University of California, Santa Barbara

234 Statistical Data Science

Yelp Review Analysis of Top 10 Restaurants in Santa Barbara

Author:	*Perm Number:*
Sunpeng Duan	7594880
Mengye Liu	5058862
Zhe Li	7595127

June 11, 2021

Contents

1 Introduction **2**

2 Exploratory Data Analysis **2**

3 Review Analysis for Brophy Bros **5**
3.1 Text Analysis 6
3.1.1 Word Cloud 6
3.1.2 Bag-of-Words and N-Grams Model 6
3.1.3 Word2Vec Model 7
3.2 Sentiment Analysis 8
3.2.1 VADER Analyzer 8
3.2.2 Support Vector Machine 8

4 Review Analysis for Top 10 Restaurants **10**
4.1 Text Analysis 10
4.2 Sentiment Analysis 12

5 Conclusion **13**

1 Introduction

Yelp is a crowd-sourced local business review and social networking software. It is currently the most widely used restaurant and merchant information app across United States. It can provide an overall summary for a specific restaurant, such as, the average rating, restaurants category and other business information. Moreover, it also records the customers' rating and review for a particular restaurant. However, the original customers' reviews are unstructured data. The objective of this project is to make use of the unstructured data and extract sentiment features, which can in turn help improve the customers' understanding of restaurants, new business owners' market knowledge, and existing merchants' awareness about restaurants' features.

In the first part of this project, we scraped and crawled the overall information of all the registered restaurants in Santa Barbara county on Yelp, and then conducted exploratory data analysis for the overall information.

In the second part, we collected all the customer reviews for the restaurants which have the top 10 largest numbers of reviews. We conducted the sentiment analysis of Yelp's customer review data with natural language processing tools as well as machine learning techniques for a seafood restaurant, Brophy Bros - Santa Barbara, and the top 10 restaurants.

2 Exploratory Data Analysis

There are total 691 restaurants registered on Yelp in Santa Barbara.

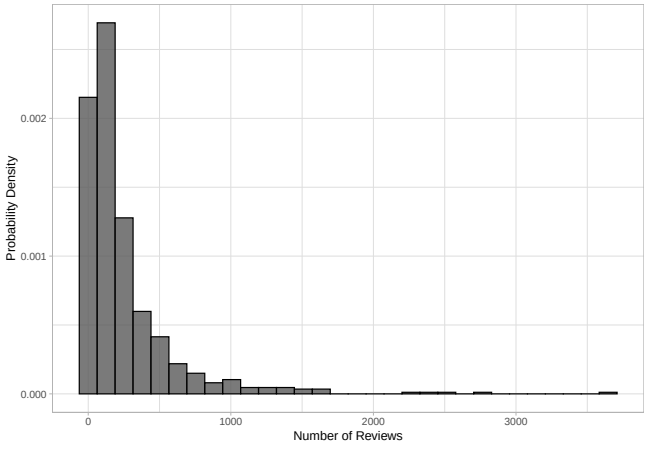

Figure 1: Histogram of Number of Reviews

According to Figure 1, the distribution of number of reviews is severely right-skewed.

And only five restaurants have over 2000 reviews. These restaurants are Los Agaves, Brophy Bros - Santa Barbara, Paula's Pancake House, Boathouse at Hendry's Beach, Santa Barbara Shellfish Company. And among these five restaurants, three of them belong to seafood restaurants.

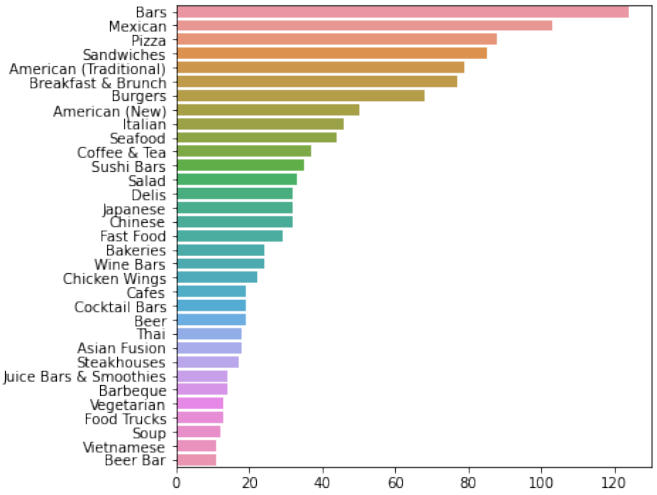

Figure 2: Most popular categories

There are 129 restaurant's categories but only 33 categories have more than 10 restaurants. The top 33 categories are shown in Figure 2. The first three most popular categories are Bar, Mexican and Pizza. Then we investigate the behavior of top 9 categories. We find that the average rating of each category is the highest in the region having the most restaurants and region with zip code 93101(Santa Barbara) has the most restaurants and highest ratings which is consistent with Figure 5.(Detail table is in Appendix: Table 1)

By looking into Figure 4, over 85% of restaurants are rated over 3. This indicates customers are satisfied with majority of restaurants in Santa Barbara.

We also investigated the relationship between ratings and number of reviews. Based on Figure 3, popular restaurants attract more customers, such as Los Agaves, which has been rated 3647 times. That is, the number of ratings for each restaurants is positively associated with its average rating score.

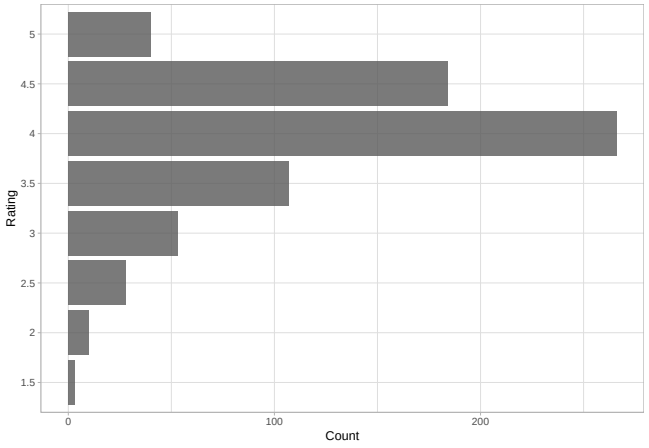

Figure 3: Bar Plot of Ratings

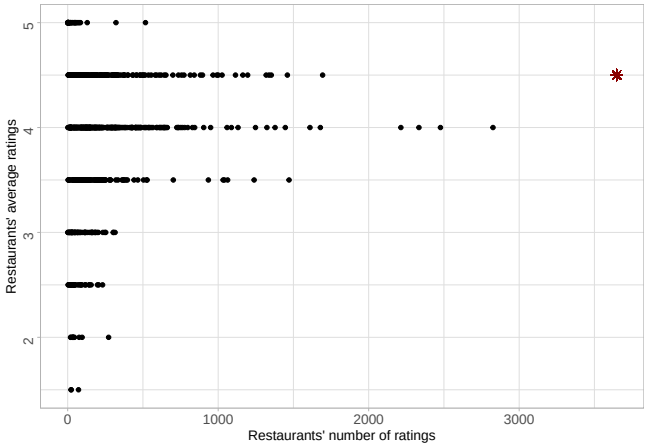

Figure 4: Ratings versus Number of Reviews

Map visualization can also provide us some information about the restaurants in Santa Barbara. Figure 5 demonstrates that most restaurants are located in Goleta and Santa Barbara Downtown. Moreover, the most popular restaurant are also in this area.

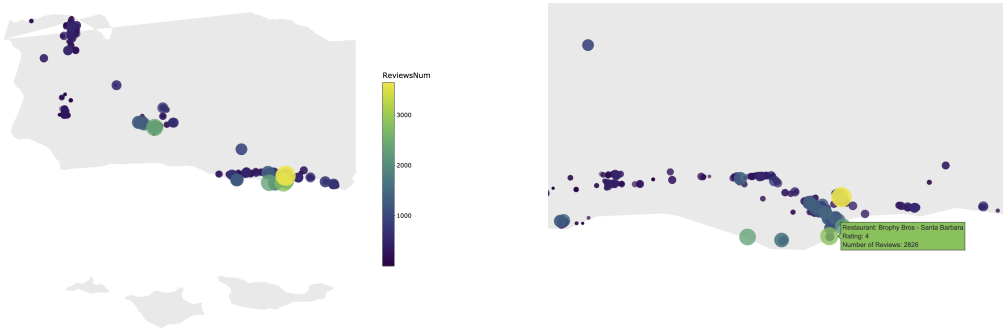

Figure 5: Maps for Restaurants in Santa Barbara

3 Review Analysis for Brophy Bros

Most of the restaurants in Santa Barbara belong to Mexican or seafood restaurants. Thus, we selected the top 1 seafood restaurant, Brophy Bros - Santa Barbara, and analyzed the reasons of its popularity. The average rating of Brophy Bros is 4 stars based on 2834 customers' reviews.

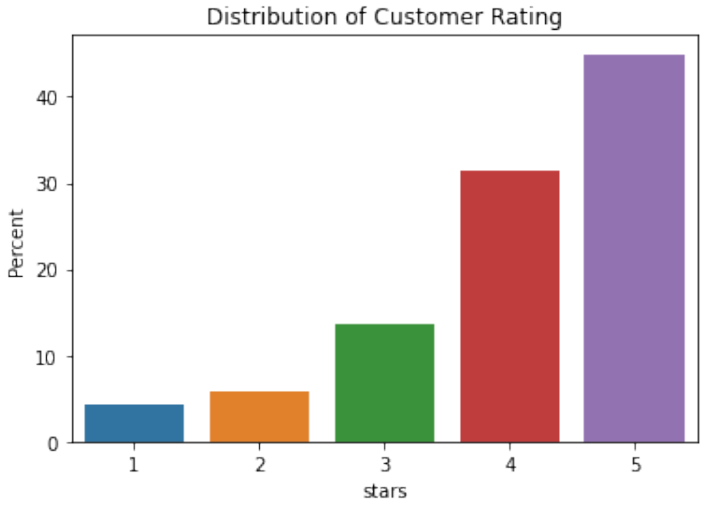

Figure 6: Distribution of Ratings for Brophy Bros

According to Figure 6, over 70% of the customers rate Brophy Bros with 4 or 5 stars. To analyze the reason of its popularity, we first examined all the reviews from its customers with word clouds, Bag-of-Words (BoW) and Word2Vec model. Then, we also conducted the sentiment analysis with all the reviews and ratings via VADER Analyzer and Support Vector Machine (SVM).

3.1 Text Analysis

In the text mining of all the reviews, some words, such as a, an, the, are not informative. We call these words stop words. We removed all the stop words in the reviews with two set of stop words from `nltk.corpus` and `sklearn.feature_extraction.stop_words` Python library.

3.1.1 Word Cloud

After removing the stop words, we visualized all the words in the customers' reviews with the following word cloud (Figure 7).

Figure 7: Word Cloud for Brophy Bros

As shown in the Figure 7, the word "good" indicates majority of the reviews are positive. These positive reviews may be due to its service, waiting time, and etc. Again, we found the word "location" occurs frequently. Moreover, Figure 7 also give us some information of Brophy Bros' menu. It seems that clam chowder is the most popular among all the items on the menu. Besides, customers of Brophy Bros also like to order some items with some kinds of fish or oyster.

3.1.2 Bag-of-Words and N-Grams Model

In addition to the word cloud, the frequency of each word among all the reviews is of interest. Bag-of-Words (BoW) model is a good choice of text representation in numbers. In the BoW model, a review by one customer is basically represented as the multiset of its words, disregarding grammar and even word order but keeping multiplicity. Then, the term frequency of each word could be measured. However,

it would be a little bit complicate due to variations of a word. For example, the base form "go" may appear as went, gone, going, and goes in reviews. Hence, to remove this effect, we applied `WordNetLemmatizer` to all the bags of the words after tokenizing all reviews and removing the stop words.

A typical BoW model only considers the word itself without taking the orders of words into account. However, the order of words matters in some cases. Hence, we also measured bigrmams frequency. 20 most frequent words and bigrams are summarized in Figure 8.

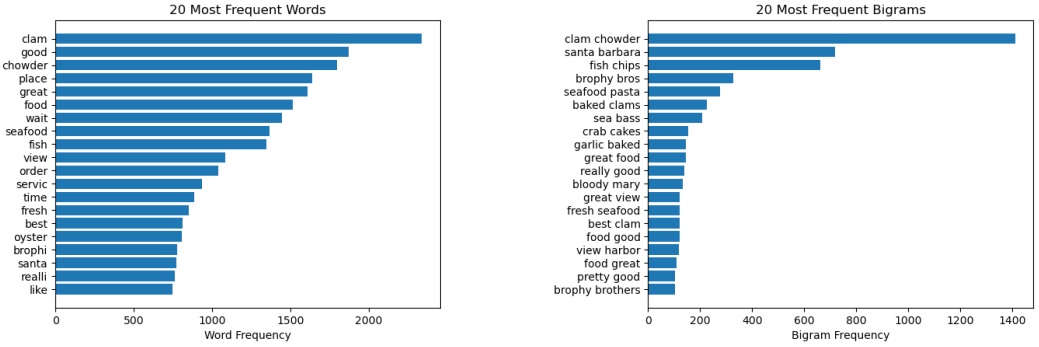

Figure 8: 20 Most Frequent Unigrams and Bigrams

The 20 most frequent words and bigrams share some similarities. Overall, most of the 20 most frequent words and bigrams are related to food. We may infer that Brophy Bros are famous for its food. And when when we look into details, "calm" and "chowder" are among the top 3 most frequent words, while "calm chowder" are the most frequent bigrams. This together with the word cloud indicate that calm chowder is most popular item among the menu. Besides, we can see more from the bigrams frequency plot. Besides clam chowder, customers also prefer fish chips and seafood pasta. Moreover, most customers also mention service. It seems that the service of Brophy Bros leads to some parts of positive reviews.

3.1.3 Word2Vec Model

To further investigate reasons for Brophy Bros' popularity, we also adopted the Word2Vec model. The Word2Vec algorithm uses an one layer neural network model to learn word associations from a large corpus of the reviews. It can help to detect synonymous words or suggest additional words for a partial sentence. From the results of BoW model, we know that the service of Brophy Bros may leads to some parts of positive reviews. To verify this, we applied the Word2Vec model with the skip-gram architecture to all the positive reviews. We want to see which words are most likely around the word "service". From the output, the word "friendly"

and "fast" occurred around "service " with the predicted probability 0.0031 and 0.0017, respectively. We may conclude that these positive reviews containing the word "service" may due to its friendly or fast service.

3.2 Sentiment Analysis

Sentiment analysis is the process of computationally determining whether a piece of text is positive, negative or neutral. In our case, the text is the review written by the consumer. With the aid of sentiment analysis, we can help us study the opinions of customers toward a restaurant. The opinions could be positive, negative or neutral. To conduct sentiment analysis, we applied VADER analyzer and support vector machine.

3.2.1 VADER Analyzer

VADER analyzer is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. With a predefined list of words with sentiment scores, VADER analyzer matches words from the lexicon with words from the review. In the `NLTK` package, given a piece of the text, the VADER analyzer returns scores with four kinds - negative, neutral, positive and compound. And the compound score ranging from -1 to 1 is a combination of positive and negative scores. If the compound score is less than 0, then this piece of text expresses negative opinion, and vice versa.

The average monthly sentiment scores by the VADER analyzer are demonstrated in Figure 9. At the beginning years of Brophy Bros fluctuated from -0.5 to 1 with large variations. After 2008, its sentiment scores are positive, indicating majority of the customers are satisfied with Brophy Bros. Moreover, after 2010, the average monthly sentiment scores tend to be stationary.

3.2.2 Support Vector Machine

Besides VADER analyzer, we also used support vector machine for sentiment analysis. However, for SVM, we should have the labels of positive or negative for model training. Rating with 4 or 5 stars is labelled as positive reviews, while rating with 1 or 2 stars is labelled as negative reviews.

Instead of using BoW model to generate the design matrix, we adopted TF-IDF model. The TF-IDF frequency is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word at the

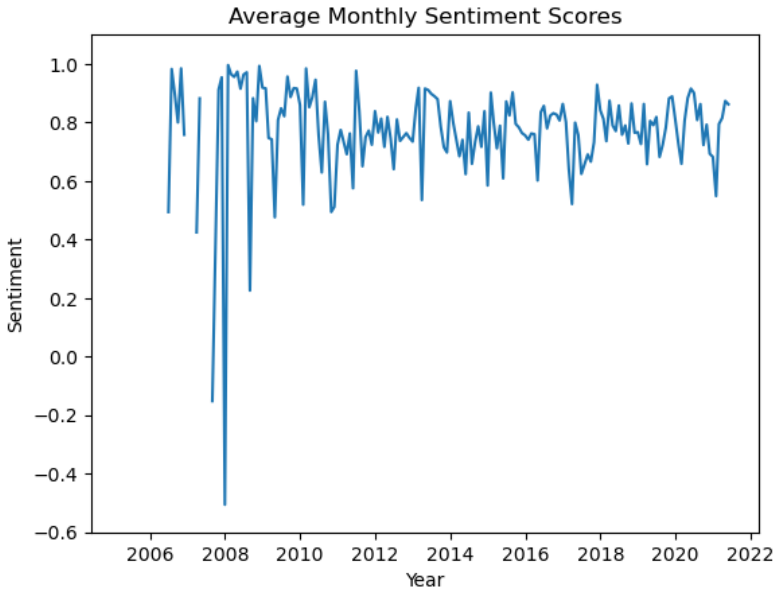

Figure 9: Average Monthly Sentiment Scores of Brophy Bros

same time. It is defined as follows,

$$td\text{-}idf_{t,d} = tf_{t,d} \cdot \log \frac{N}{df_t},$$

where $tf_{t,d}$ is the frequency of term t in document d, df_t is the number of documents containing term t, and N is the total number of documents.

We built the SVM models with both both tf-idf frequency matrix of unigram and bigrams. And the results are summarized in the following Figure 10. The bar chart visualizes the coefficients of important features with the top 10 largest absolute size of coefficients. Blue bars correspond to important unigram/bigram which may indicating positive reviews, while red bars correspond to important unigram/bigram which may leading to negative reviews.

Based on Figure 10, the model with unigram tf-idf matrix does not provide us more valuable information, since most words are obviously subjective. On the contrary, the model with bigram tf-idf matrix indicates the bigrams such as clam chowder, fresh seafood and baked clams are associated with the positive reviews.

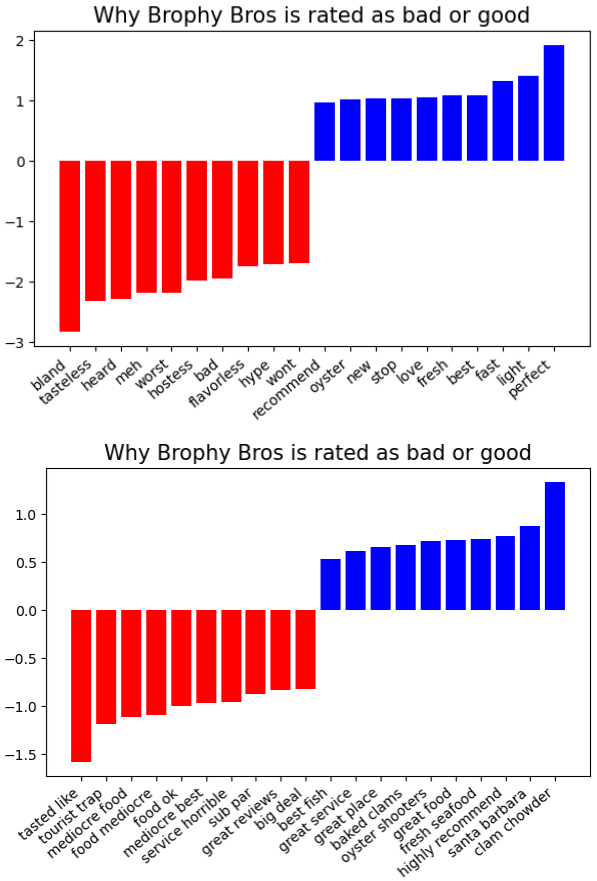

Figure 10: Reasons for the Popularity of Brophy Bros

4 Review Analysis for Top 10 Restaurants

We also analyzed reviews of the top 10 restaurants in Santa Barbara. Figure 11 shows the distribution of customer ratings for top 10 popular restaurants in Santa Barbara. As shown, over 70% reviews are rated with 4 or 5 stars. This is similar to the distribution of ratings for Brophy Bros - Santa Barbara.

4.1 Text Analysis

As shown in the word cloud (Figure 12), the word "food" and "place" are most frequent among the reviews of top 10 restaurants. This may indicate that food and place are common factors affecting the customers' reviews. The word "great" and "good" also indicate majority of the reviews are positive.

We also applied the BoW model to record the frequency of each unigram and bigram

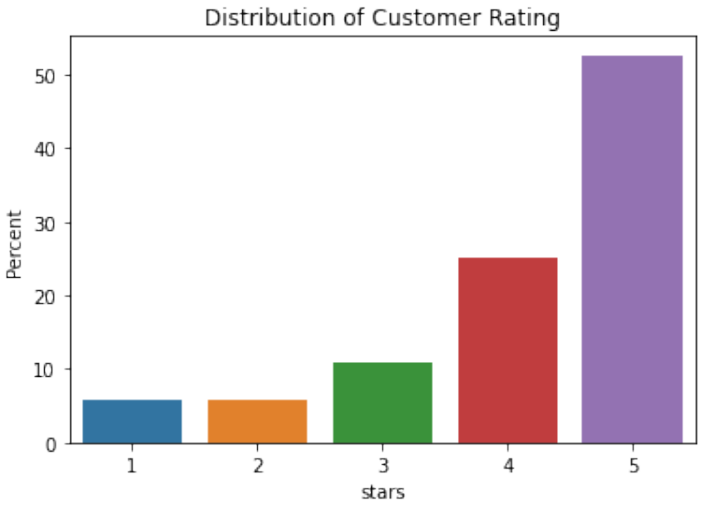

Figure 11: Distribution of Ratings for Top 10 Restaurants in Santa Barbara

Figure 12: Word Cloud for Top 10 Restaurants in Santa Barbara

of the reviews. The unigram plot shows the similar results of word cloud.

According to the first 5 most frequent bigrams, we can find that the bigrams "clam chowder", "Mexican food", and "danish pancakes" correspond to Brophy Bros, Los Agaves and Paula's Pancake House, respectively. These restaurants are the three most popular restaurants. Moreover, "great food", "food good", "food great", "good food", "service great", and "great service" may be the reasons for top 10 restaurants' popularity.

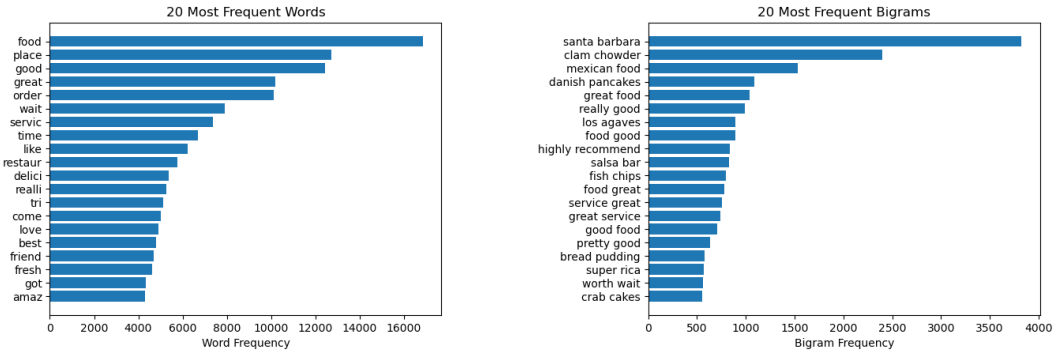

Figure 13: 20 Most Frequent Unigrams and Bigrams

4.2 Sentiment Analysis

Besides the word cloud and BoW models, we also want to use historical review with known sentiment to predict the sentiment of a new review. Rating with 4 or 5 stars is labelled as positive reviews, while rating with 1 or 2 stars is labelled as negative reviews. However, according to Figure 11, this dataset is unbalanced. To deal with unbalance, we randomly selected 5000 positive reviews among all the positive reviews. At the same time, we kept all the negative reviews. Then we combined these positive reviews and negative reviews, and use 70% of them as training data and 30% of them as validation dataset.

We built two models, logistic regression and Naive Bayes model. The ROC curve of logistic regression and Naive Bayes model based on validation data set are summarized in Figure 14. Both logistic regression and Naive Bayes have similar performance. And their AUC is 0.861.

We use G-Mean to find the cut-off probability for classification. And G-Mean is defined in the following,

$$\text{G-Mean} = \sqrt{TPR \cdot (1 - FPR)},$$

where TPR and FPR are true positive rate and false positive rate.

With the optimal cut-off probabilities, the confusion matrices are shown in Figure 15. The left panel is for logistic regression and the right panel is for Naive Bayes model. And both models have good performance. They also have the same accuracy. That is 78%.

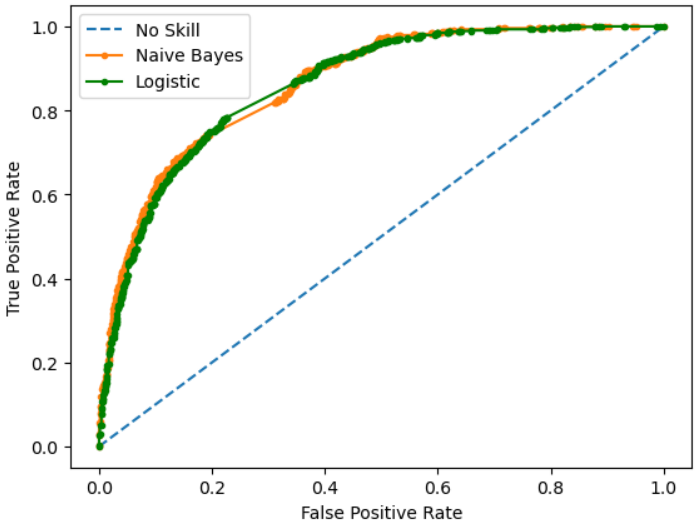

Figure 14: ROC Curve for Logistic Regression and Naive Bayes

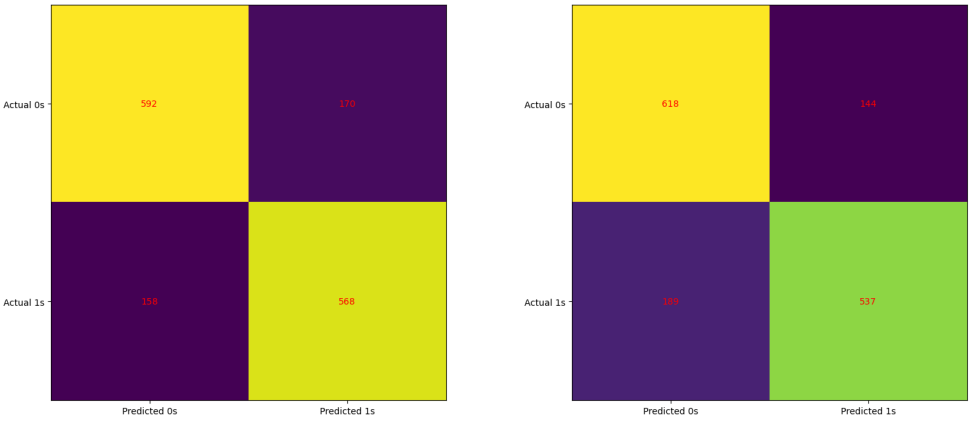

Figure 15: Confusion Matrices for Logistic Regression and Naive Bayes

5 Conclusion

In this project, we analyzed the Yelp review data of restaurants in Santa Barbara. We fist visualized the overall data of restaurants, and found that most restaurants are located in Goleta and Santa Barbara Downtown. Then we analyzed the most popular seafood restaurant, Brophy Bros. According to our analysis, the positive reviews are largely due to clam chowder, fresh seafood and baked clams. Moreover, we applied text mining tools to the reviews of top 10 popular restaurants. The food quality and service quality may be the reasons for top 10 restaurants' popularity. Besides, we also built logistic regression and Naive Bayes model to predict the sentiment of a

new review. These two model have similar performance. And their overall accuracy are both around 78%.

APPENDIX:

Table 1: Summary of Top 3 region of top 9 categories

Categories	zip	count	mean rate
Seafood	93101	13	4.115385
Seafood	93117	5	3.900000
Seafood	93109	4	4.125000
Sandwiches	93101	14	4.250000
Sandwiches	93117	12	3.750000
Sandwiches	93454	10	3.800000
Pizza	93101	14	4.142857
Pizza	93436	14	3.750000
Pizza	93117	13	3.769231
Mexican	93117	21	4.023810
Mexican	93436	18	3.750000
Mexican	93013	12	3.583333
Italian	93101	9	4.055556
Italian	93463	6	4.083333
Italian	93013	5	4.200000
Burgers	93436	13	3.153846
Burgers	93454	12	3.375000
Burgers	93117	11	3.909091
Breakfast & Brunch	93117	14	4.142857
Breakfast & Brunch	93101	12	4.208333
Breakfast & Brunch	93454	9	3.500000
Bars	93101	33	4.075758
Bars	93436	15	4.100000
Bars	93463	13	4.153846
American (Traditional)	93454	13	3.307692
American (Traditional)	93117	11	3.954545
American (Traditional)	93013	8	3.875000
American (New)	93101	16	4.250000
American (New)	93463	7	4.071429
American (New)	93117	6	4.000000

count is the amount of restaurants in corresponding region. *mean rate* is the average ratings of the restaurants in corresponding region.