Restaurants in Hoboken, NJ: Predicting review ratings based on review text terms

Authors: Daniel Hunt and Sufyan Gulab Shaha

Web Mining

BIA-660

May 10, 2024

Introduction

Hoboken is known for having the most bars per capita of any city in the United States. Combined with delis, coffee shops, pizza parlors, and byob restaurants, Hoboken has a robust and highly competitive restaurant ecosystem.

Hoboken is also home to several students, workers, families, and tourists. To find the BEST option for food, people often rely on word of month or other endorsements for places to eat (Jolson and Bushman, 1978). In steps Yelp.

Yelp is a platform that allows users to leave ratings and reviews for restaurants so food seekers can make decisions based on a vast resource of recommendations. Good reviews lead to more business!

The Problem

For existing business owners, they will want to find ways to increase their Yelp ratings or maintain their premium status to maximize foot traffic.

For new business owners, they may want to find a the quickest path to high quality reviews to become relevant in the cluttered ecosystem.

Both owners need to understand what makes a good and bad review and can they forecast ratings changes based on sample reviews?

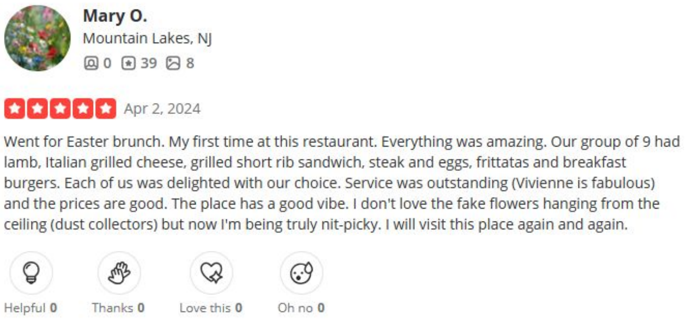
The Data

Data Scope:

- Restaurants in Hoboken
- Most Reviewed restaurants
- Up to 9 pages of reviews (40+ individual reviews)

Data:

- Rating
- Review Text



	Rating	Review
0	4	The paella was delicious, my only comment is t...
1	4	Excellent food. Highly recommend the Mar Y Tie...
2	5	It had been a while since I've been back to Ho...
3	4	This Cuban restaurant is a surprise at the beg...
4	4	In addition to the paella which was a bit too ...

Methodology

A selenium based tool was used to scrape the data from yelp after navigating the website and sorting the restaurants by most number of reviews. The raw data was then cleaned to make it suitable for use, which also included text processing like removing stopwords and punctuation.

Once it was cleaned, we used a TF-IDF function to understand the importance of words which are more important or meaningful in the corpus. We then used text classification models to analyze our data and to map the texts to a rating. The models we used for our analysis were an SVM and a NB. They were chosen because they are relatively efficient compared to other models especially when it comes to dealing with high-dimensional data.

Once both models were trained, we assessed the performance of each using precision, recall, accuracy, F-Score, and AUC scoring.

NB Model

A Multinomial Naive Bayes (NB) classification model predicts outputs based on probabilistic connections between features and labels. The model calculates the probability of each class given the features and selects the class with the highest probability as the predicted class for the input.

It assumes that the features are independent of each other given the class label.

We used MLE alongside our model with alpha set to 0. Since we already knew about the imbalance in our review data, as yelp tends to display positive raings most of the time to incentivize it's platform with positive reviews, or the community being simply too positive, we factored in a class weighting system to update the probabilities based on samples and class counts to balance the labeling frequency.

SVM Model

The second model was an SVM. It is designed to classify data into classes and create maximum separation between each class using a vector as the divider to find a hyperplane. This hyperplane is chosen to maximize the margin between the closest points of different classes, which helps improve the generalization of the model.

To address the stated imbalance issue, we incorporated class weights to make sure the ratings with fewer reviews received a higher weight against their probabilities.

This meant a lift on 1 and 2 star review text.

Performance

Both models were accessed using precision, recall, accuracy, F-Score, and AUC scoring. We found out that the most robust measurement for performance in this report was the AUC scores for each class.

AUC scores indicate how good the model is at separating the classes.

In our analysis, we found out that the AUC scores carried between the two models.

Naive Bayes model had moderate score for one, two, three and five star ratings, for the four star rating however, the scores were quite low

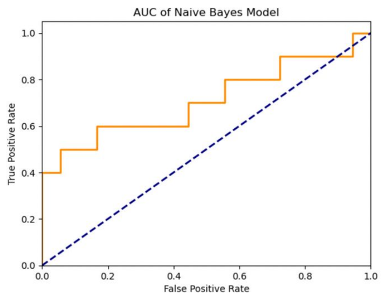
AUC for class 1: 71.88%

AUC for class 2: 71.88%

AUC for class 3: 68.00%

AUC for class 4: 46.94%

AUC for class 5: 71.11%



Performance

AUC scores for the SVM Model were lower than the scores of our NB Model.

The results showed that SVM did good at classifying the ratings for three, four, and five star reviews, but performed poorly when it came to classifying one and two star reviews.

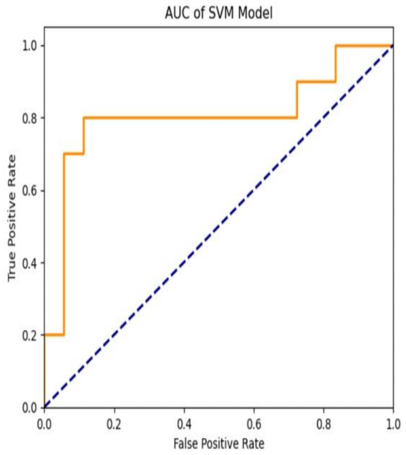
SVM - AUC for class 1: 57.29%

SVM - AUC for class 2: 58.33%

SVM - AUC for class 3: 76.00%

SVM - AUC for class 4: 67.35%

SVM - AUC for class 5: 80.56%



Performance-NB

Precision, recall and F-Score indicate the performance at a specific class threshold.

The Naive Bayes model showcases low performances with weighted average scores around 30%.

The results were **Precision - 30%, Recall - 36%, and F-Score - 29%**

We can assess that the model does moderately well to separate the data but performs poorly when it comes to predicting the ratings. Thus the model does have potential and can identify some ratings, but not the others.

The model fails to classify four star ratings due to high overlaps with five star and three star ratings.

Performance- SVM

The SVM model showcases even lower performance than the Naive Bayes model with weighted average scores ranging from 20% - 35%.

The weighted average scores were **Precision - 19%, Recall - 26%, and F-Score - 24%.**

One and Two star ratings received ‘0’ scores across precision, recall and F-Score.

Five star reviews however performed far better with **39% precision, 90% recall and 55% F-Score.**

Result and Outcome

In a test to see if the models could accurately predict the rating of text, we selected at random two reviews:

Dori Q.
Perth Amboy, NJ
16 9 2

Feb 15, 2022

I have to start off by saying that the only good part about this restaurant is the drinks. The service was ok but the prices for the food were too expensive for the amount they put on your plate. Wasn't worth the money. To top it off we felt rushed out of the restaurant couldnt eat the food on your plate because they kept asking if I was done knowing that I was still eating. The restaurant closed at 10pm and started shutting off the lights at 9:30pm. We literally felt kicked out. Will not go there again.

Mimi N.
Northeast Philadelphia, Philadelphia, PA
0 10 27

Jul 2, 2021

This place is the worst representation of Cuban food. We ordered delivery and got the paella for $29, shredded beef for $27, churros for $10. The Paella was bland and there was little to no sea food. Also the Portions are extremely small for the price. We got 1 small cup of rice for the shredded beef and the container was mostly liquid instead of meat. Huge disappointment. The Churros were soggy and only had four small pieces. Overpriced bland food. They should really work on their portion size. Literally ripping people off.
I will never order from this place again and will not recommend to anyone.

Result and Outcome

Both models came to the same conclusion, however they were both incorrect. In the second sample, the models predicted the opposite ranking of label.

```
new sample tf_idf size: (1, 1375)
Naive Bayes Prediction:  ['3']
SVM Prediction:  ['3']
Actual: 2
```

```
new sample tf_idf size: (1, 1375)
Naive Bayes Prediction:  ['5']
SVM Prediction:  ['5']
Actual: 1
```

Result and Outcome

As we take a closer look at the second sample, we see many words that were associated with high ratings - menu items like **beef, churros, paella, portion, and recommend**.

The algorithms will have scored these words as highly probable to appear in five star reviews. Our assumption is this is due to the imbalance in reviews so common food terms are more associated with high ratings purely based on the volume of the five star reviews.

In Conclusion

Due to the low performance of the models, more training is required and additional segmenting should be applied to achieve better results:

- Sentiment analysis to isolate sentiment based on rating to better categorize text
- Bigram analysis to see if grouping of words changes the outcome of the algorithm
- Inclusion of additional features that may influence quality of review: type of cuisine, location of reviewer (does distance travelled influence perception), or a reviewers reviewing history which might influence the weighting parameters

However, the tool does **highlight most common terms** which can be an indicator of themes an owner should allocate more attention to to maximize (positive) conversation around.

Reference

Jolson, M. A., & Bushman, F. A. (1978). Third-party consumer information systems: the case of the food critic. Journal of Retailing, 54(4), 63–63.