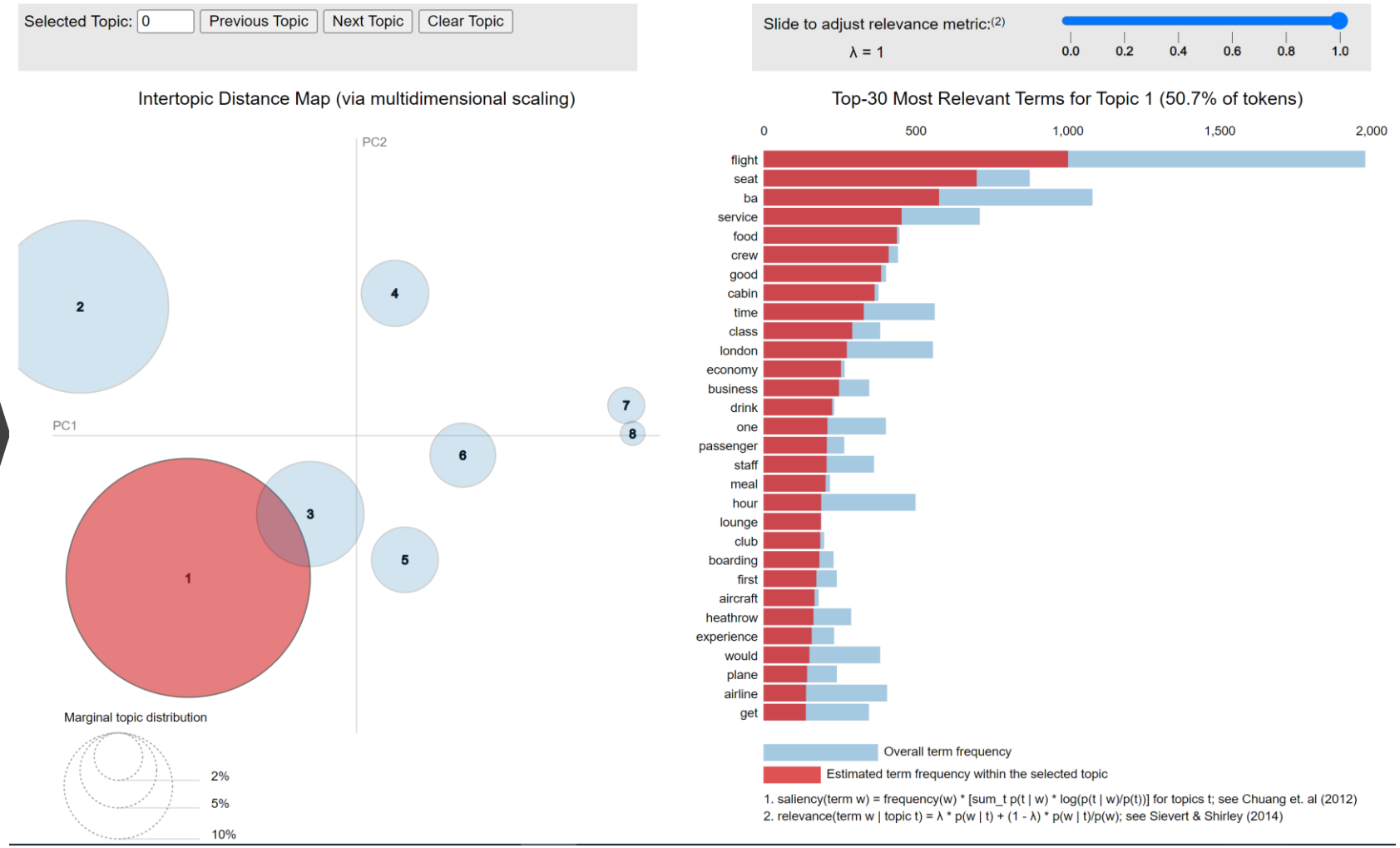
Web Scraping To Gain Company Insights

Wordcloud

The image below shows the most important words in the customer reviews at a glance. The bigger words are the most mentioned words. The next two slides show the relative frequency of use of these words. The ideal wordcloud for the BA brand would be one in which negative words such as refund, poor, and flight cancelled are so small they can't be easily spotted. I believe work needs to be done to reposition the brand in such a way that it will mostly elicit positive sentiments.

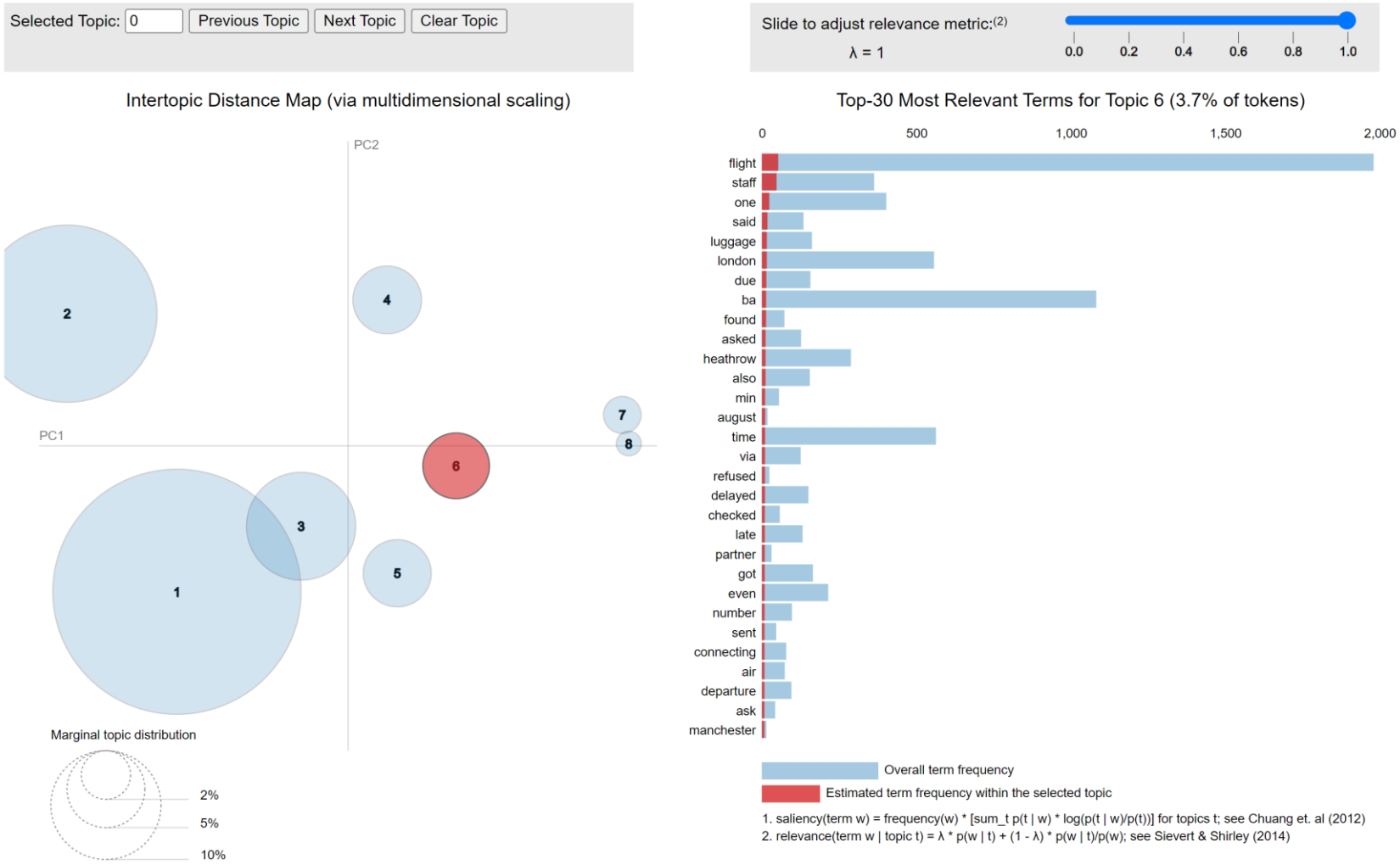
TOPIC MODELLING

The image on the right shows probabilistic distribution of words i.e. the most important words and their frequency in the reviews. It can be seen that, other than the brand name BA and terms like flight and seat which are unsurprisingly common, customers are mostly concerned about the quality of service, the food, the crew, and time. It also shows that the adverb, GOOD, is a frequently used term which probably indicates relative positive sentiment toward the brand.



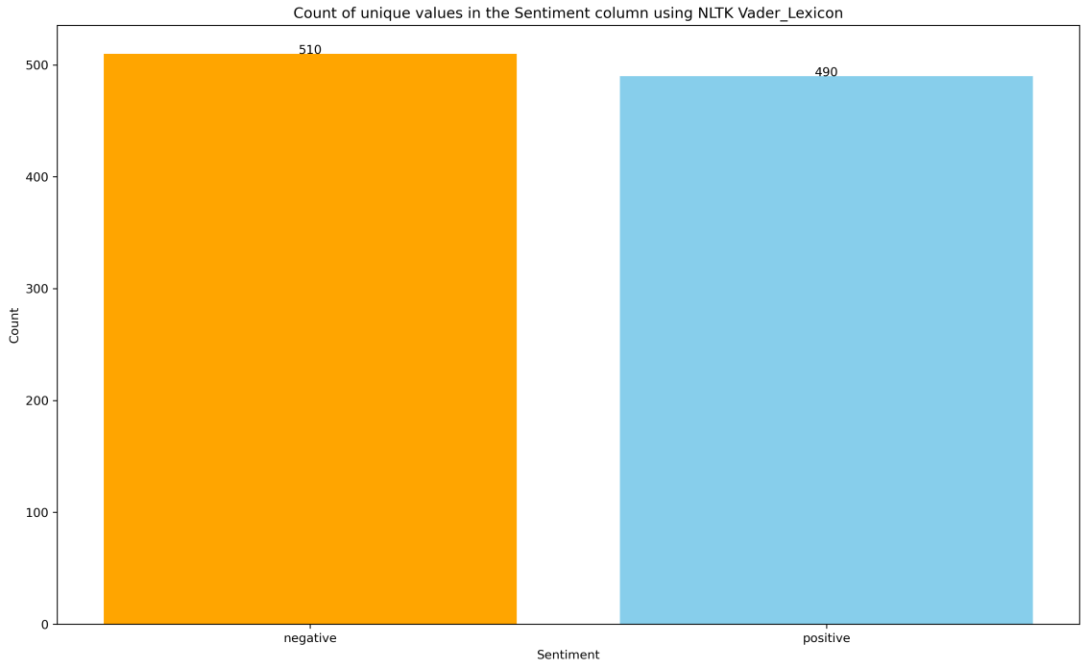
TOPIC MODELLING

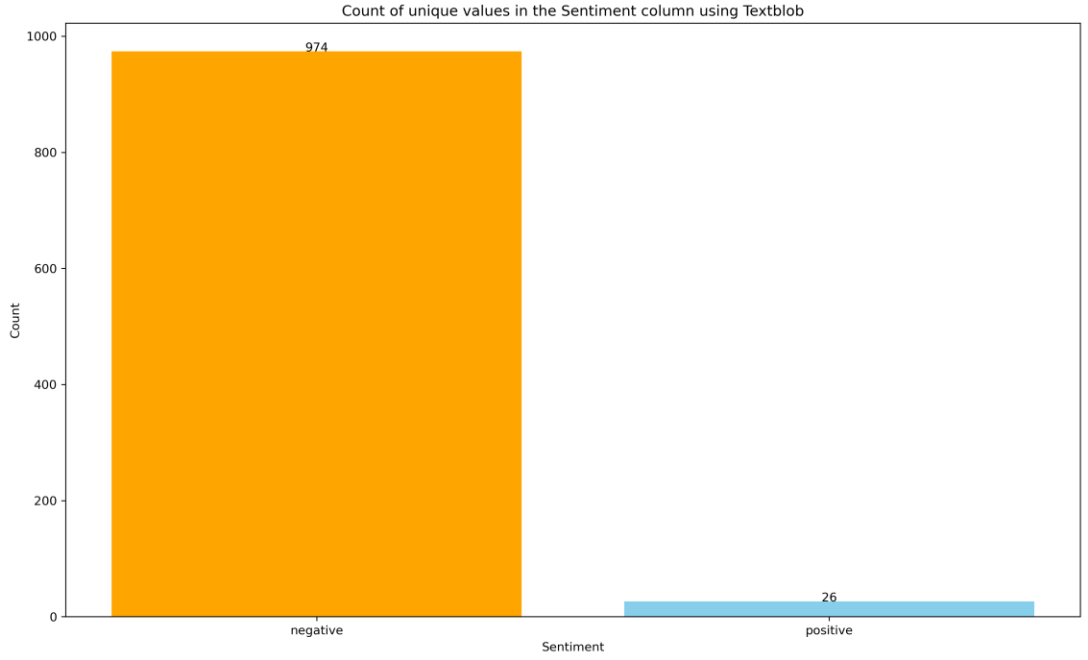
The image on the right shows probabilistic distribution of words. In this case, these words are of lower import to the customers than the words in the previous slide. It can be seen that there are more words with negative connotation. These are words such as refused, delayed, and late. Even though these negative words are used here by the customers, their relative importance is low. It is however important that the overeall sentiment of customers will be better conveyed using sentiment analysis which we would be looking at in the next couple of slides.



Sentiment Analysis

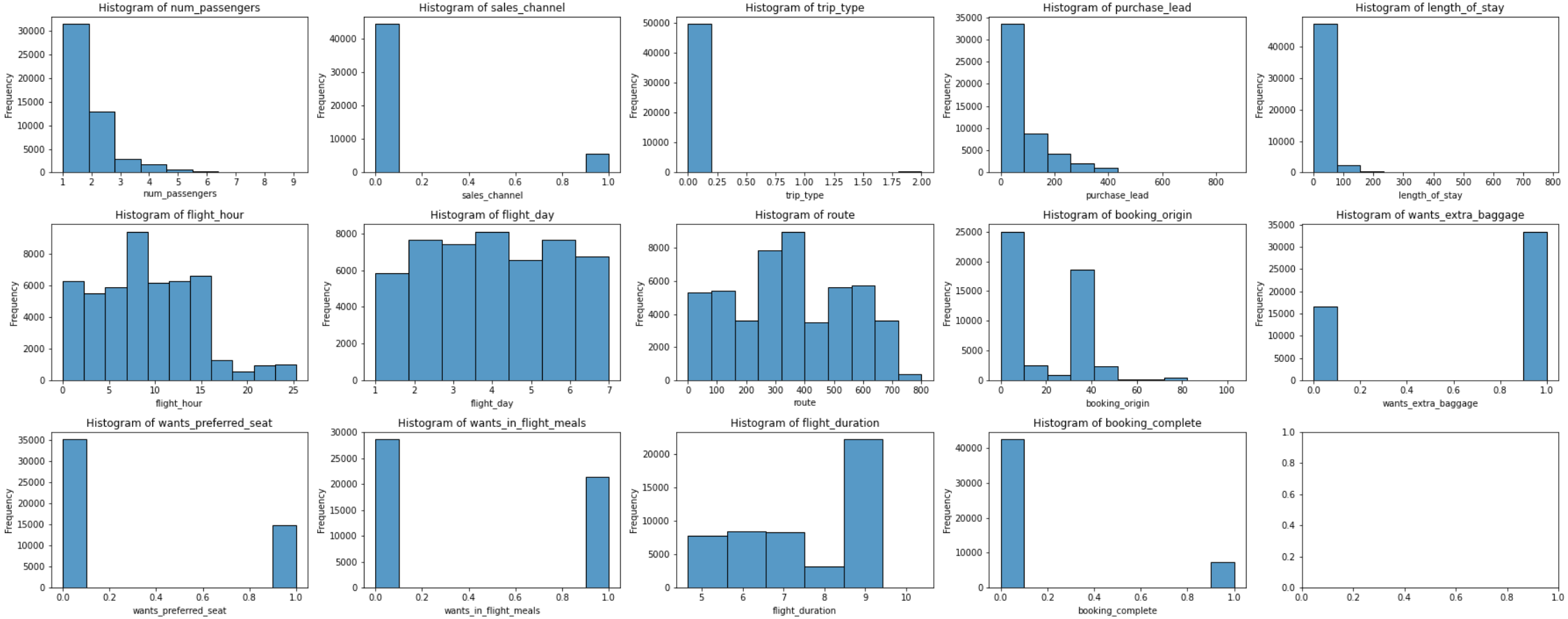
The barplots below is a visual representation of the cumulative negative and positive sentiments of the customer reviews. It can be seen that there are more negative sentiments than positive in both models. This net negative sentiment might been customers are generally unhappy with the service.





DATA DISTRIBUTION PER COLUMN

The graph of Histogram of num_passengers below shows there are more individual travellers (1-travellers) than any other n-travellers where n is the number of travllers. The higher n is the less we find n people traveling together in our dataset. It can also be seen that over 40,000 people make booking enquiry via the internet as opposed to those who use mobile which is less than 5000. There are also more round_trip potential bookings than circle_trip and one-way trip. There are also a lot of potential travelers who want extra luggage as opposed to those who are content with the standard luggage allowance. This is expected as these are potential holiday bookings.



Evaluation Metrics

Why do we choose Precision to Recall? Precision asks the question 'What proportion of positive identifications was actually correct?'. It takes into consideration the quantity of false positives. We don't want the classifier to mistakenly identify incomplete_booking as a complete_booking while. This is because BA is more concerned with our model's ability to accurately classify complete bookings. It is perharps more tolerable for some bookings to be wrongly classified as incomplete bookings than the other way round. In this case the precision of our tree classifier is 64%. Accuracy is not a reliable measure of the performance of the model since we have class imbalance in our data.

	precision	recall	f1-score	support
0	0.86	1.00	0.92	10651
1	0.64	0.05	0.08	1849
accuracy			0.86	12500
macro avg	0.75	0.52	0.50	12500
weighted avg	0.83	0.86	0.80	12500

FEATURE IMPORTANCE

The graph on the right shows the relative importance of the features to complete booking. As can be seen from the graph, Purchase Lead is the most important feature in determining whether a customer buys a holiday with BA or not.