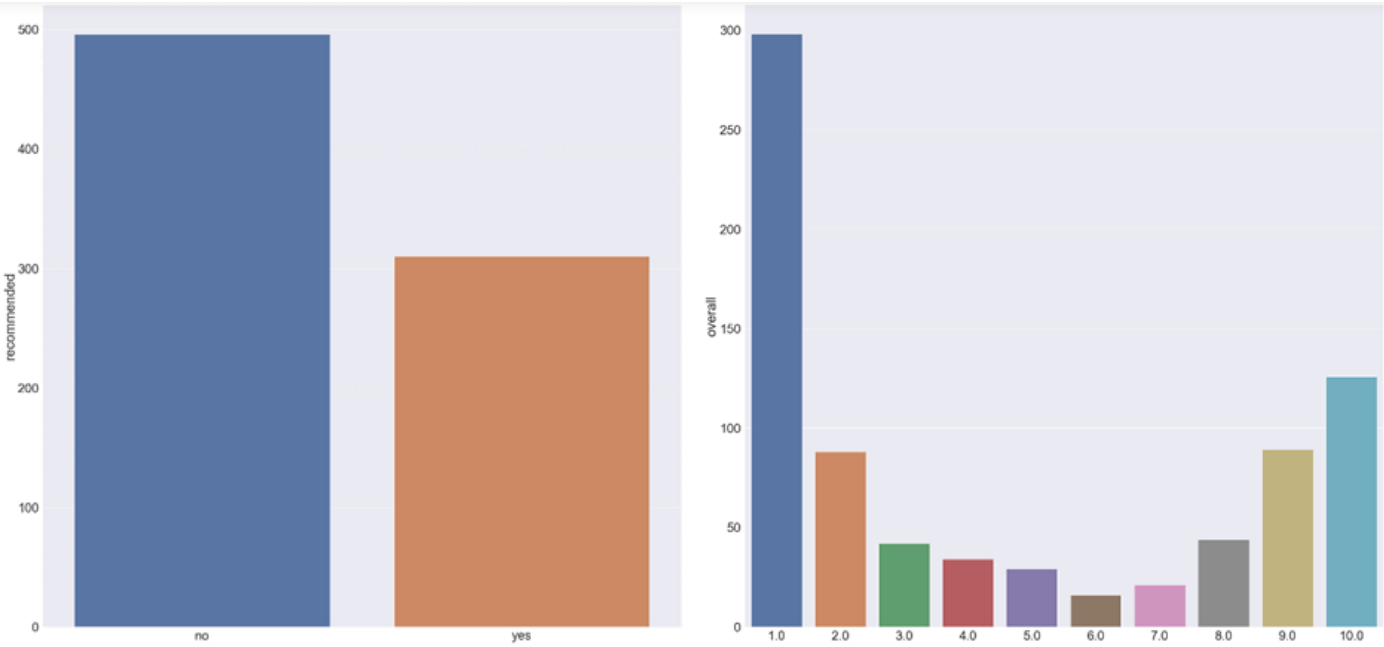
There are also some COVID-related N-gram topics:

```
{'flight due covid19 united': 4,

'cancelled flight due covid19': 4,

'cancelled due covid refund': 4}
```

Besides the EDA for whole dataset, we also did EDA for each airline to compare their performance. Specially, Alaska has the most percentage of ‘Recommend’ and positive score for ‘Overall’ rating comparing to other airlines. But one of the reasons could be Alaska has the lowest number of customer reviews:



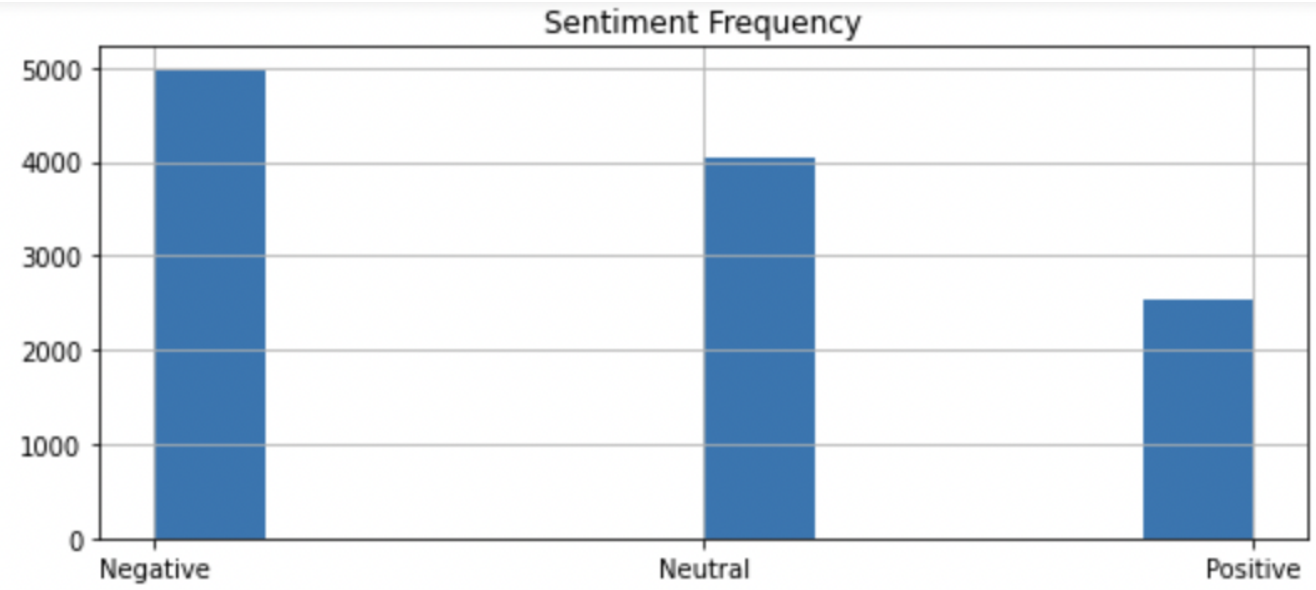
'Recommend' and 'Overall' for Alaska

Textblob

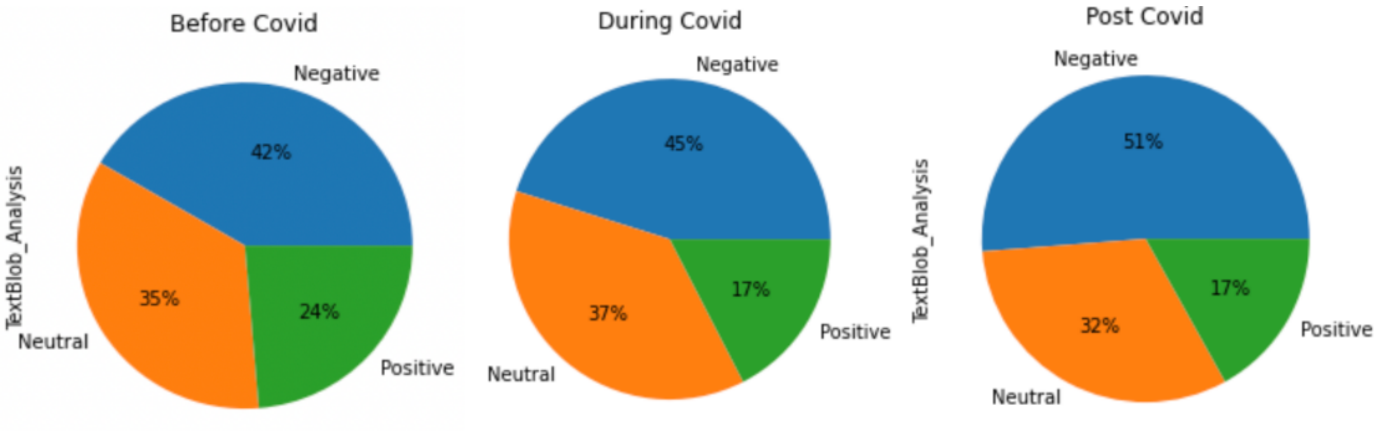
We use Textblob to do sentiment analysis, and got subjectivity and polarity of each customer's review. From those scores, we derived customers' sentiments of positive, negative, and neutral:

	airline	review_date	customer_review	TextBlob_Subjectivity	TextBlob_Polarity	TextBlob_Analysis
0	United Airlines	2022-05-21	trip verified worst experience i have ever h...	0.595370	-0.322222	Negative
1	United Airlines	2022-05-21	trip verified very displeased with the fact ...	0.335714	0.106429	Neutral
2	United Airlines	2022-05-21	not verified absolutely terrible experience b...	0.518182	-0.390909	Negative
3	United Airlines	2022-05-21	trip verified we were scheduled to leave pit...	0.498347	0.092890	Neutral
4	United Airlines	2022-05-19	trip verified first leg of trip to kul dread...	0.463757	0.057421	Neutral
...	...	...	...	...	...	...
11568	American Airlines	2013-12-09	flew from seattle on december 3rd arriving to ...	0.462955	-0.120530	Negative
11569	American Airlines	2013-12-07	traveled one-way moving back home from miami t...	0.578472	0.145833	Neutral
11570	American Airlines	2013-12-06	i would never have booked a flight with americ...	0.403333	-0.013333	Negative
11571	American Airlines	2013-12-05	had to cancel my trip because of american's mu...	0.439924	-0.040076	Negative
11572	American Airlines	2013-12-04	flight that we were boarding in denver arrived...	0.532692	-0.128846	Negative

11573 rows × 6 columns



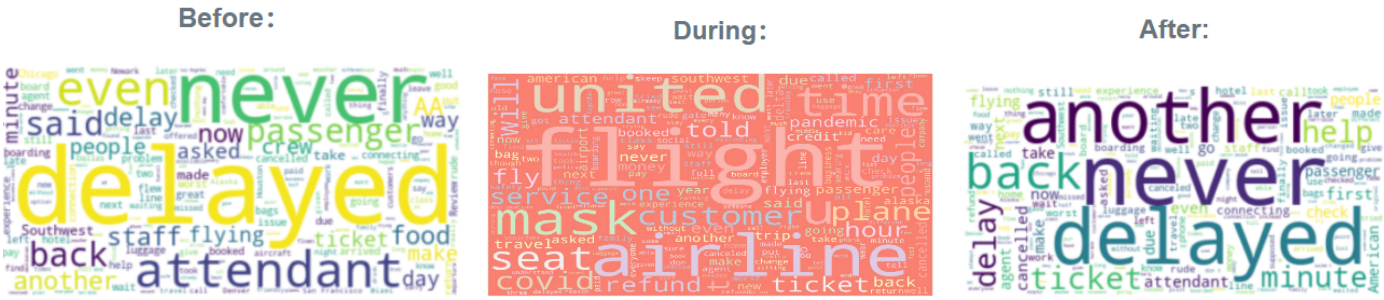
We defined three periods of COVID spread. Before 01/01/2020, the virus had not yet begun to spread, or people were not yet aware of its harmful effects so there was no corresponding policy. During 01/01/2020–12/27/2021, the virus has spread wildly and mutated into many variants. Each region has corresponding measures for travel and health protection. After 12/27/2021, the CDC has relaxed restrictions on isolation and mask wear. This indicates that the pandemic is nearing an end. We can see that as these three periods pass over time, the negative passenger reviews are increasing, and the positive reviews are decreasing. This indicates that the COVID-19 has a clear negative impact in the airline field:



Before 01/01/2020 01/01/2020 - 12/27/2021 After 12/27/2021

Word Cloud

Next method we used was Word Cloud, to find the common keywords in the three COVID periods. From the images below, we can see there are many COVID related words like “mask”, “refund” or “pandemic” in the ‘During’ period, while “delay” and “never” are the most common complaints for ‘Before’ and ‘After’ periods. Although one of the reasons could be ‘During’ is one short period of time that has less travelers and reviews, customers definitely commented a lot about COVID policies and related services in their reviews during the pandemic.



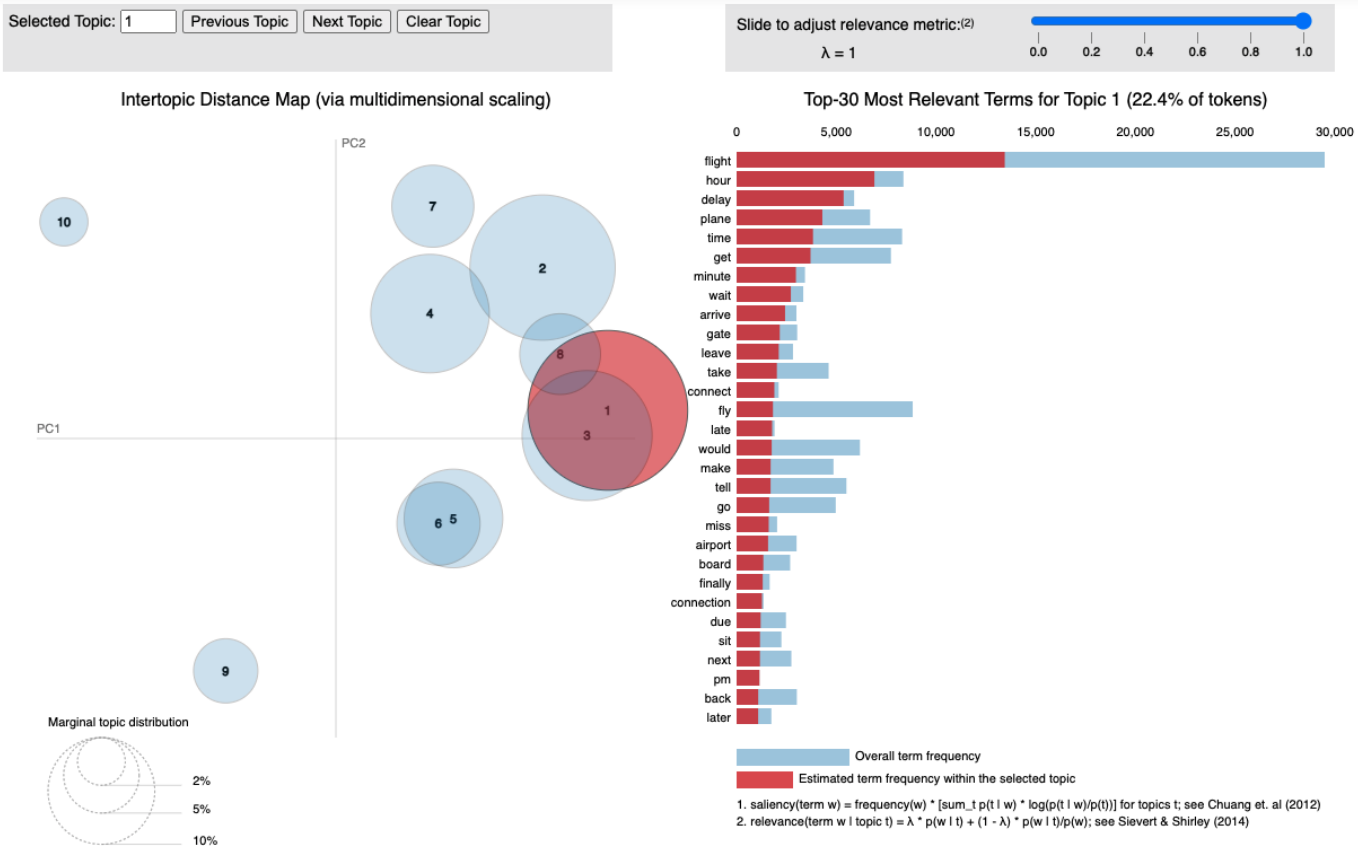
Word Cloud

LDA Model

We used LDA topic model to pick the top topics and understand customers’ propensity. For the number of topics we choose, we tried several numbers, and the number of topics 10 gave us the highest coherence score of 0. 4212:

```
[(0,
  '0.146*"seat" + 0.029*"pay" + 0.024*"sit" + 0.021*"plane" + 0.018*"extra" + '
  '0.017*"row" + 0.014*"together" + 0.014*"get" + 0.013*"board" + '
  '0.013*"first"'),
 (1,
  '0.140*"bag" + 0.125*"check" + 0.055*"luggage" + 0.039*"baggage" + '
  '0.029*"carry" + 0.013*"overhead" + 0.013*"fit" + 0.011*"space" + '
  '0.010*"deliver" + 0.010*"item"'),
 (2,
  '0.045*"flight" + 0.036*"fly" + 0.029*"airline" + 0.025*"time" + '
  '0.014*"make" + 0.011*"american" + 0.010*"travel" + 0.010*"year" + '
  '0.009*"trip" + 0.009*"last"'),
 (3,
  '0.035*"say" + 0.034*"ask" + 0.028*"go" + 0.025*"tell" + 0.020*"line" + '
  '0.017*"help" + 0.017*"get" + 0.016*"staff" + 0.016*"could" + '
  '0.013*"people"'),
 (4,
  '0.039*"passenger" + 0.015*"safety" + 0.014*"attendant" + 0.010*"crew" + '
  '0.010*"game" + 0.009*"music" + 0.009*"play" + 0.009*"collect" + 0.009*"bed" '
  '+ 0.008*"adequate"'),
 (5,
  '0.086*"flight" + 0.044*"hour" + 0.034*"delay" + 0.028*"plane" + '
  '0.025*"time" + 0.024*"get" + 0.019*"minute" + 0.017*"wait" + 0.016*"arrive" '
  '+ 0.014*"gate"'),
 (6,
  '0.043*"flight" + 0.025*"ticket" + 0.024*"call" + 0.024*"customer" + '
  '0.021*"would" + 0.021*"tell" + 0.018*"book" + 0.016*"service" + '
  '0.013*"could" + 0.013*"day"'),
 (7,
  '0.070*"service" + 0.068*"customer" + 0.055*"airline" + 0.049*"bad" + '
  '0.041*"fly" + 0.029*"experience" + 0.029*"never" + 0.027*"ever" + '
  '0.025*"rude" + 0.020*"staff"'),
 (8,
  '0.046*"cancel" + 0.043*"day" + 0.042*"hotel" + 0.042*"flight" + '
  '0.030*"voucher" + 0.023*"next" + 0.021*"night" + 0.018*"due" + '
  '0.016*"airport" + 0.014*"american"'),
 (9,
  '0.043*"flight" + 0.030*"seat" + 0.024*"class" + 0.023*"food" + 0.021*"crew" '
  '+ 0.020*"service" + 0.019*"good" + 0.018*"cabin" + 0.015*"first" + '
  '0.012*"drink"')]
```

We printed out the keywords for each topic, and then we visualized the topic content according to the keywords. For example, topic 1 has keywords like seat, sit and row, hence we can say that airline reviews under topic 1 are about seating. The top three topics that customer most talking about are “Delay” (22.4%), “Travel experience” (18.5%), and “Booking service” (14.8%). Through the LDA topic model, the attention to airline-related



Prediction Modeling

Before building our model, we first did some preprocessing like normalize, tokenize, remove stop words and lemmatize:

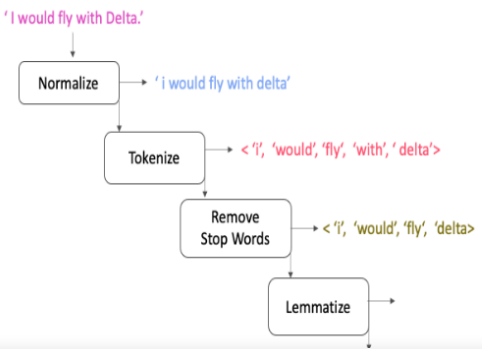
```
def CleanText(text):

    # Remove non-character
    letters = re.sub("[^a-zA-Z]", " ", str())

    # Convert to lower case
    lower = letters.lower().split()

    # Remove stopwords
    stops = set(stopwords.words("english"))
    words = [w for w in lower if not w in stops]

    # Lemmatization
```



Then we build a logistic regression model as the baseline, see whether TF-IDF or count on the validation set performed better. In our case, the model using CountVector has a higher accuracy and F1 score for class 1, so we use the CountVector model to test further.

```
Result Using CountVector:

Accuracy on validation set: 0.9033

AUC score : 0.7199

Classification report :
              precision    recall  f1-score   support

           0       0.91      0.98      0.95      1569
           1       0.84      0.46      0.59       283

    accuracy                           0.90      1852
   macro avg       0.87      0.72      0.77      1852
weighted avg       0.90      0.90      0.89      1852


Confusion Matrix :
 [[1544   25]
 [ 154  129]]
```

```
Result Using TFIDF:

Accuracy on validation set: 0.8947

AUC score : 0.6656

Classification report :
              precision    recall  f1-score   support

           0       0.89      1.00      0.94      1569
           1       0.93      0.34      0.49       283

    accuracy                           0.89      1852
   macro avg       0.91      0.67      0.72      1852
weighted avg       0.90      0.89      0.87      1852


Confusion Matrix :
 [[1562    7]
 [ 188   95]]
```

Vectoization

Based on that, we built three models: Weighted Logistic Regression, Naïve Bayes, and Random Forest:

Weighted Logistic Regression:

```
{'class_weight': {0: 1, 1: 1}} 0.8786114900298063

Accuracy on validation set: 0.8947

AUC score : 0.6656

Classification report :
              precision    recall  f1-score   support

           0       0.89      1.00      0.94      1569
           1       0.93      0.34      0.49       283

    accuracy                           0.89      1852
   macro avg       0.91      0.67      0.72      1852
weighted avg       0.90      0.89      0.87      1852


Confusion Matrix :
 [[1562    7]
 [ 188   95]]
```

Naive Bayes:

```
Accuracy on validation set: 0.8785

AUC score : 0.6358

Classification report :
              precision    recall  f1-score

           0       0.88      0.99      0.93
           1       0.78      0.29      0.42

    accuracy                           0.88
   macro avg       0.83      0.64      0.68
weighted avg       0.87      0.88      0.85


Confusion Matrix :
 [[1546   23]
 [ 202   81]]
```

Random Forest:

```
Fitting 5 folds for each of 16 candidates, totalling 80 fits

Accuracy on validation set: 0.8747

AUC score : 0.5945

Classification report :
              precision    recall  f1-score   support

           0       0.87      1.00      0.93      1569
           1       0.95      0.19      0.32       283

    accuracy                           0.87      1852
   macro avg       0.91      0.59      0.62      1852
weighted avg       0.88      0.87      0.84      1852


Confusion Matrix :
 [[1566    3]
 [ 229   54]]
```

and we can see the Weighted Logistic Regression model has the best performance with an accuracy score around 90%.

First, we found the pandemic has reduced customer satisfaction in many aspects. Also, during the pandemic, flight delays and bad service quality were increased due to various policy restrictions and limited number of staff.

Thus, we recommend that airlines should improve their services and ensure more on-time flights after the pandemic, to attract more customers and also positive reviews. Besides, given that various variants of the virus are still spreading, our another recommendation for airline company is to continue maintaining a good level of health policy and pay more attention to sanitizing processes, so that customers feel safer traveling with them.

Challenges

Our first challenge during the project was to work on another person's code. Since people have different coding styles, so sometimes we had misalignments for the analysis results

processes, which bring us a lot challenges to explore and try different methods to finish our tasks.

Conclusion

We chose to perform sentiment analysis on airline reviews to understand the impact of different features on the airline industry. We used web scraping to get our data, then we performed EDA, Textblob, Word Cloud and LDA model to analyze our data. Then we ran Logistic Regression, Naïve Bayes and Random Forest to see how these models performed. We concluded that Logistic Regression with countvector had the highest accuracy of 90% in correctly predicting the sentiment of the comments.

After doing this project, we had a better understanding on class aspects, and become more familiar with NLP methods. Regarding the current global epidemic, we hope that our project could bring timely insights and areas of improvement to the airline industry, so that they could achieve better customer satisfaction rate and quickly recover for post-pandemic.