Capstone Project Finale Report

Yelp Sentiment Analysis
Serik Omarov

1. Background

In a world where data is becoming increasingly more important, people have a prerequisite demand to have more information before spending their money on a product or service. This is especially evident in review-based applications like Yelp. Yelp, founded in 2004, is an online social network service in which consumers provide reviews on their local restaurants by describing their individual experience with a review and rating on a 1 to 5 star-rating scale. Consumers can see what a new restaurant will entail in terms of type of food, amenities, etc. before commuting. Restaurant owners benefit by receiving valuable, direct customer feedback on service, quality, location, amenities, etc. Reviews may contain key factors consumers are looking for in their restaurants that the restaurants may be currently missing or are performing poorly on. Analyzing these reviews are an integral part of continuous improvement because they could result in future, higher reviews and positive feedback. Higher reviews will also enable the restaurant to become more recommended and popular in search engine entries.

Sentiment analysis is a vital technique in which I will be classifying the review text as either positive or negative by incorporating aspects of Natural Language Processing (NLP) and machine learning. Specifically, I will conduct logistic regression and determine the sentiment based upon the evaluation of a restaurant review. From the Yelp review data, consumers discuss their experiences and opinions through their reviews, thus natural language processing and sentiment analysis can be utilized.

In this project, i conduct data cleaning and processing, exploratory descriptive analytics, modeling, results, and conclusions. Overall, sentiment analysis can help understand consumers, detecting business' day-to-day successes and faults, and remedying them over time.

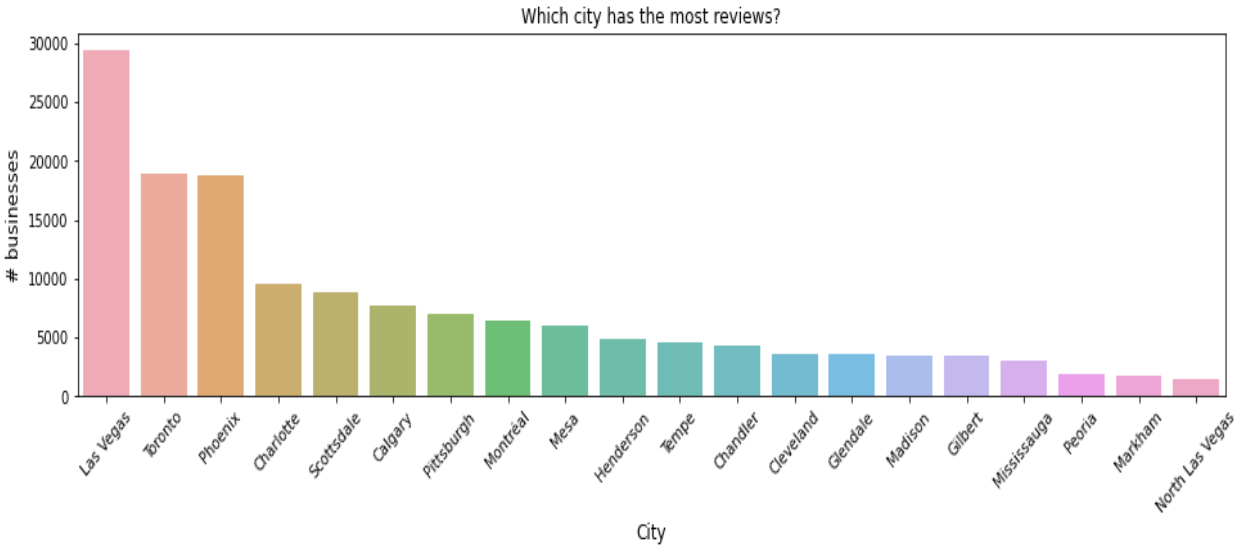
2. Data

The dataset that I used for our research was the Yelp Dataset that included JSON files of business, checkin, reviews, tip, and users. In analysis I limited dataset to that of the business and review JSON files. The business file includes business_id, stars, attributes, and categories, with around 8 million records. The review file includes the business_id and reviews at about 200,000 records. The two datasets will be combined to produce one dataset to work from. My focus on Yelp is to understand the relationship between text reviews and ratings provided by the consumers' experiences at their local restaurants. I incorporate sentiment analysis to this research to create sentiments based on text reviews provided by the consumer.

3. Data Cleaning

One of the bigger concerns with the dataset was the size and volume of the data. With the two utilized datasets, the task of joining both the business data and the review data proved to be very expensive. As each of the 209,393 unique businesses in the dataset had hundreds to thousands of reviews, it necessitated the requirement to narrow down our analysis to study a certain niche of Yelp restaurants. This would make both topic modeling and the application of the business problem more viable. Ultimately the study was conducted on Japanese

restaurants' reviews in the city of Las Vegas, Nevada. In this city was the most heavily concentrated number of Yelp user reviews for establishments. Regarding the business dataset, all geographic and location-based features were dropped such as address and coordinates. Next, filtering was utilized to reduce the dataset down to Japanese restaurants in Las Vegas. After filtering, the dataset comprised 437 individual restaurants. Further dataset cleaning was done such as querying for NA values. 80 missing values were found in the column "Attributes", and these records were removed. Finally, once the business dataset was cleaned and filtered, this made joining with the review dataset less demanding. The business dataset and the review dataset were joined on the column business_id. Finally, the joined dataset resulted in 148,224 records. This represented all the Japanese-style restaurants in Las Vegas and all their respective Yelp reviews.



4. EDA

Distribution of star ratings was viewed by a bar plot. Through the below figure, I can see comparisons among discrete categories of star ratings. Each category in the bar plot is a respective star rating. It is clear that the majority of reviews for Japanese-style restaurants in Las Vegas range from 3 to 4.5 star ratings. Based on this, I can assume that most of the reviews will be relatively positive. This may be due to performance or because consumers who seek out Japanese-style food in a city like Las Vegas may be more familiar with the style of food and know that they already enjoy it.

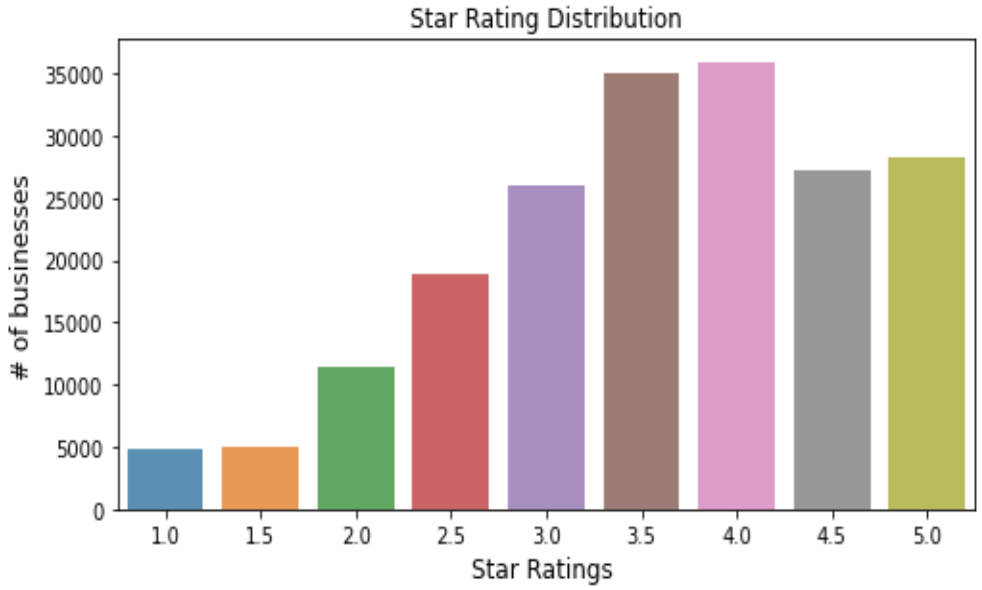


Figure2 Review Stars Barplot

Continuing on the previous assumption, in the Appendix Figure 1 the top 10 restaurants based on review count. The majority of this view are restaurants that cater towards the most well-known types of Japanese-style foods like BBQ, sushi, and Ramen. Distribution of the count of reviews for each restaurant was viewed by a histogram. Through the below figure show that a majority of Japanese-style restaurants have less than 500 reviews. Lower review counts may be correlated to the generally lower, overall popularity of traditional Japanese food in a city like Las Vegas, where Japanese-style food is tended to be fused with familiar styles of food in the area.

Table 1. EDA

	stars	review_count	is_open
count	381.000000	381.000000	381.000000
mean	3.837270	300.503937	0.566929
std	0.602431	403.386569	0.496152
min	1.000000	3.000000	0.000000
25%	3.500000	49.000000	0.000000
50%	4.000000	153.000000	1.000000
75%	4.000000	396.000000	1.000000
max	5.000000	3449.000000	1.000000

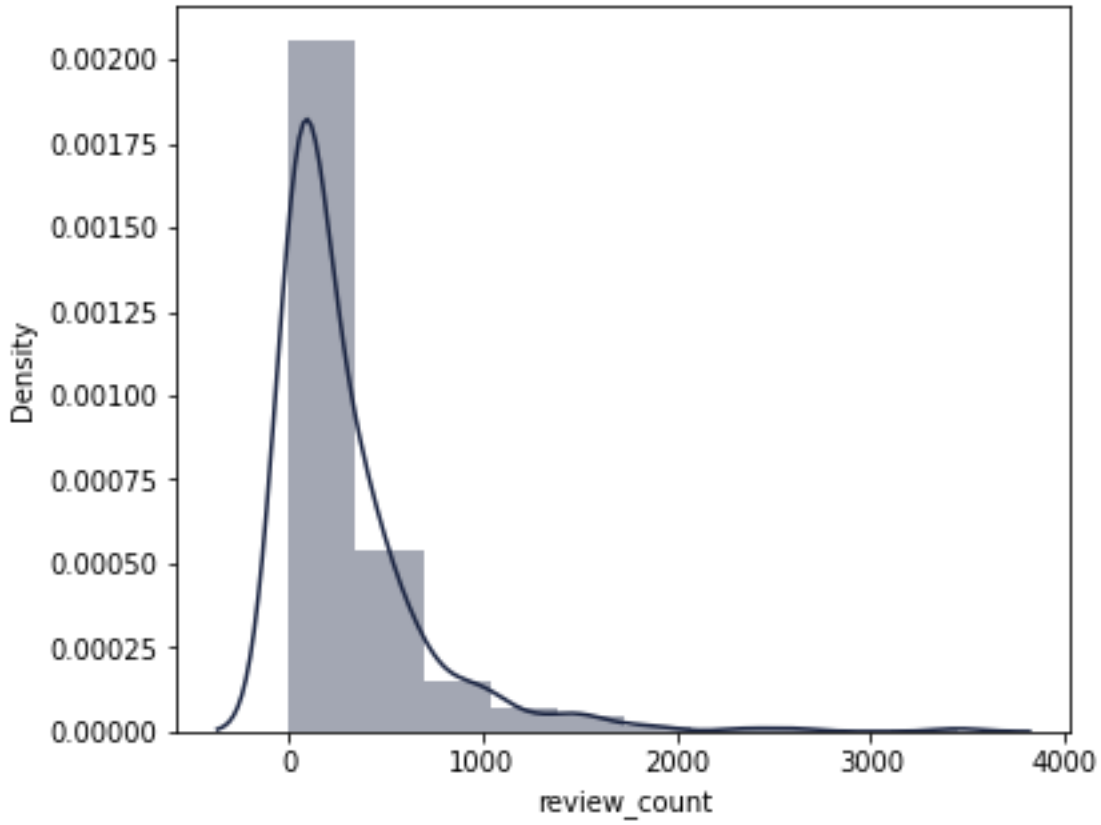


Figure 3 Review Count Histogram

World clouds were produced in descriptive analytics to quickly visualize the most frequent words, in respect to top positive words and top negative words. These allow Japanese-style restaurant owners to swiftly view concepts or items consumers use most in their reviews

These can provide nimble insights on what people talk about most in their positive or negative reviews. Quick insights can be made due to the nature of the Wordclouds. Word clouds visualize the most frequent words in a set of text easily by manipulating size. The larger the size of the text words in the word cloud, the higher the frequency of the words in the set of text. Also, stop words were filtered out of the reviews in the NLP cleaning step, so unmeaningful words were filtered out. In the below “Top Positive Words” word cloud, Japanese-style restaurant owner could quickly see the types of items or concepts consumers talked most positively about from Japanese-style restaurants. It should be made clear that these items or concepts are notin relation to a specific restaurant but to all of them included in the filtered dataset. For example, in the “Top Positive Words” word cloud, “ice cream”, “sushi place”, and “happy hour” are some of the most frequent words in positive reviews. These are just the concepts that consumers talked most about in positive reviews overall. A Japanese-style restaurant owner could integrate having quality ice cream on the menu or having competitive happy hours to possibly increase his or her number of positive reviews.

Figure 4. Positive Word Cloud

In the below “Top Negative Words” word cloud, Japanese-style restaurant owner could swiftly see the most frequent words consumers talk most negatively about. Many of the most frequent, negative words are food types like “miso soup”, “sushi roll”, and “fried rice” as well as words like “customer service”. Japanese-style restaurant owners can be informed that food quality and customer service are some of the attributes that consumers talk most negatively about in reviews. Through this, restaurant owners can make the appropriate adaptations towards these aspects.

Figure 5. Negative Word Cloud

5. Modeling

The predictive task for our study aims at predicting the sentiments of a user's experience at a Yelp restaurant solely given the textual data of their respective review. Our goal is to build a model that classifies the sentiment of the review being positive or negative, without the supplemental categorization of a Yelp star or review
rating. Additionally, I am interested in identifying the textual features that serve as the most significant for this method of classification. For our model, I utilize the discriminative model Logistic Regression. Logistic Regression encompasses the use of the sigmoid function, outputting a probability between 0 and 1.Furthermore, a binomial logistic regression predicts the probability that a specific observation is classified in one of the two categories, positive or negative.

5.1 Splitting Data

The train-test split ratio was set at 80/20, respectively. The training data consisted of 118,579 reviews while the testing set consisted of 29,645.

5.2 Feature Selection

An important parameter for constructing our model is the number of features selected for the vectorized tokens in the corpus. We evaluated the accuracies based on the varying numbers of features allowed using another pipeline where the Logistic Regression model was fit to the training data and ran with 500 iterations using trigrams from the Count Vectorizer. The evaluation of the selecting of n-grams will be discussed later in the study.

Table 2. Feature Selection Accuracy

# of Features Accuracy	Accuracy
5000	0.9199
7500	0.9218
10000	0.9222
12500	0.9231
15000	0.923

With the results above, the higher number of features starting from 5,000 results in higher accuracy. However, once the number of features exceeded 12,500, the accuracy dropped off in value. I decided to choose12,500 features for further implementation.

5.3N-Grams

When utilizing predictive modeling to predict sentiment labels of reviews, there can be potential drawbacks in using only single words as features or unigrams. Being able to understand certain positions such as "not good" and "not bad" will not be accounted for, possibly leading to poor classification. To take this into proper consideration, n-gram usage as features can lead to increased accuracy of the sentiment prediction model.

Table 3. N-Gram Selection

# of N-grams Accuracy	Accuracy
1	91.49%
2	89.26%
3	82.00%

The results provided above were produced from a pipeline similar to the one used for evaluating features election. However, for this pipeline, the parameter for n gram range in the TFIDF vectorizer was run for ngrams1, 2, and 3 or single feature, bigrams, and trigrams. From the results show that single features provide the most accurate model, with the use of trigrams having the least. Attaching a word such as a negation to another word that precedes or follows it can be a beneficial procedure that enables the improvement of a classification problem's accuracy since negation can factor into opinion or sentiment expression. Many studies evaluate the performance of sentiment classification models based on n-grams. In a similar case that classifies polarity on Tweet sentiments, it was found that the best performance was achieved when using bigrams. For this project , though unigrams had the higher accuracy, choosing bigrams for the model will prove much more effective. It will provide more conclusiveness in identifying phrases of words that are associated with both the positive and negative sentiment class.

6. Model Results

I conduced feature importance for the purpose of gaining insights into the data and as well as the model. The implementation of feature importance assigns scores to input features based on how impactful they are predicting a target variable. This is done by splitting the coefficients from the classifier into positive and negative features and plotting them to see their impacts on positive or negative sentiments. From the top positive and negative words, the most important features are labeled with scores on the y-axis.

From Appendix Figure 2, two of the most important features are "definite back" and "highly recommend". These inform Japanese-style restaurant owners that when positive experiences are had, consumers are more likely to return or spread their positive experiences with their communities which can be a form of free advertising by word-of-mouth. From Appendix Figure 3, the opposite can be informed. Words like "won't back" and "never come "inform Japanese-style restaurant owners that consumers with negative experiences will not return and that they may spread those negative experiences with their communities which results in negative advertising byword-of-mouth. It can also be seen that "food poisoning" and "service horrible" are aspects that are highly correlated to negative sentiments. Japanese-style restaurant owners can adjust their business to steer away from these aspects.

7. Model Evaluation

Using the test data, the performance results from the predicted and actual values are visualized in the confusion matrix in Figure 6.

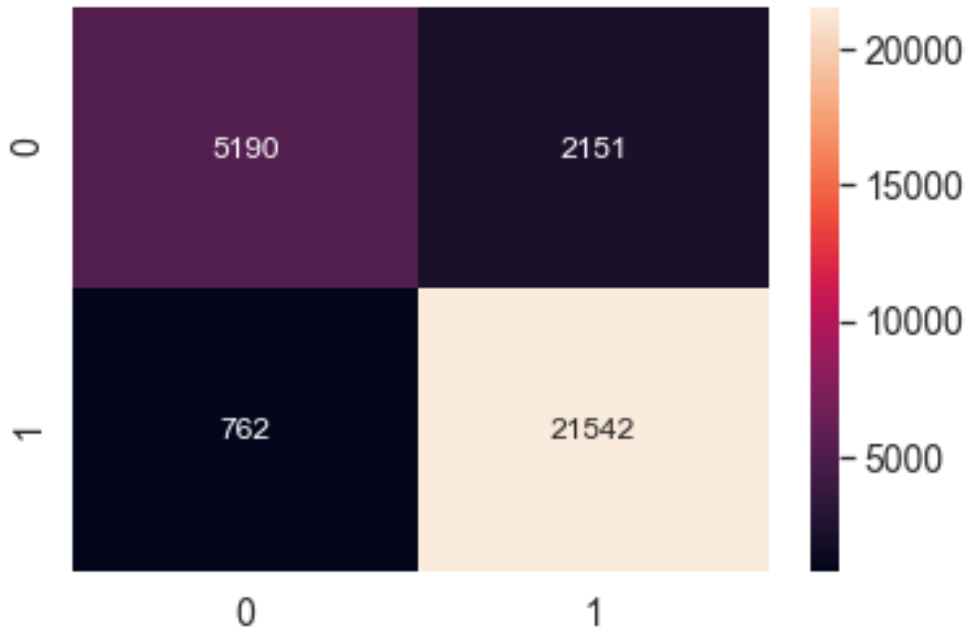


Figure 6. Logistic Regression Confusion Matrix

A confusion matrix is helpful for retrieving performance measures such as accuracy, recall, precision, and F1-score. Accuracy answers the question: “Out of all the classes, how much did I predict correctly?”. Recall measures how much the model was able to predict correctly out of all the positive classes; ahigh value is desirable. Precision measures how much actual positive predictions were made from all the correct positive class predictions. The F1-measureallows the ability to measure both precision and recall simultaneously. These measures along with the classification report are provided in Table 4 and Table 5 respectively.

Table 4. Model Performance Measures

Measures	Score
Accuracy	90%
Recall	97%
F-1	94%
Precision	91%

Table 5. Model Classification Report

	Precision	Recall	F1-Score	Support
0	88%	69%	78%	6194
1	90%	97%	93%	17567
accuracy			89%	23761
macro avg	89%	83%	86%	23761
weighted avg	89%	89%	89%	23761

Another important performance indicator is known as the AUC-ROC curves. ROC is defined as a probability curve that plots the True Positive Rate (TPR) on the x axis
against the False Positive Rate (FPR) on the y-axis. The AUC conveys the model’s capability of classifying the distinction between classes. A higher AUC that’s close to 1 represents a more effective model in predicting the sentiment between negative and positive.

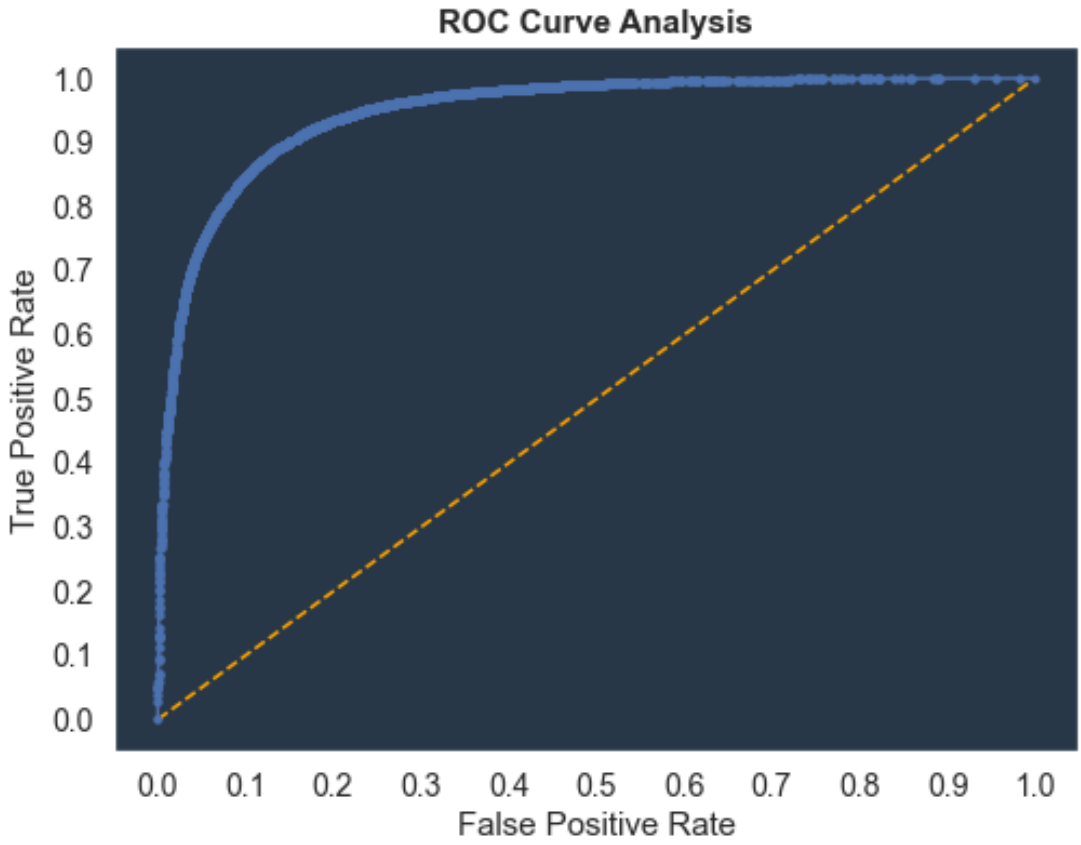


Figure 7. ROC-AUC

7.1 Performance Evaluation

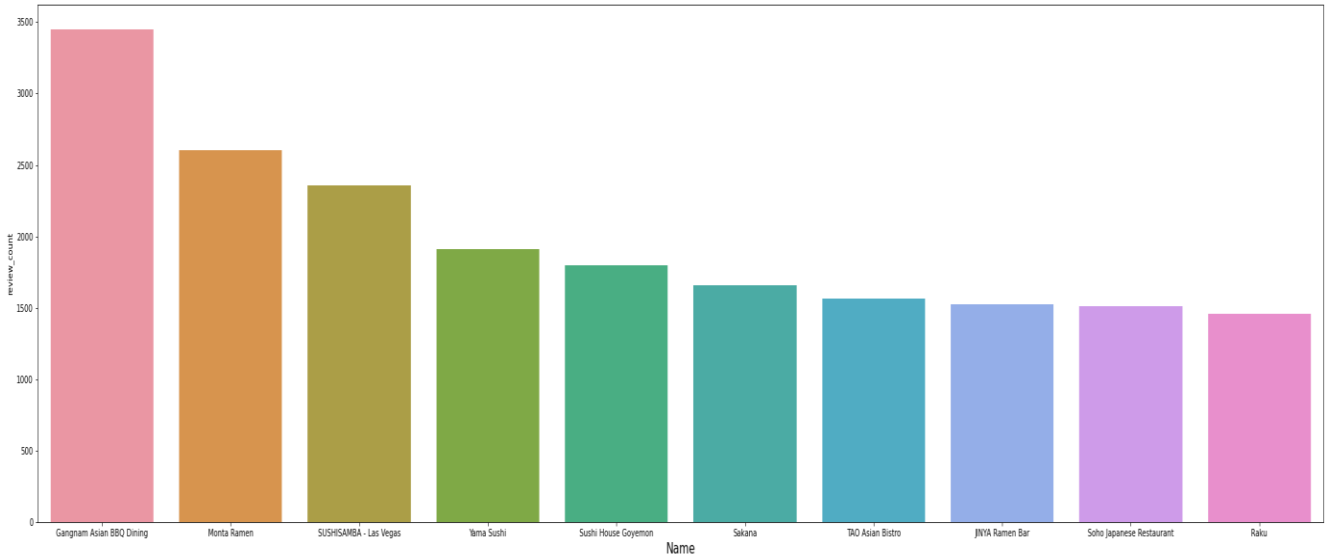
The model produced very good results with the overall accuracy of the model being 90%. The recall score is also very high at 96%, denoting that our model is able correctly identify positive classes. Finally, the F-1 score at 93% conveys that the model is able to correctly predict positive classes very well. Looking at the classification report, we see that there are 22,304occurrences of the positive class and only 7,341 of the negative class, even despite assigning the neutral star rating reviews to negative sentiment. I expected this imbalance of distribution of classes from exploratory analytics, as most of the reviews for these Las Vegas restaurants were positive. Ultimately, this is a concerns unbalanced support in the data may prove a weakness and could possibly necessitate resampling or further feature extraction. By acknowledging our AUC score of 0.949 and its closeness to 1, I interpret this as a very good measure of separability, meaning there is a 95% chance the model will be able to distinguish between the two classes.

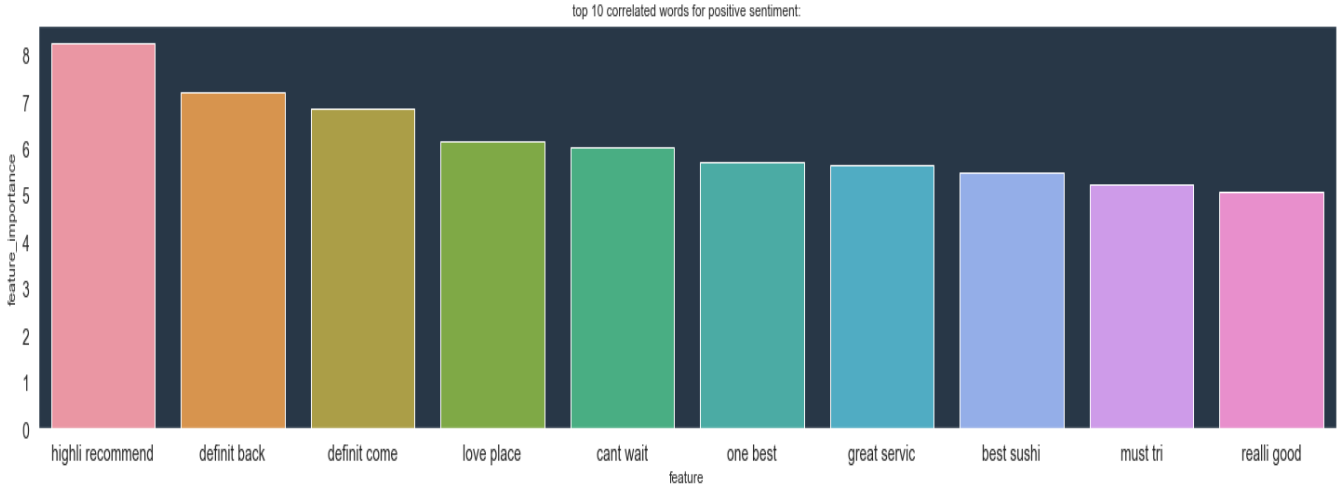
8. Conclusion

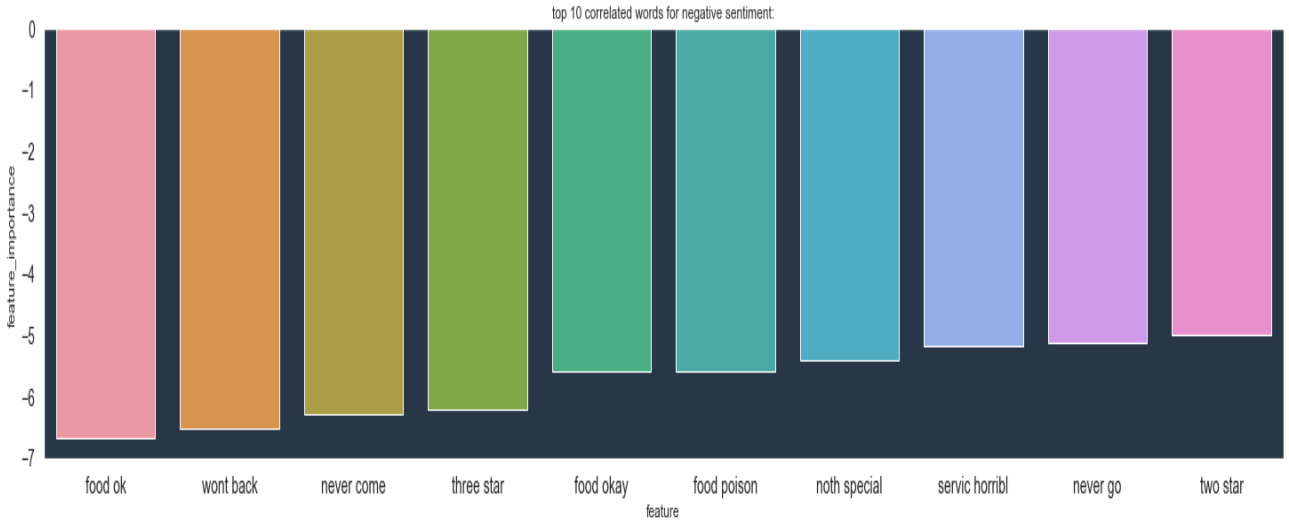
Throughout this analysis I found that sentiment words were more positive than negative of Yelp users ‘experiences at the designated restaurants. In relation to the positive sentiments of users, there is a positive correlation between that of positive reviews with high ratings, and negative reviews with lower ratings. By segmenting the area to Las Vegas and categorizing Japanese restaurants, i was able to gain insight on how they operate.

Each of the individual users provide their opinions throughout their reviews; as the positive outweighs the negative, Japanese restaurants are providing great dining services and food to their customers, which increases the positivity disclosed in their review, as well as an input of a higher rating. Although, through this project, Japanese restaurant owners can also view aspects that drive more negative sentiments which they can take initiative on and remedy over time.

Appendix


review_count
3500
3000
2500
2000
1500
1000
500
0
Gangnam Asian BBQ Dining
Monta Ramen
SUSHISAMBA - Las Vegas
Yama Sushi
Sushi House Goyemon
Sakana
TAO Asian Bistro
JINYA Ramen Bar
Soho Japanese Restaurant
Raku
Name


top 10 correlated words for positive sentiment:
feature_importance
8
7
6
5
4
3
2
1
0
highli recommend
definit back
definit come
love place
cant wait
one best
great servic
best sushi
must tri
realli good
feature


top 10 correlated words for negative sentiment:
feature_importance
0
−1
−2
−3
−4
−5
−6
−7
food ok
wont back
never come
three star
food okay
food poison
noth special
servic horribl
never go
two star
feature