Yelp Review Sentiment Analysis

The Problem:

The Black Pearl, a new hotel and casino plans to open up a location right on the Las Vegas Strip. As they continue to plan out the design and amenities for their hotel, they want to take a closer look at what the public is saying about their soon to be rivals in the area. They want to understand what topics are correlated with good and bad reviews, and to be able to classify sentiment of their own reviews on any platform to track their performance. By analyzing the yelp reviews of the current resorts and casinos on the strip they'll be able to gain insight into their competition, and the star ratings will provide a good benchmark to analyze sentiment moving forward.

The Data:

The Yelp public dataset is available online through their dataset challenge webpage, and is also available for download on Kaggle. The complete dataset contains 5 different tables, but the two that are being used in the scope of this project are the business table and the reviews table. The business table contains company specific information for the different businesses in the Yelp public dataset such as name, location, company star rating, and review count. Each business is given a unique business ID that allows you to join information among the different tables in the dataset.

The other table being used for analysis is the reviews tabel. This table contains reviews from 2005-2018 and each review has a unique review ID. The business ID the review was regarding, the text of the review, review rating, and the information on if that particular review was voted as cool, helpful, or funny is included as well.

Both of these tables were downloaded in a JSON format. The lines were read in one by one, and then converted to a JSON string. This JSON string was then used to create a dataframe using the .read_json() method available in Pandas.These steps were completed separately for both the business and review files. The next step was to filter out the business dataframe to only include businesses in the hotel category. This way when the business data is merged with the review data, only the reviews for hotels would be included. This made the review file, originally containing 4.39 GB of data, more manageable for analysis.

The filtered hotel data and the reviews data were then merged together on the unique business_id variable using an inner join. Because the The Black Pearl is specifically interested in resort casinos on the Las Vegas Strip, another step of filtering was done to make sure we were only working with joint hotel/casinos and only those located on the Las Vegas Strip.

As both the business and review dataframe had a star variable, those column names had to be changed to review rating and company rating to differentiate the two metrics, and avoid

confusion moving forward. Afterwards, the unneeded columns were dropped to make sure we were only focusing our analysis on useful information. The attributes column, containing valuable information for restaurant type businesses, was dropped along with state, city, and postal code information that was the same for every business in this view. Latitude and longitude were also dropped as the focus of this project is in one specific area. Null values were then filled in for the address field with the string 'Not Available' to make it clear this information was not provided. Lastly, the index was set to review_id, and a column called text length was added so the length of each review could be compared between different review ratings as well.

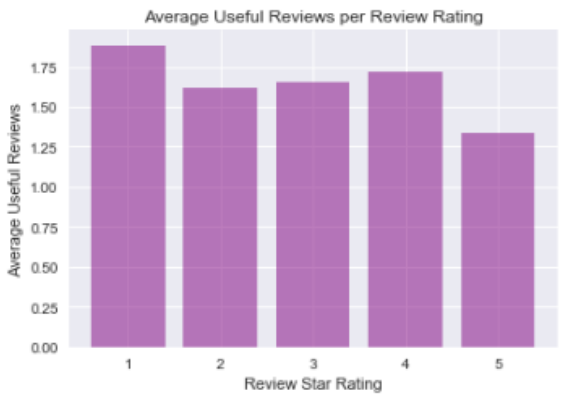
The final step before diving into the exploratory analysis was to check for any outliers in the data. I started by looking for any companies with 0 reviews, and although I didn't find any, I found a big range in the number of reviews per company with 3 as a minimum and 4,041 as the max. Although these are outliers, they do not indicate any errors, and were not moved from the dataset. Next I checked to make sure the review ratings and company ratings all fell within the 1-5 star possibilities. All of the ratings fell within the expected threshold, and I moved on to checking the useful, funny, or cool designations for the yelp reviews. Each of these 3 variables should only have be marked with a 1 or a 0 for each review. A 1 in the useful column means that the review was considered useful, for example, and 0 means it was not considered useful. Each of the 3 variables, useful, funny, and cool only contained either 1's or 0's as expected. Lastly, I took a look at the text length variable to check if there were any one or two word reviews. The minimum text length for any review was 15 which is still long enough to convey sentiment. At this point the data was considered ready to explore further.

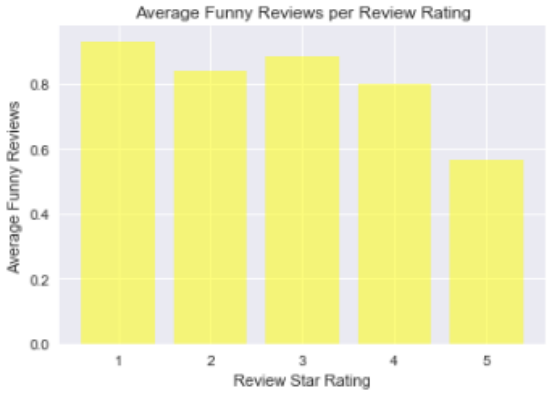
Analysis and Testing:

In order to get a better understanding of what contributes to high and low rated reviews, we'll first have to take a look at the makeup of our of our sample.The analysis started off by breaking down the total number of reviews by review rating, and I found that 1 and 4 star reviews are the most common in this dataset with 15,639 and 15,556 reviews respectively. 2 star reviews had the smallest number at 9,139.

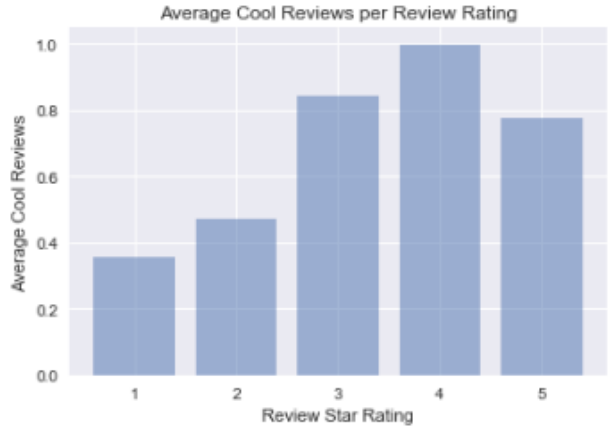
When I looked at the number of useful, funny, and cool labeled reviews by review rating I found that 1 and 4 star reviews contained the most amount of these designations. As the 1 and 4 star reviews had a bigger pool to draw from, I compared the averages of the useful, funny, and cool reviews across the star ratings as well.

When looking at the average usefulness across the different star ratings, 1 star reviews ranked the highest. Overall the lower rated reviews contained higher useful scores than higher rated reviews did. This was the case with the average funny reviews as well. 1 star reviews were again ranked the highest, and like the useful designation, funny reviews tended to have higher average scores in the low review ratings compared to the high ones.





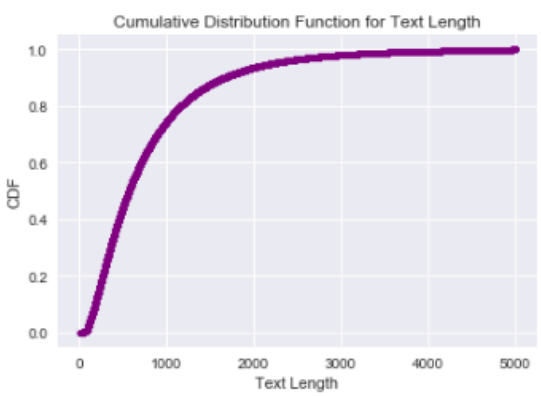
Cool reviews, however, had the highest average score in 4 star reviews. Unlike reviews marked as useful or funny, the cool reviews were found in greater numbers on the higher rated reviews.

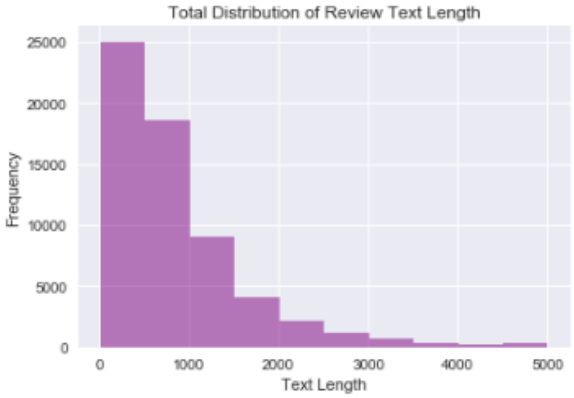


As the data seems to represent lower reviews being more frequently marked as useful or funny, and higher reviews more frequently marked as cool, t-tests were run to confirm whether or not their was a significant difference in these designations between high (above 3 stars) and low (below 3 star) review ratings. The t-tests were run for each of the 3 designations, and the results all pointed to a significant difference between the high and low rated review sets. After generating bootstrap samples for the mean, the probability of useful and funny designations being higher in lower rated reviews and cool designations being higher in higher rated results was tested, and the results confirmed our initial analysis.

The next variable of interest was the text length of each review. When we look at average text length by review rating we find an overall trend of shorter text lengths being associated with higher review ratings. It seems that when something has gone wrong for a customer, it requires more explanation than a positive experience does. There are a number of outliers for text length for each of the review ratings though , so a hypothesis test was conducted to determine if there truly is a significant difference in text length for high and low rated reviews.

The results of the t-test assuming the average text length was equal in high and low review ratings showed a p-value far below the .05 significance level. We can confidently say there is a significant difference in the text length between review ratings. The text lengths range from a minimum value of 15 to a maximum value of 5,000, but the average was found to be a length of 864.





A cumulative distribution function was generated for the text length variable and we see that 70% of the reviews fall below a length of 1000, and 40% of the reviews fall below 500. The distributions of review lengths holds the same shape with low, high, and the middle 3 star ratings.

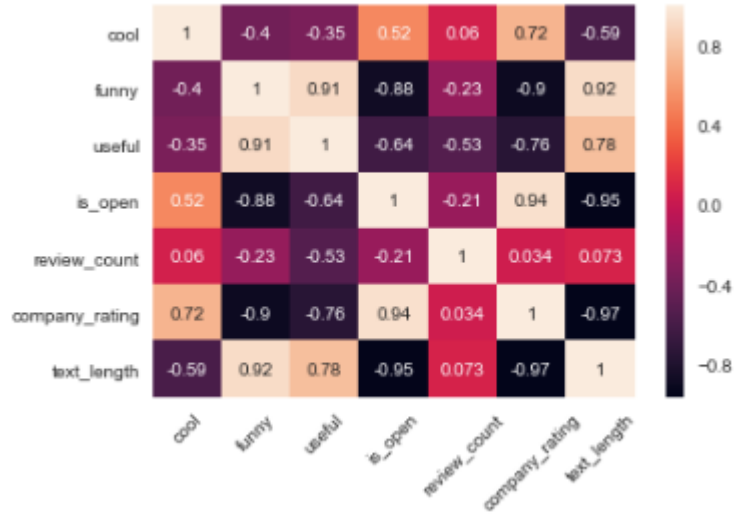
Looking at the average text length by company rating it became apparent that the higher text lengths were more frequently associated with lower rated companies. This builds off of our findings that text length is longer for lower star reviews. In regard to the number of reviews by company rating, you'll see that companies with ratings between 2.5 and 3.5 have the highest number of ratings. Companies rated on either extreme of the spectrum tend to have a fewer number of total reviews. Generally speaking, this pattern makes sense as more reviews give the opportunity for both positive and negative experiences to be represented, and can cause the average to fall towards the middle of the rankings. To check if there was an important correlation

between the number of reviews and the company rating however, a t-test was performed to see if there was a significant difference between the number of reviews in high and low rated companies.

The resulting p-value computed from the t-test was above the .05 significance level at .25. These results did not allow us to reject the null hypothesis, and did not allow us to find that the number of reviews had a significant impact on the overall company rating.

I found that the review counts ranged from a minimum value of 3 to a maximum value of 4,041 reviews across all companies in this dataset. The average company review count is 760, and we find the majority of review counts falling below 500. Although we previously looked at the useful, funny, and cool designations as they related to review ratings, it made sense to see if the patterns we found held true for company ratings as well. T-tests were again run for the three variables to determine if there was a significant difference between high (above a 3 star rating) and low (below a 3 star rating) companies. The results of our testing found that there actually isn't a significant difference in average amount of useful or funny ratings between high and low rated companies. For cool reviews however, the results proved that the average cool rating was more likely to be associated with high rated companies.

The last step before examining the text itself was to look closer at correlations between all the numeric variables.



The data was grouped by review rating, and a heat map was generated to easily see how the variables related to one another. A number 1 represents a 100% positive correlation and - 1 would represent a 100% negative correlation. From this heatmap we can see that text length is positively correlated with funny and useful reviews, and is very negatively correlated with company ratings.

We can also see that the company rating is positively correlated with cool reviews, but negatively correlated with useful ones.The review count showing almost no correlation with company ratings, backs up our findings regarding this question posed earlier. One other thing to note is that funny reviews are often found useful, but they have a strong negative correlation with company ratings. Customers seem to appreciate honest reviews about negative experiences so they can avoid making the same mistakes themselves. Positive experiences don't relate to the useful designation in the same way.

After looking through the other variables in the dataset, the text of the reviews themselves were analyzed further. The text was split up into two groups, one star and five star reviews. These were chosen in hopes of look at the most polarizing language for customer experiences.

Each of the two groups were tokenized separately, converted to lowercase, and filtered to only contain alpha characters. I then used the english stopwords from the nltk.corpus package to make sure redundant and unimportant words were removed. Words like hotel, stay, stayed, and 'u' were added to this set of stopwords after an initial run through. I found that they were not helpful in finding common themes, and were frequently appearing in both 1 and 5 star reviews. Once the stopwords were removed, each group was lemmatized and converted to a bag of words.

Looking at the most common words in 5 star reviews I found that words like; nice, pool, good, service, clean, really, love, staff, also, and would were most prominent. In the 1 star reviews words like even, never, back, could , told, service, front, desk, said, and check were appearing most often. There were some common themes in both the 1 star and 5 star review texts. Room was mentioned most frequently in both groups of reviews, and time, like, and service were both found in the top words as well.

Word clouds were generated for each group to better visualize the words associated with both the positive and negative reviews. You'll find the word cloud for the 1 star reviews directly to the left, and the word cloud representing the the 5 star reviews below. Looking at these world clouds we see some additional words of importance stand out for positive and negative reviews. Words like called, nothing, and fee popping up in in the 1 star word cloud are easy to associate with negative hotel experiences. On the positive side, words like restaurant, buffet, food, and

drink appearing in the 5 star word cloud bring our attention to other important factors that go into a positive experience.

In addition to looking at the 1 and 5 star review text, I also wanted to gain some insight on the reviews for top rated companies currently on the Las Vegas Strip. While looking closer at who the top competition will be once The Black Pearl opens its doors, all companies having less than 1000 reviews were taken out of the equation. This left 21 companies, and only 4 of them had a company rating above 3.0. These four were M Resort Spa and Casino, ARIA Resort and Casino, South Point Hotel, Casino & Spa, and Red Rock Casino Resort & Spa. The M Resort Spa and Casino had a company rating of 4.0 and the remaining three had company ratings of 3.5.

Looking at all of the reviews for these companies with over a 3 star rating, I was able to gain some insight into what the positive reviews for these companies consisted of. The common words in the general 5 star review group were found in these reviews as well. Words like great, nice, like, pool, good, service, would, really, also, and clean were most common in the top companies positive reviews. Another few words that appeared more frequently specifically for these companies however, were restaurant, buffet, and food, letting us know what these companies are doing really well.

A word cloud was generated for the top companies positive reviews as well. Looking closer at the world cloud below, bed, spa, and bathroom were also mentioned frequently in these company's reviews.

After analyzing the tokens as individual words I decided to look at the review text as bigrams so that we could gain a little more insight into the topics that were appearing most frequently. As food, restaurant, and spa were words appearing frequently in the top rival reviews, I focused on separate sets of reviews that each contained one of those top words. I also created bigrams for the top rivals negative and positive reviews.

While breaking out the reviews containing the word food into bigrams, I noticed that food court, room service, monte carlo, and mandalay bay appeared very frequently. The reviews containing restaurant had the same top words mentioned in the food reviews, but also contained red rock, and hard rock, other businesses in the area. The top rivals positive reviews mentioned the front desk, customer service, sky suite, movie theatre, bowling alley, and a clean room very frequently, while the negative reviews also had customer service and front desk as appearing most often. It seems that the interactions with the staff, good or bad, often find their way into the reviews.

Summary of Findings:

- 1 and 4 star reviews are most common
- Reviews marked as useful or funny are more frequently associated with low review ratings while reviews marked as cool are more frequently associated with high ratings
- Longer review lengths are associated with low rated reviews, and this feeds into longer review lengths being associated with lower rated companies as well
- The number of reviews alone does not have a significant impact on the overall company rating
- There isn't a significant difference in the average useful or funny reviews between high and low company ratings, but reviews marked as cool are more likely to be associated with high rated companies
- Text length is positively correlated with funny and useful reviews, and is very negatively correlated with company ratings.
- Funny reviews are positively correlated with useful ones
- Common themes in five star reviews: room, time, service, nice, good, love, everything, staff, clean, pool, restaurant, buffet, food, and drink
- Common themes in one star reviews: room, time, even, never, back, could , told, service, front, desk, said, checked, called, nothing, and fee
- Common themes in top companies positive reviews closely resembled the total 5 star reviews, but had restaurant, buffet, food appear more often and they also included bed, spa, and bathroom frequently
- Reviews mentioning food were more often about food courts or room service than any specific restaurants

Predictive Models:

Although we've gathered valuable information on the competition's Yelp reviews, and the factors that go into positive or negative reviews, not all platforms will be accompanied by a star rating. With The Black Pearl's goal in mind of gathering valuable feedback on a variety of online platforms, a supervised machine learning model was created to classify reviews as positive or negative based on the text they contain.

The first step in this process was to prepare the data for our models. Each review was given a label so we can make a judgement on the model's accuracy after the predictions are made. Reviews with a star rating above 3 were given the label 1 and represent positive reviews, while the reviews with a star rating below 3 were given the label 0 and represent negative reviews. All of the reviews that had 3 star ratings were considered neutral and were removed from the dataset used for modeling so the focus can remain on truly positive and negative experiences. The remaining reviews and their corresponding labels were then split into training and testing

groups. This will help avoid overfitting the model by testing on a different dataset than the model was trained with.

In order for our classification model to be able to read the review text it will need a feature vector to work through the classification task. Vectorizing our reviews will break up the text into individual words (tokens), and convert the text data into a matrix of token counts. The rows will be sparse arrays that represent the different reviews, and the columns will represent the words in the vocabulary. This way of representing text however, does not preserve the order of the text and is often referred to as a bag of words representation.

Scikit - learn provides a couple of options to help vectorize the review text. Count vectorizer and tf-idf vectorizer were both used for this project. The count vectorizer uses the term frequency as described above, but the tf-idf puts a weight on the term frequency based on how often that term appears in all of the reviews. The more popular that term is, the less predictive power it's thought to have, and the more it is downweighted. It provides a good measure of how important each term is.

Before transforming our review data with the vectorizers, a pre-processing function was created so that the text could be cleaned up before it was separated out in the vector. The pre-process function brought all the text to lowercase, removed stopwords, and removed all non alphabetic characters. The cleaning steps will make sure that each feature in the vector is easily interpretable.

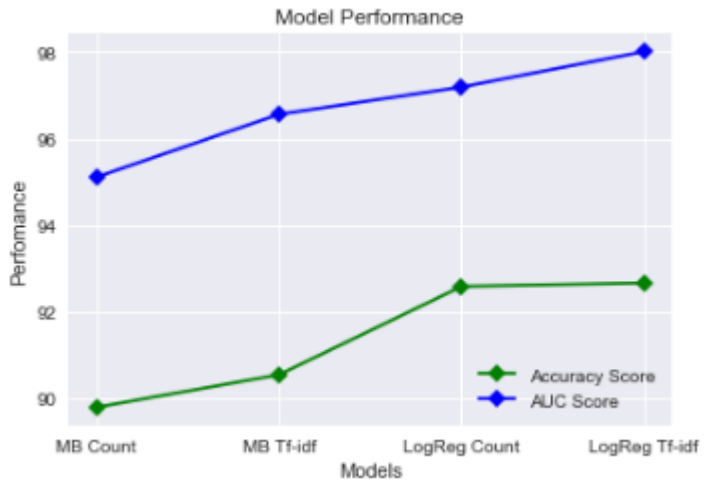
Once the data was cleaned and represented in both the count and tf-idf vectors we were able to fit our data to the classification models. As this is a binary text classification problem we first looked at a Multinomial Naive Bayes model. This model is commonly used in natural language processing classification problems because of its basis in probability, its speed, and its simplicity. In our case it will make a decision on how likely a review is positive or negative based on the words in that review.

The Multinomial Naive Bayes model produced an accuracy score of 89.79% with the count vector and a 90.54% accuracy score with the tf-idf. ROC curves were plotted for both vectors, and the area under the curve(AUC) was computed. Both produced high AUC scores with .9511 for the count vectorizer and .9656 for the tf-idf.

Another model was introduced to help compare performance and see if we could improve our results. The Logistic Regression model was chosen because it is another model that specializes in binary classification, and like the Multinomial Bayes model, it will produce a probability that an item in question should be classified in a specific category. If the probability is greater than .5 it labels the data as 1 (positive review) and if it less than .5 it will label the data as 0 (negative review).

After running the data through the Logistic Regression classifier we found a slight improvement over the Multinomial Naive Bayes models. The accuracy score produced for the count vectorizer was 92.39% and tf-idf scored at 92.67%. For both models the accuracy scores using the count vector and the tf-idf vector were almost identical. Again, ROC curves were drawn for each group, and the logistic regression classifier scored higher in the AUC scores as well. The count vectorizer received a score of .9719 and the tf-idf produced a .9801.

The next step was to see if we could further tune the parameters of our model, and the impact of a different alpha (an additive smoothing parameter) was looked at for the multinomial bayes model, and a different C value (inverse of regularization strength) was looked at for Logistic Regression. There were no parameter adjustments that caused any large leap in performance for any of the models. The three lower scoring models were improved by only hundredths of a percent, and the default parameters produced the best performance for the highest performing model, the logistic regression model using the tf-idf vector. You can find a performance



summary to the left. Once the top model was identified it was further analyzed to see if it's high scores help up in cross validation. Using five fold cross validation the accuracy scores and AUC scores were tested and compared to the results we got from our original test set. The cross validation accuracy scores ranged from 91.8% - 93.5%, and the AUC scores for the area under the curve ranged from .975 - .982. Both of the cross validation score sets represent a tight grouping around the results from our original test set, and we can be confident that the model will produce similar performance with new data.

Next we can look at the feature weights given by our top model in predicting positive and negative reviews. We find that the words toothpaste, obsolescent, tmi, grateful, massages, merciful, and loomed were highly predictive for negative reviews and the words criticism, portrait, chuckles, disembark, barbecued, jewels, chargin, and what is most likely the misspelling of a popular flower, mangolias, were highly predictive for positive reviews. Keep in mind that these words, as part of the tf-idf vector, are weighted higher when they appear less often. You will notice a difference in highly weighted words here compared to the most frequent words associated with positive and negative reviews mentioned previously.

Recommendations and continued improvements:

1. Logistic Regression Model:

As the Logistic Regression classifier with the tf-idf vectorizer outperformed the other models, and its performance metrics held up during our cross validation, we recommend this model for continued analysis of competitor Yelp reviews. This model can be fit with the data on The Black Pearl's future Yelp reviews as well after opening. This model will provide a solid foundation for text analysis on other platforms as well, but adjustments will have to be made based on the intricacies of the specific platform. It's not recommended to use this model to jump between online platform without taking the time to tune the model appropriately.

2. Focus on the following competitors:

- M Resort Spa Casino
- ARIA Resort & Casino
- South Point Hotel
- Red Rock Casino Resort & Spa

Through our analysis we've found that many businesses above a 3.5 rating did not have a sufficient amount of reviews to fairly capture both positive and negative experiences. When looking for The Black Pearl's top competition only companies with 1,000+ reviews were considered. The aforementioned businesses were the only ones in this category above a 3.0 company rating. It will be important going forward to monitor sentiment regarding these businesses as a benchmark for success, and look to gain competitive advantages based on their customers experiences.

3. Food:

Food wasn't only one of the most frequent topics mentioned in 5 star reviews, but it appeared even more frequently when we looked at the top rival reviews. The importance of great restaurants can't be overlooked to provide a positive customer experience, but even more frequently the reviews mentioned the food court, room service, and buffets. Paying special attention to promoting quality in these areas can be a major factor for customers sharing positive experiences about your business.

4. Customer Service:

Customer Service was one the most frequent topics in both positive and negative reviews, and front desk appeared very frequently in the top rival positive reviews as well. As a lot of the negative reviews contained words like called, told, and resort fee it's crucial to make sure the staff behind the front desk are properly trained to handle a multitude of different types of customer requests quickly, and that they are comfortable explaining the reasoning behind any pricing the customer does not expect. Customer service has been frequently mentioned in the competitor reviews whether it was good or bad, so this can be an area to shine in as the experience with the staff at the front desk will make it into the majority of reviews.

Continued Improvements:

As analyzing customer feedback and making improvements based on customer needs is a continual process, there will always be ways to further refine the process. One area of improvement in the future can be increasing the amount of n-grams studied. This analysis looked at individual words as tokenz and bigrams as tokens to help pull more insight from the reviews. Although it will take more computational power, analyzing 3 or 4 gram tokens will help provide more context in the topics that appear most frequently, and using a higher n gram range while vectorizing the review data can help the accuracy of the sentiment model as well.

Another idea to look into is tagging the reviews using named entity recognition in the future. This can allow you to extract information about the most talked about people or organizations in these reviews, it may help your staff provide suggestions for places to go or things to do based on what the majority of customers are writing about at that time.