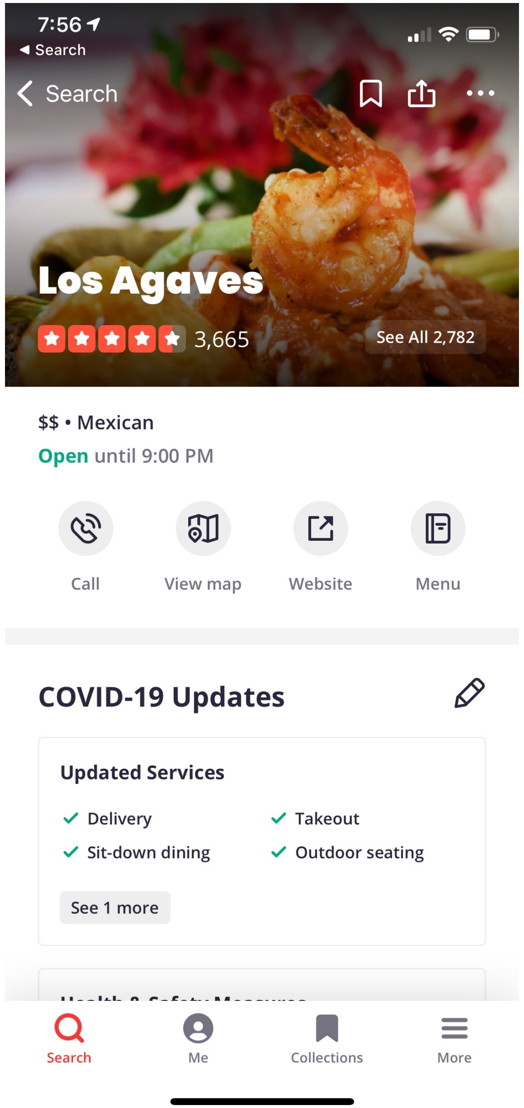
Yelp Review Analysis of Top 10 Restaurants in Santa Barbara

Sunpeng Duan
Mengye Liu
Zhe Li

Introduction



7:56

◂ Search

Search

Los Agaves

★★★★★ 3,665

See All 2,782

$$ • Mexican

Open until 9:00 PM

Call · View map · Website · Menu

COVID-19 Updates

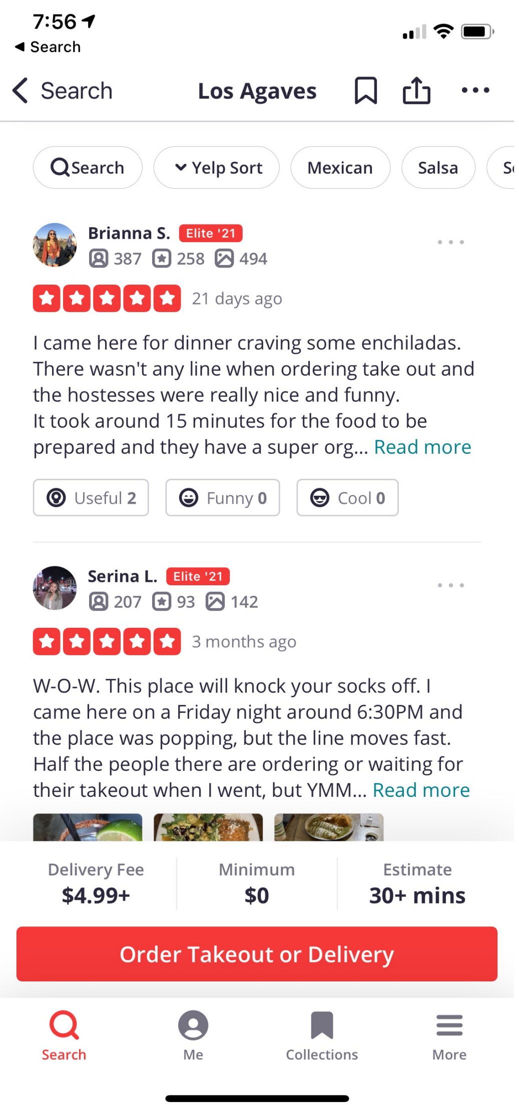
Updated Services

✓ Delivery ✓ Takeout

✓ Sit-down dining ✓ Outdoor seating

See 1 more

Search · Me · Collections · More



7:56

◂ Search

Search **Los Agaves**

Search · Yelp Sort · Mexican · Salsa

Brianna S. Elite '21

387 · 258 · 494

★★★★★ 21 days ago

I came here for dinner craving some enchiladas. There wasn't any line when ordering take out and the hostesses were really nice and funny.
It took around 15 minutes for the food to be prepared and they have a super org... Read more

Useful 2 · Funny 0 · Cool 0

Serina L. Elite '21

207 · 93 · 142

★★★★★ 3 months ago

W-O-W. This place will knock your socks off. I came here on a Friday night around 6:30PM and the place was popping, but the line moves fast. Half the people there are ordering or waiting for their takeout when I went, but YMM... Read more

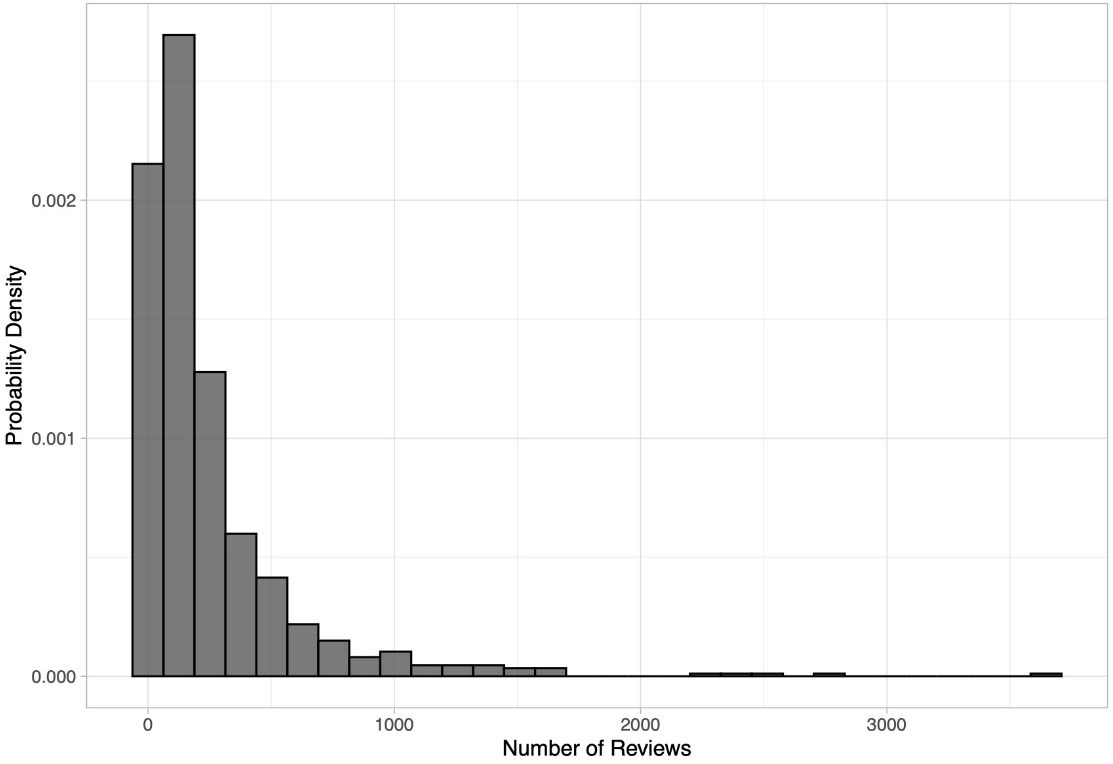
Delivery Fee	Minimum	Estimate
$4.99+	$0	30+ mins

Order Takeout or Delivery

Search · Me · Collections · More

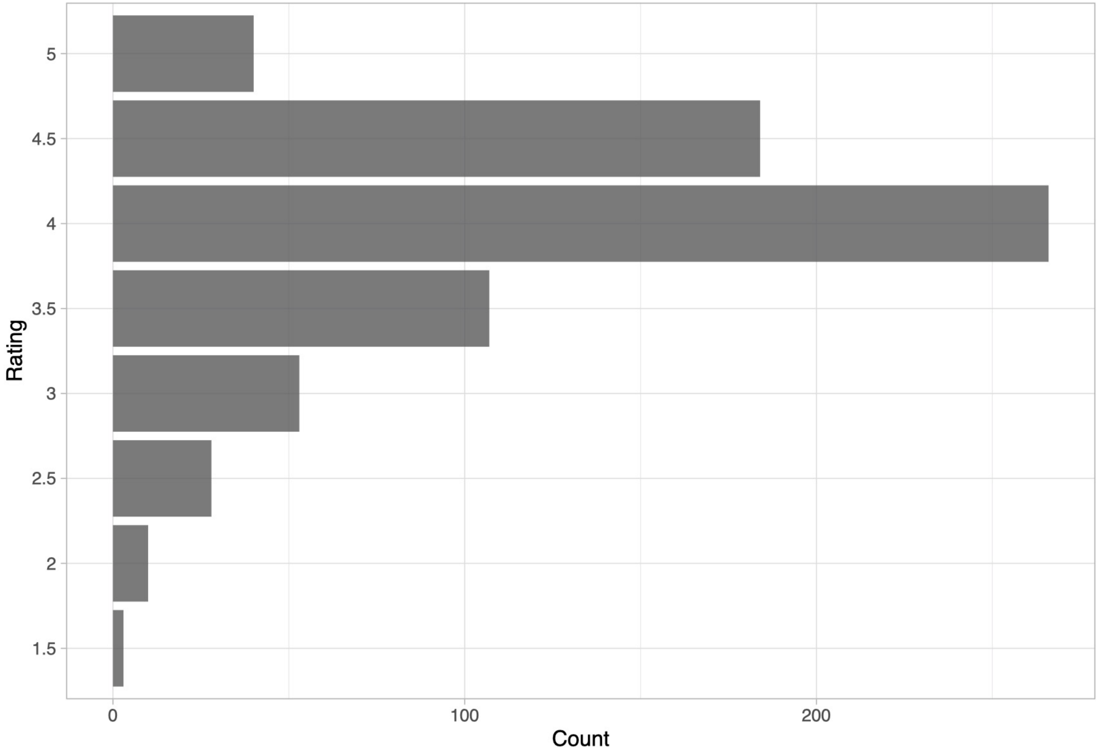
- Yelp is a crowd-sourced local business review and social networking software.
- Provide an overall summary for a specific restaurant.
- Records the customers’ rating and review for a particular restaurant.
- **Goal** : extract sentiment features from review data, and identify positive and negative reviews.

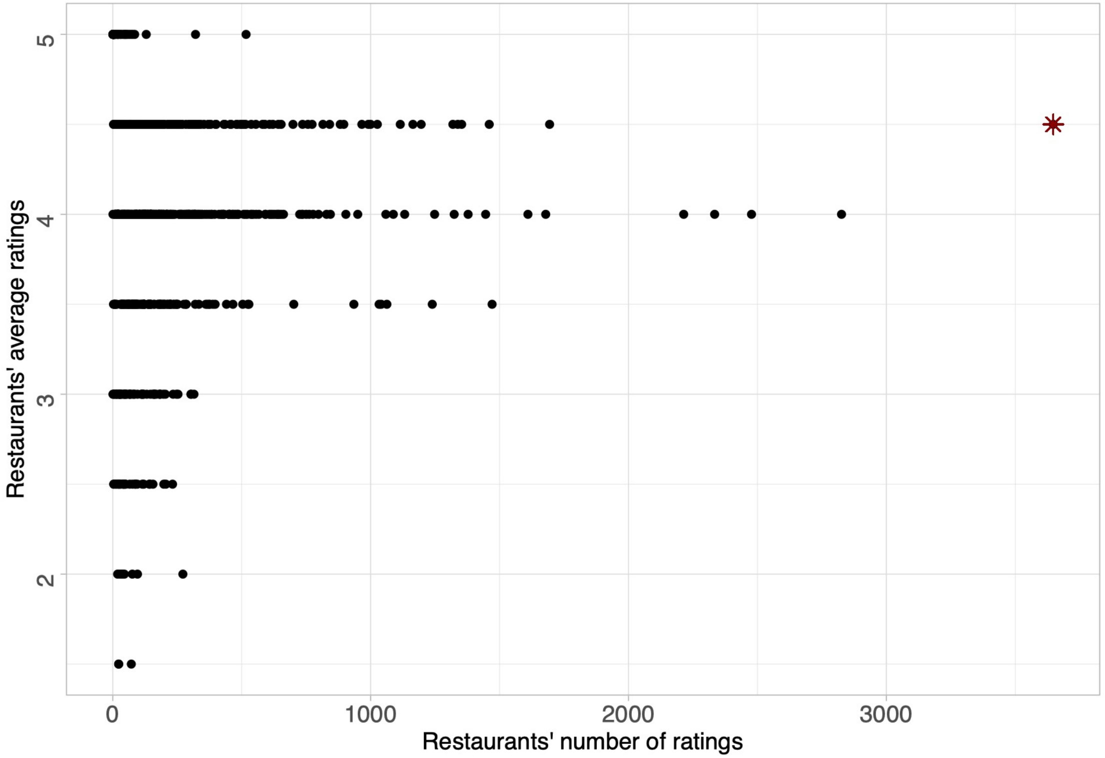
Data Description



- Scrap and crawl the overall information of all the registered restaurants in Santa Barbara county on Yelp.
- Collected all the customer reviews for the restaurants which have the top 10 largest numbers of reviews.
- Total 691 restaurants registered on Yelp in Santa Barbara.
- The distribution of number of reviews is highly right-skewed. And only five restaurants have over 2000 reviews.

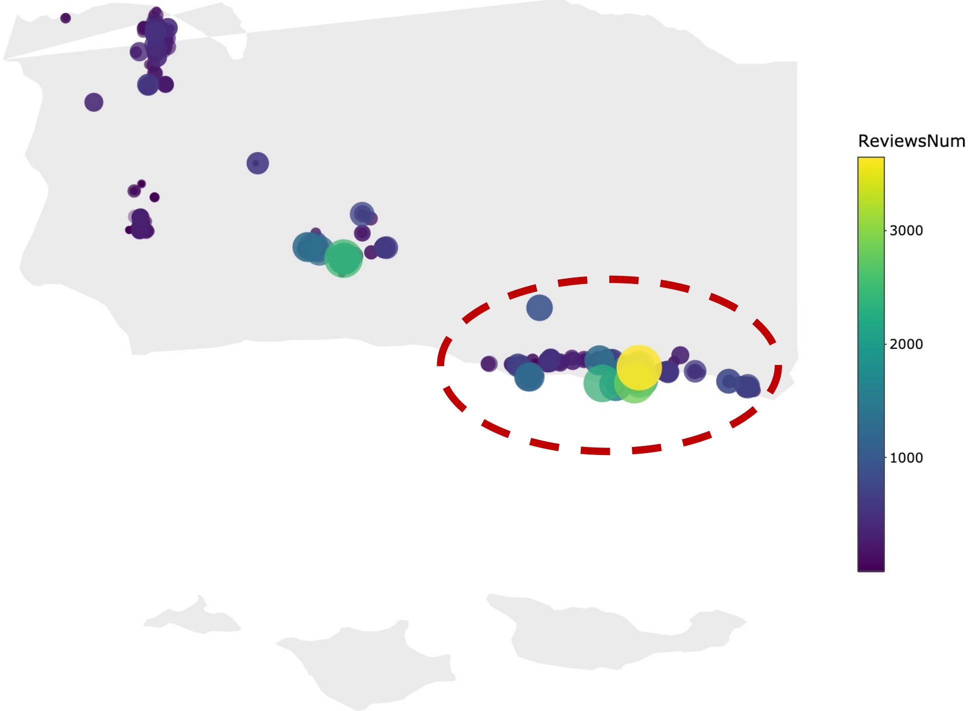
Data Description

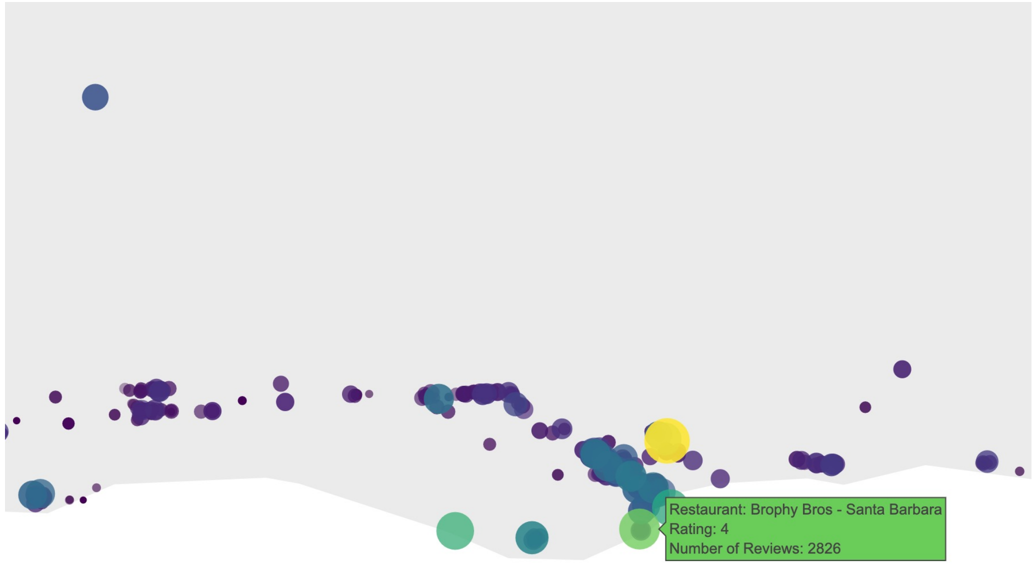




- Over 85% of restaurants are rated over 3. Popular restaurants attract more customers.
- The number of ratings for each restaurants is positively associated with its average rating score.

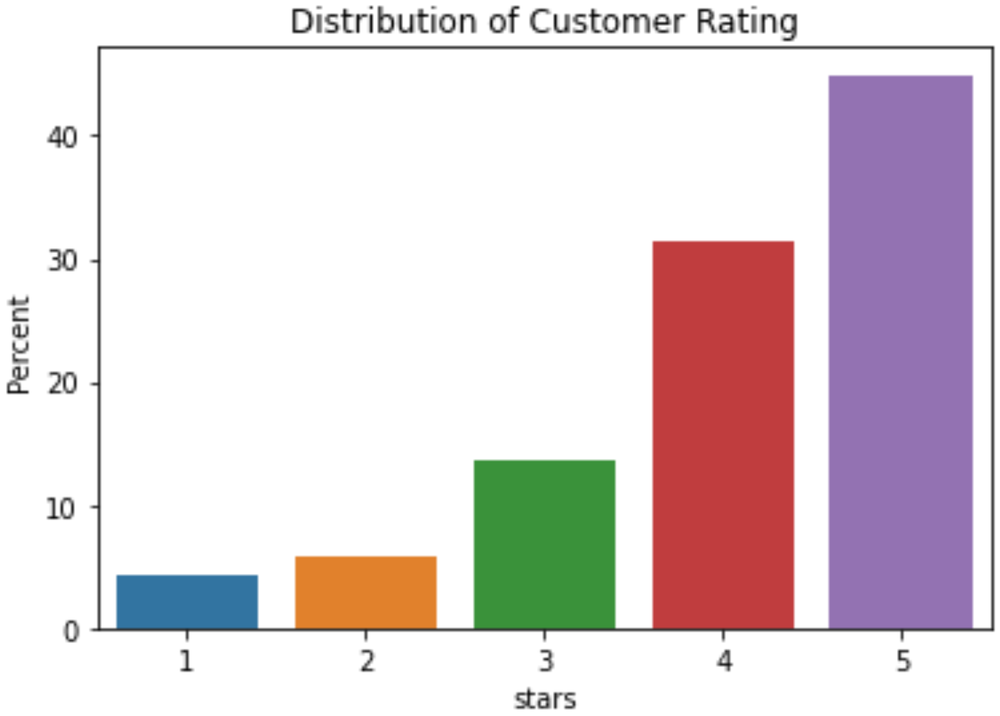
Data Description




Restaurant: Brophy Bros - Santa Barbara
Rating: 4
Number of Reviews: 2826

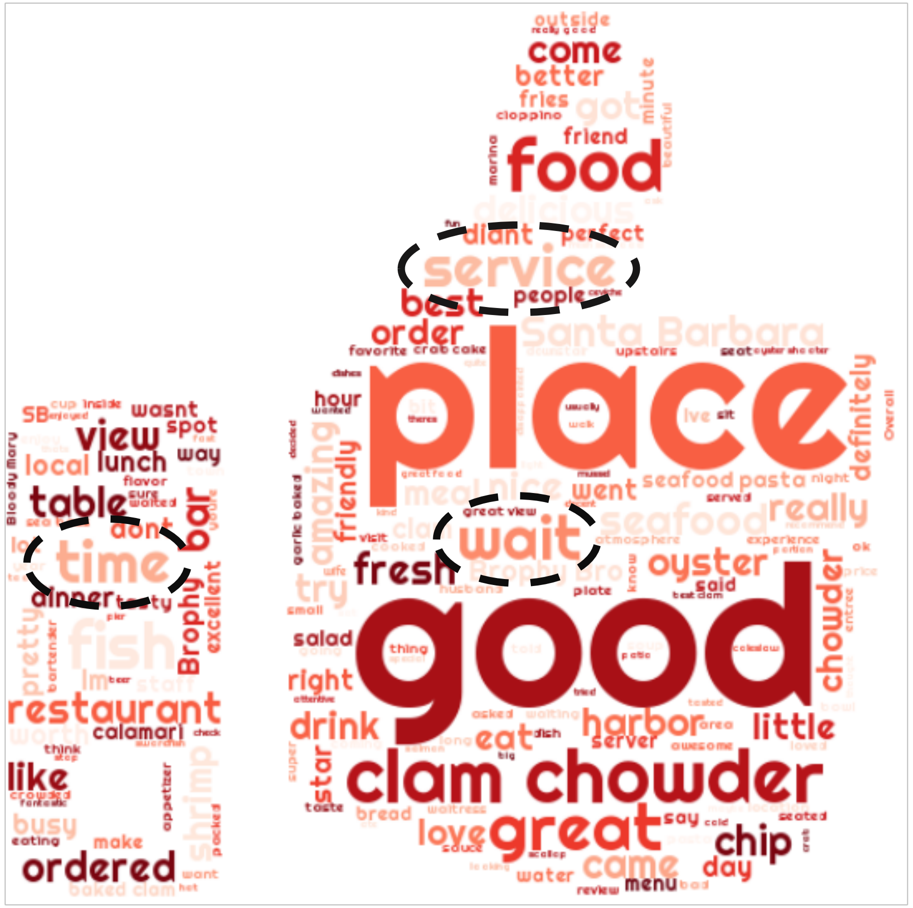
- Over 85% of restaurants are rated over 3. Popular restaurants attract more customers.
- The number of ratings for each restaurants is positively associated with its average rating score.

Review Analysis for Brophy Bros



- Brophy Bro: the most popular seafood restaurant in Santa Barbara.
- The average rating of Brophy Bros is 4 stars based on 2834 customers' reviews.
- over 70% of the customers rate Brophy Bros with 4 or 5 stars.
- **Goal** : analyze the reason of its popularity.

Text Analysis



- ❖ The word “good” indicates majority of the reviews are positive.
- ❖ These positive reviews may be due to its service, waiting time, and etc.
- ❖ 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.

Bag-of-Words

Model

- Bag-of-Words (BoW) model: text representation in numbers.
- A review by one customer is basically represented as the multiset of its words, disregarding grammar and even word order but keeping multiplicity.
- The term frequency of each word could be measured.

Data Cleaning

- Process our text data to convert text into vector format by splitting a review into individual words and returning a word list.
- Remove punctuations and stop words, such as, a, an, the.
- Deal with variations of each word.
- For example, the base form “go” may appear as went, gone, going, and goes in reviews.

N-grams

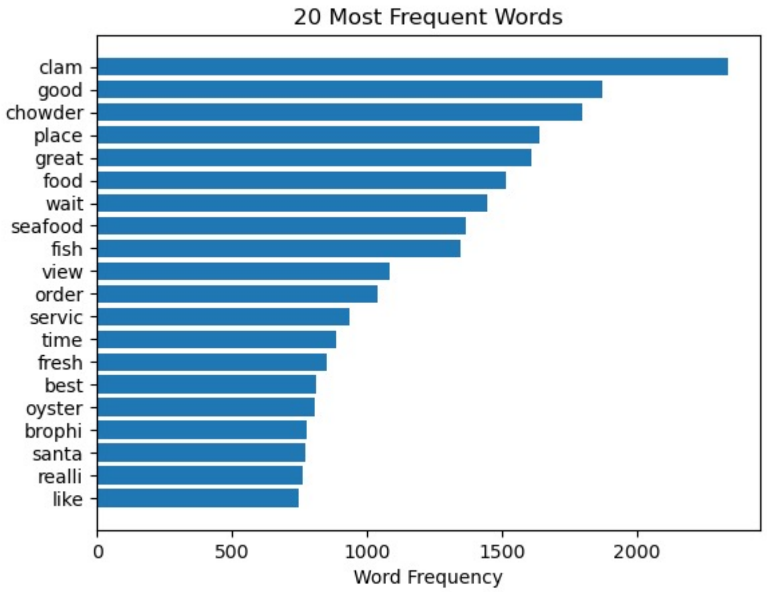
Any limitation of BoW? Order of words matters!

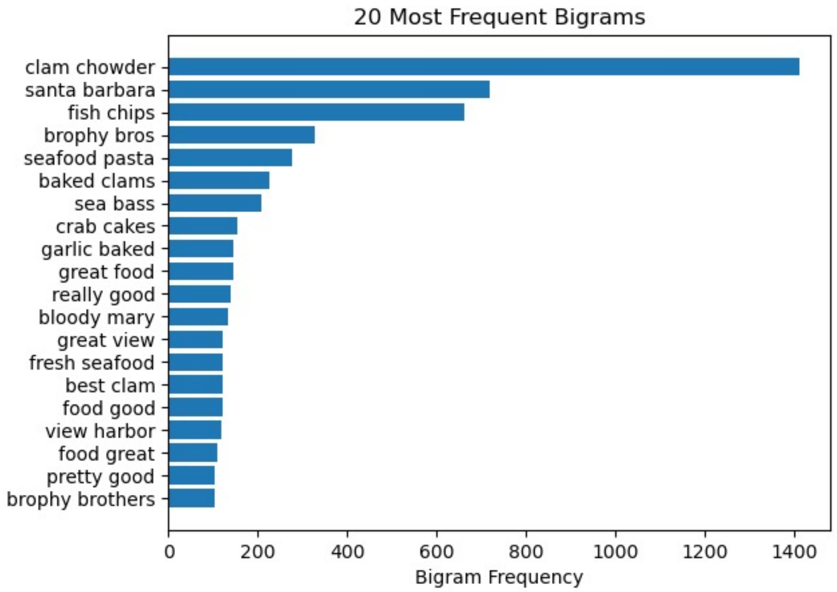
Example:

I am happy, not sad. **I am sad, not happy.**

- 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.
- We consider bigrams!

BoW vs N-Grams





- The 20 most frequent words and bigrams share some similarities.
- Overall, most of the 20 most frequent words and bigrams are related to food.
- "calm" and "chowder" are among the top 3 most frequent words, while "calm chowder" are the most frequent bigrams. ➡ Popular item: clam chowder.
- Customers also prefer fish chips and seafood pasta.
- Most customers also mention service.

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.

Results

- 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.

Sentiment Analysis

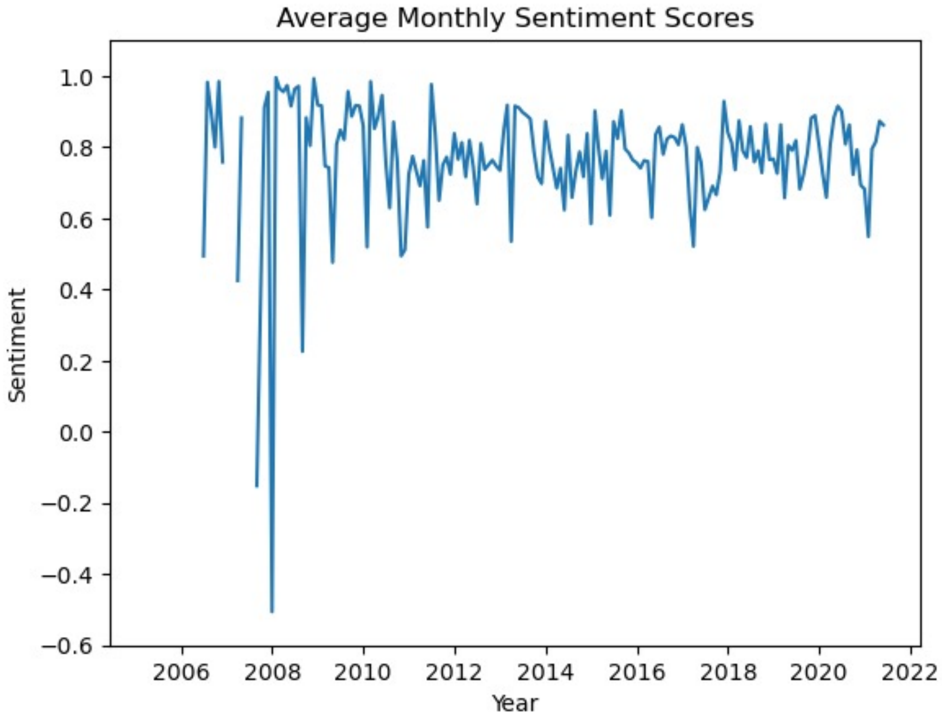
VADER Analyzer

- 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.
- Given a piece of the text, the VADER analyzer returns scores with four kinds - negative, neutral, positive and compound.

SVM

- Sentiment label : positive (4 or 5 stars) vs negative (1 or 2 stars)
- Design matrix: tf-idf frequency matrix

VADER Analyzer



- The compound score ranging from -1 to 1 is a combination of positive and negative scores.
- 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.
- After 2010, the average monthly sentiment scores tend to be stationary.

SVM: TF-IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

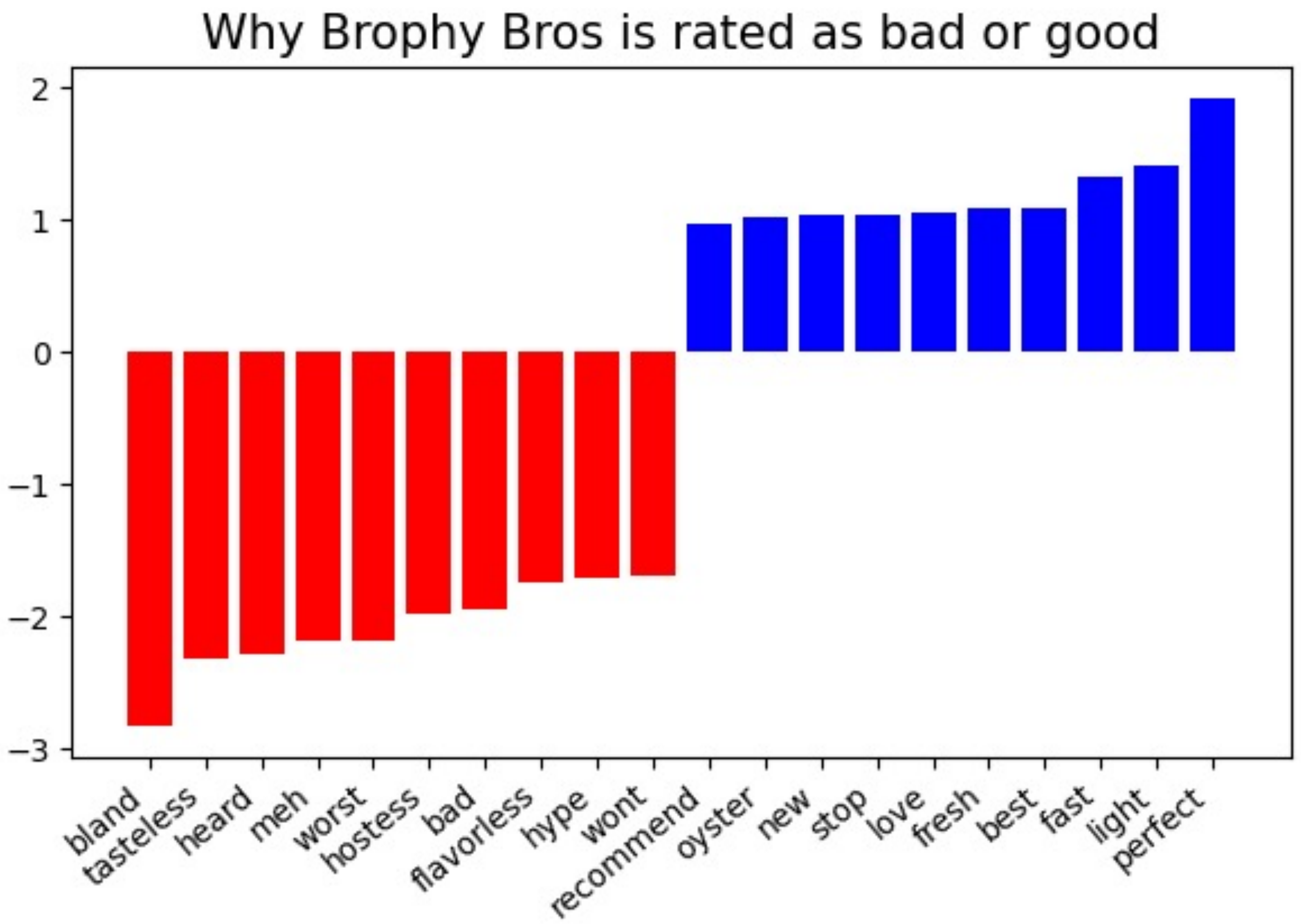
df_x = number of documents containing x

N = total number of documents

	Review 1	Review 2	Review3
{perfect}	tf-idf scores	…	…
{light}	…	…	…
{fast}	…	…	…

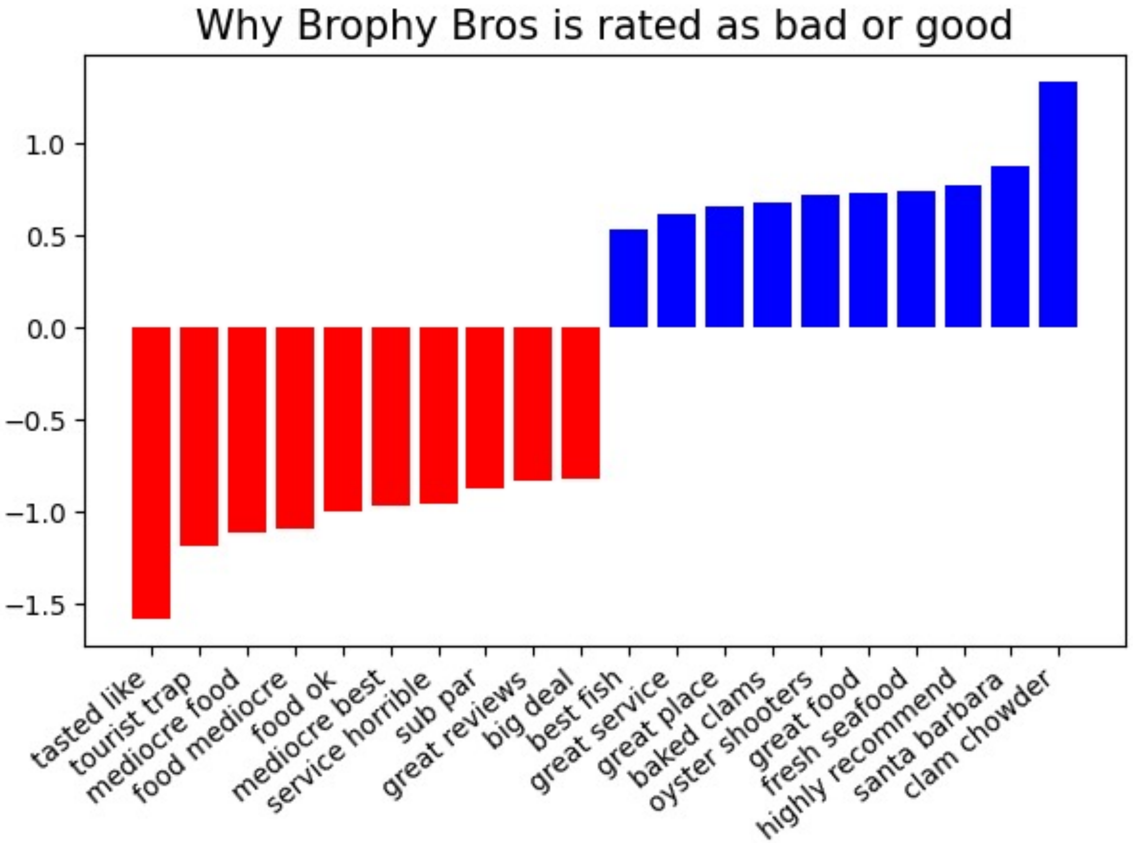
	Review 1	Review 2	Review3
{calm chowder}	tf-idf scores	…	…
{fresh seafood}	…	…	…
{great service}	…	…	…

SVM



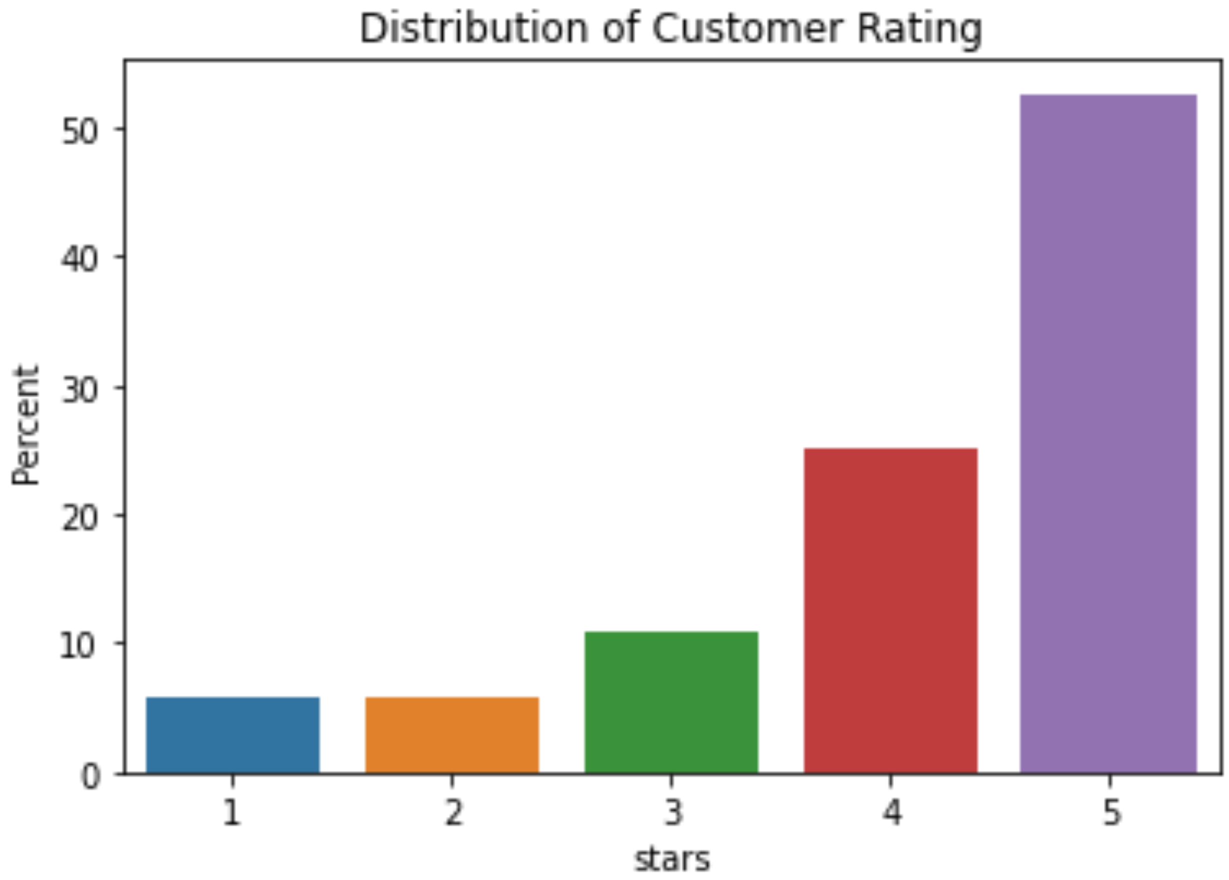
Uninformative!

SVM



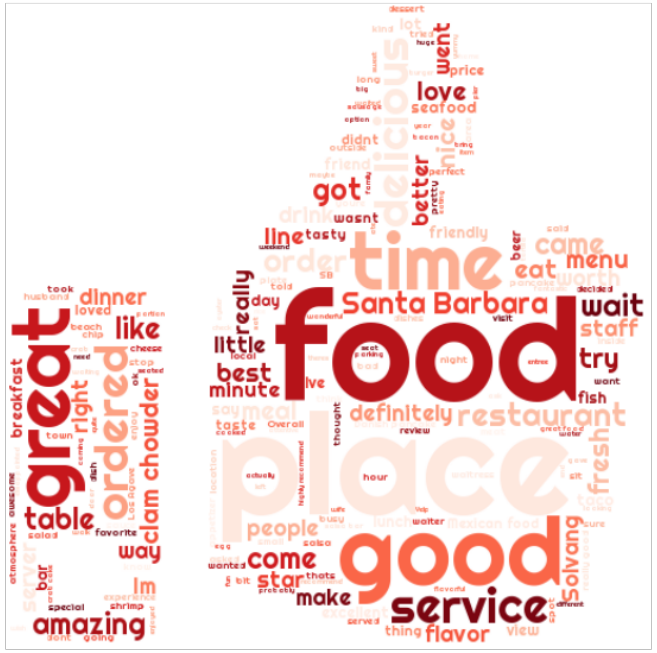
- The model with bigram tf-idf matrix indicates the bigrams such as clam chowder, fresh seafood and baked clams are associated with the positive reviews.

Top 10 Restaurants



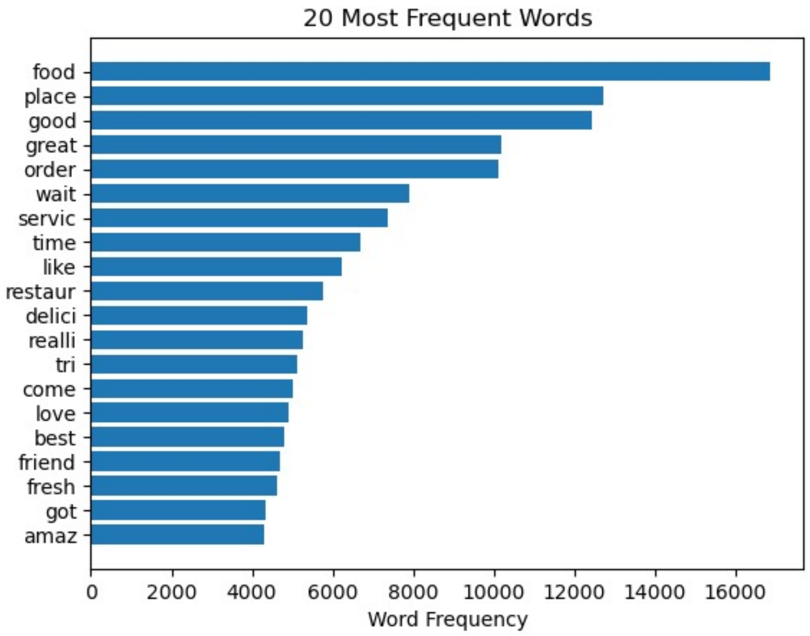
- ❖ Over 70% reviews are rated with 4 or 5 stars. This is similar to the distribution of ratings for Brophy Bros - Santa Barbara.

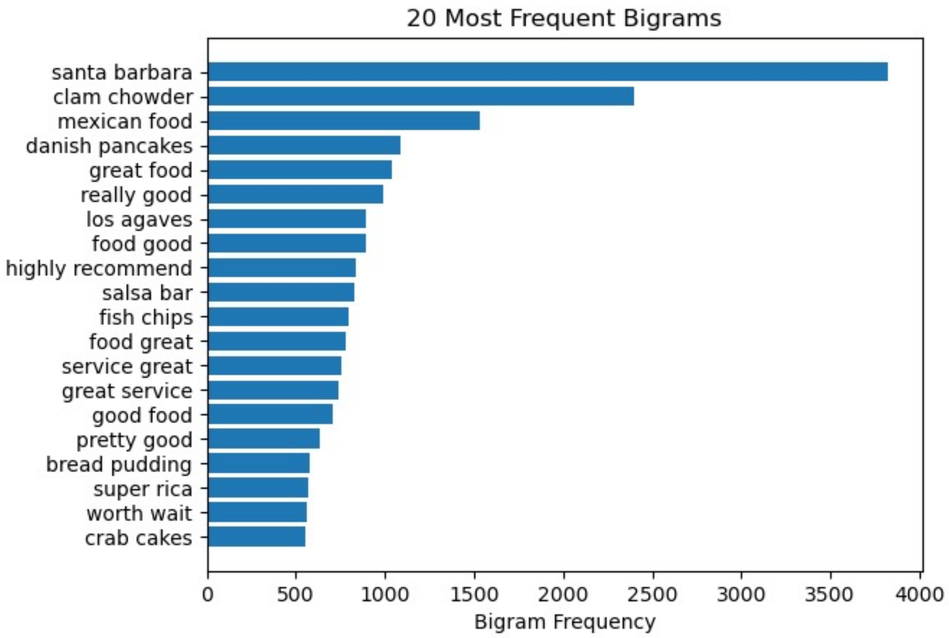
Text Analysis



- ❖ The word “food” and “place” are most frequent among the reviews of top 10 restaurants, indicating that food and place are common factors affecting the customers’ reviews.
- ❖ The word “great” and “good” also indicate majority of the reviews are positive.

Text Analysis


20 Most Frequent Words

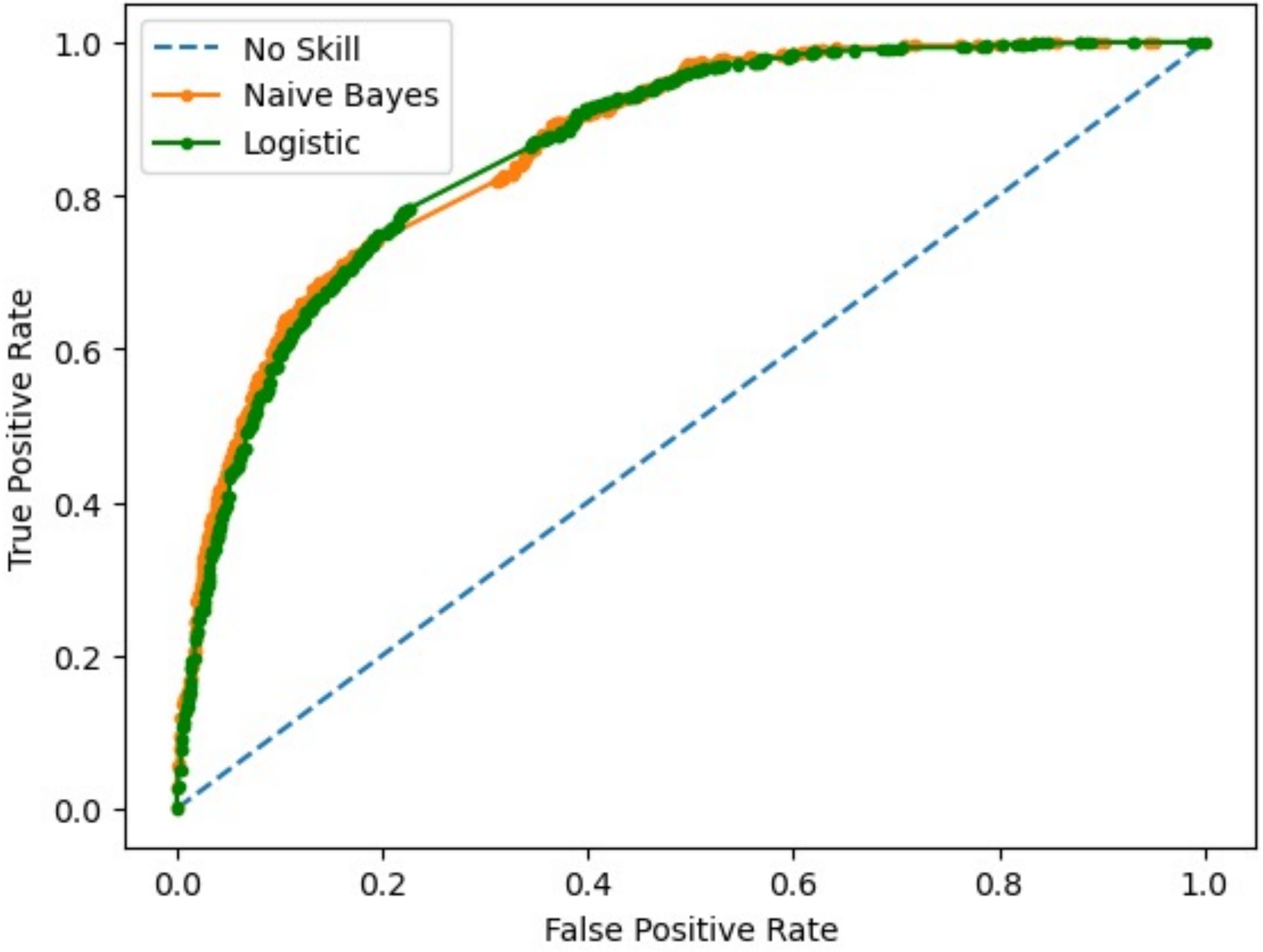

20 Most Frequent Bigrams

- 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.
- Moreover, “great food”, “food good”, “food great”, “good food”, “service great”, and “great service” may be the reasons for top 10 restaurants’ popularity.

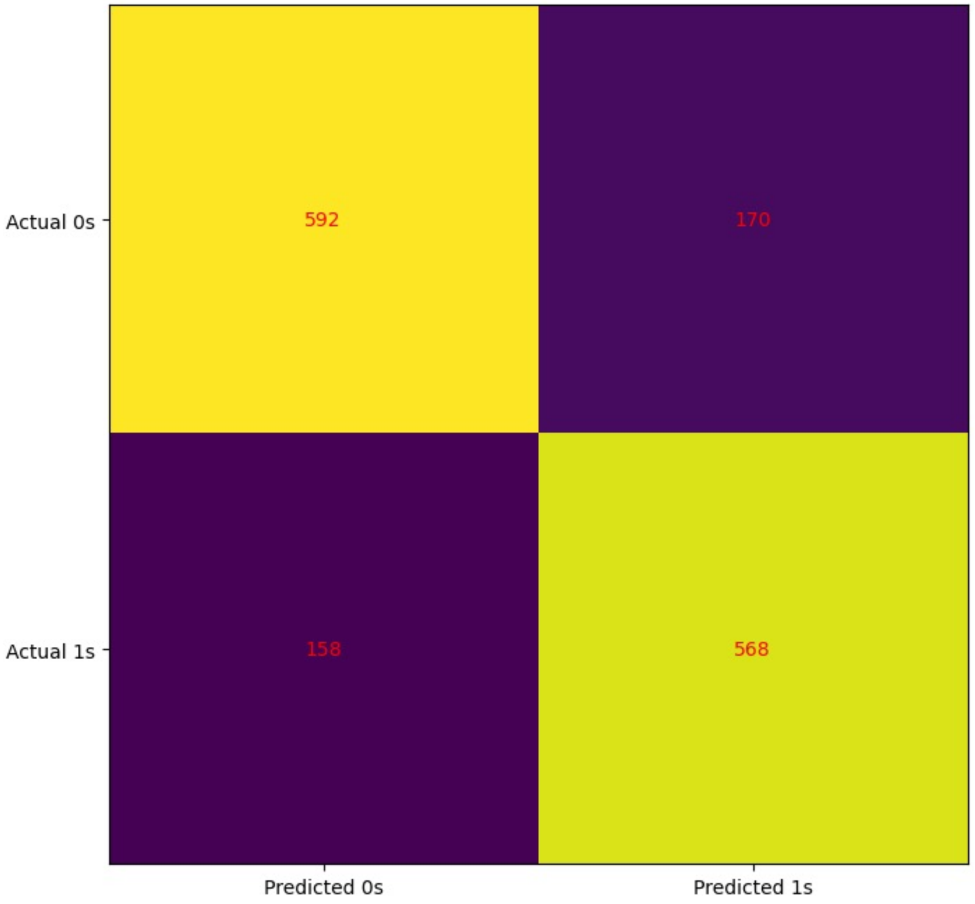
Sentiment Analysis

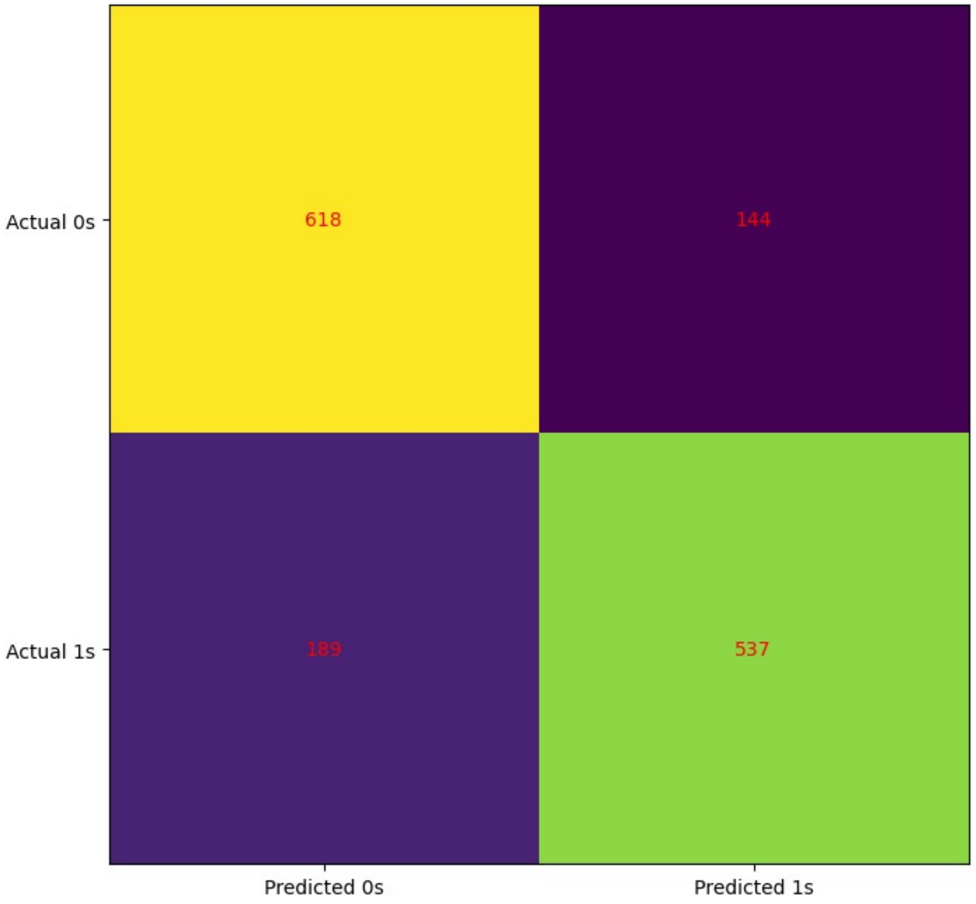
- Sentiment label : positive (4 or 5 stars) vs negative (1 or 2 stars).
- Design matrix: tf-idf frequency matrix.
- Unbalance: randomly select 5000 positive reviews among all the positive reviews and keep all the negative reviews.
- Use 70% of them as training data and 30% of them as validation dataset.
- Two models: logistic regression and Naïve Bayes model.
- Use G-Mean to find the cut-off probability for classification.
- G-Mean = √ TPR · (1 − FPR).

ROC Curve


No Skill
Naive Bayes
Logistic
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Confusion Matrix





Thank You