Analysis of Yelp reviews for Cafés in NYC

Course: BIA-660 Web Mining
Under Prof. Rong (Emily) Liu
Vaibhav Shanbhag, Krishma Shah, Sukit Kajonpradapkul, Jui Pendharkar
Stevens Institute of Technology
Hoboken, NJ

Motivation

Social media reviews play a vital role in customers' decision making when selecting places to eat/ drink. Reviews from customers can generate a positive impact on the business as a whole. ... It's also about what potential customers can experience from your brand. Reviews on social media enable the customer to highlight a service that stands out. Many people use Yelp to find a good restaurant. Nonetheless, with only an overall rating for each restaurant, Yelp offers not enough information for independently judging its various aspects such as environment, service or flavor. This will help companies improve their business model by understanding customer feedback and expectation through analysis of unstructured text data using NLP.

Objectives

Online reviews and customer feedback can give businesses great insight in order to improve the business model of different cafés. Our goal is to analyze the reviews for Coffee shops and Cafés in Manhattan, NY and develop methods to extract business intellect from reviews, to understand what factors cause customers to be satisfied or dissatisfied. This can help them to improve customer satisfaction and their profit.

Introduction

Yelp is an American multinational corporation founded in 2004 which aimed at helping people locate local business based on social networking functionally and reviews. The main purpose of Yelp is to provide a platform for customers to write review along with providing a star-rating along with an open-ended comment. Yelp data is reliable, up-to-date and has a wide coverage of all kinds of businesses. Millions of people use yelp and empirical data demonstrated that Yelp café reviews affected consumers' food choice decision-making. What makes a café good? What are the major concerns of customers regarding a café? Through this project, we are going to extract the common features like food/coffee, services, experience, etc. behind all kinds of cafés with different analysis on Yelp review data.

The project is organized as follows: Part 1 gives the overview of the Yelp data and introduces NLP to extract the sentiment of the review and gather information. Part 2 describes this extracted data and classifies the features using many classification algorithms with discoveries with different cafés. We conclude and imply different future applications to gain knowledge for cafés.

The Dataset

For this analysis, we are focussing on the Selenium package to get the dataset of all cafés from Yelp using web scraping. We are concentrating on the list of cafés on the website and extracting these links to get other important attributes necessary for our analysis. Our data contains 10,269 reviews, with the following information for each one:

1. **cafe_name** (name of the café being reviewed)
2. **avg_rating** (Overall rating of the café being reviewed)
3. **review** (Review text)
4. **indv_rating** (rating from 1–5 for that review)
5. **review_date** (Day the review was posted)
6. **useful {1/0}** (Comments on the review, given by other users)

7. **funny{1/0}** (Comments on the review, given by other users)
8. **cool {1/0}** (Comments on the review, given by other users)
9. **latitude** (geographical location co-ordinate of the café)
10. **longitude** (geographical location co-ordinate of the café)

	A	B	C	D	E	F	G	H	I	J
1	cafe_name	avg_rating	review	indv_ratin	review_date	useful	funny	cool	latitude	longtitude
2	11th Street C	4	Wouldn't let me order a coffee w	1	12/15/2018	1	2	0	40.73578	-74.0073
3	11th Street C	4	The selection of drinks and smoo	1	4/15/2017	0	0	0	40.73578	-74.0073
4	11th Street C	4	I've been coming here for years a	1	10/19/2016	0	0	0	40.73578	-74.0073
5	11th Street C	4	The food is not good, the space is	1	9/15/2015	0	0	0	40.73578	-74.0073
6	11th Street C	4	First time here. I'll be here again.	1	09-08-2011	1	0	0	40.73578	-74.0073
7	11th Street C	4	Great place for lunch. Last time I	1	02-08-2012	2	0	1	40.73578	-74.0073
8	11th Street C	4	I was so happy to have discovere	1	12/29/2010	1	0	0	40.73578	-74.0073
9	Ancolie	4.5	So disappointed I wasted 10$ on	1	04-02-2018	1	0	0	40.73325	-73.9991
10	Ancolie	4.5	So I walked by this place a few tir	1	3/15/2018	0	0	0	40.73325	-73.9991
11	Ancolie	4.5	So fresh and delicious. You might	1	12/27/2016	0	0	0	40.73325	-73.9991
12	Bean & Bean	4	I work in the building next door s	1	5/30/2017	0	0	0	40.70731	-74.0122

Figure: dataset formed after web scraping in a csv file

Data Processing

The dataset is scraped from Yelp consisting of 10 attributes and we only focussed on the cafés and coffee shops in the Manhattan area of New York. For this analysis, we only took customer reviews from the Yelp cafés. The emphasis of the study is more on polarising ratings (highly positive and highly negative) and less on neutral ratings (e.g. 3-star ratings). Thus, the reviews have been classified into two categories: Positive (Ratings 4 and 5) and Negative (Ratings 1 and 2).
A subset of about 4000 reviews (positive and negative) were selected from the total scraped reviews. The analysis also calculates the distribution of Positive and Negative rated reviews from the selected subset. Specifically, review content was the corpus for our analysis; rating was the identifier for discriminating positive or negative sentiment and grouping.(See below)

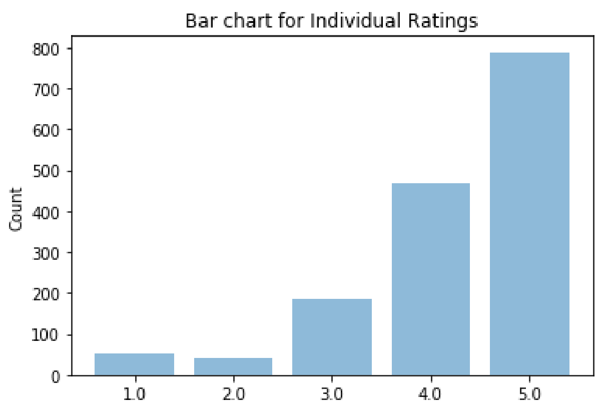


Figure: Distribution of reviews based on each rating

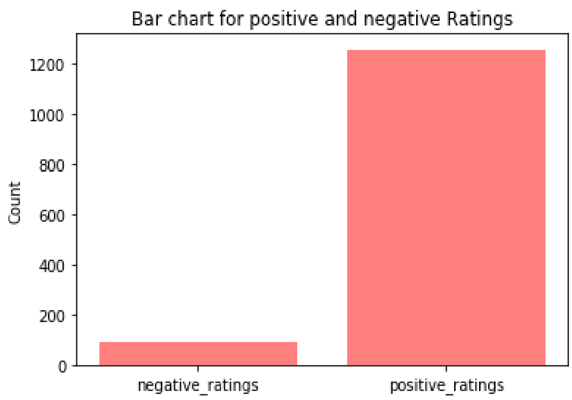


Figure: Distribution of reviews into group of rating

To get a better understanding of the cafés and coffee shops in the vicinity of the New York area, we analyzed these cafés with their ratings to get knowledge of the area which has all the good rating cafés. Below shown map concludes that the Midtown area of Manhattan possess most of the cafes with 4.5 and 5-star ratings.

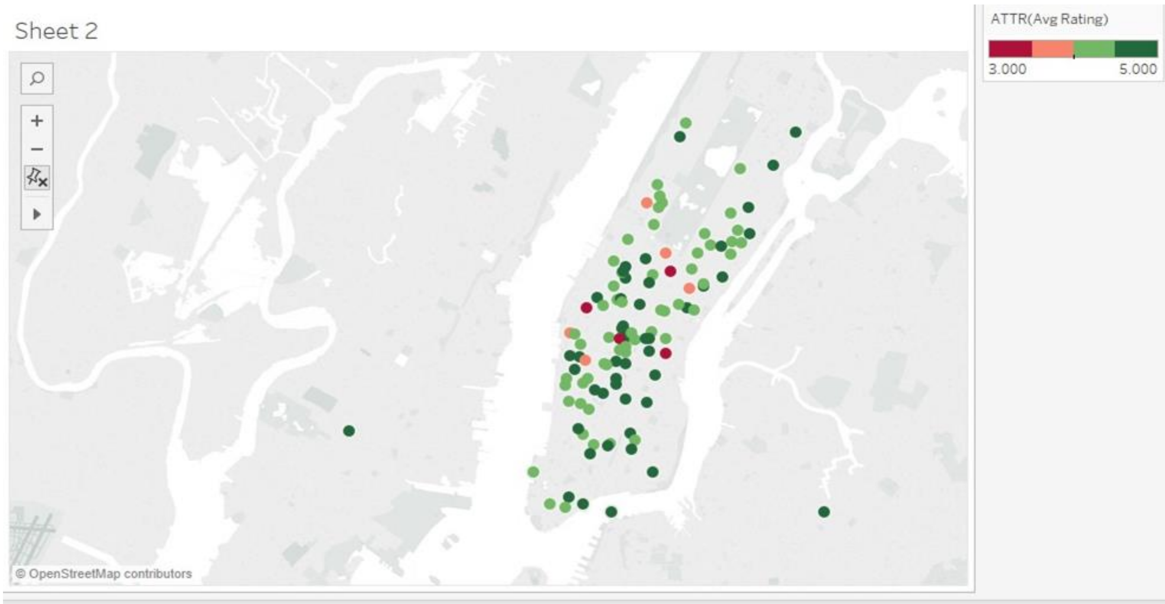


Figure: Distribution of cafes and coffee shops in Manhattan based on overall ratings

The analysis of the data further consists of implementing the data exploration on the length of the review versus the rating of each of those reviews. By analyzing that, it seems like overall, the distribution of text length is **similar** across all five ratings. However, the number of text reviews seems to be skewed a lot higher towards the 4-star and 5-star ratings.

From the box plot too, looks like the 1-star and 2-star ratings have much longer text, but there are many outliers (which can be seen as points above the boxes). Because of this, maybe text length won't be such a useful feature to consider after all. (See graphs below)

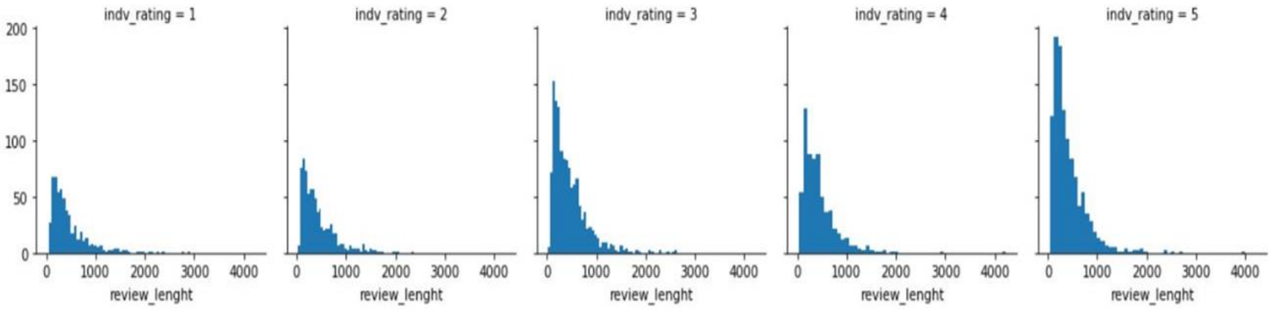


Figure: Histograms of text length distributions for each star rating. Notice that there is a high number of 4-star and 5-star reviews

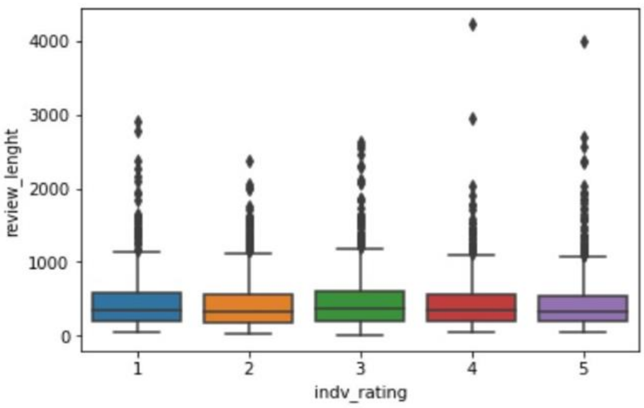


Figure: Box plot of text length against star ratings

Data Cleaning

Review analysis was performed on the review text to get the background overview of the text. Review analysis consists of the following parts: words in each review were separated and the punctuations were removed so that a "bag of words" was generated for each review, then we lemmatized and filtered out the stop words in each bag of words using both the in-built list in Python's SpaCy package.
Specifically, we assumed and labeled reviews with ratings greater or equal to 4 as positive while the rest as "negative". This decision was made based on our observation of the distribution of ratings.

Figure: Top 20 words before lemmatization

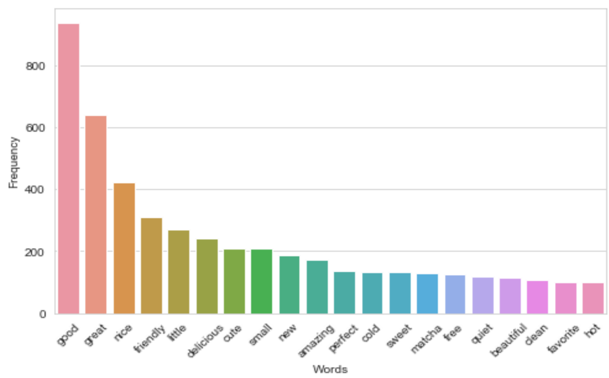


Figure: Top 20 words after lemmatization

We further, classified the review date attribute from the dataset and grouped the months into different classes based on the seasons (e.g. months of January, February, November and December were marked as Winter). Hence, the seasons Summer, Winter, Spring and Fall were concluded, also created bigrams of words and the main aim to perform such an analysis was to gather information from the review and verify if seasons actually impact the customers behaviour and ordering styles. We got to know that customers order hot chocolate in majority during winters than in summer and cold brew, cold coffee and avocado toast in much higher quantities in summers which can be viewed as follows:

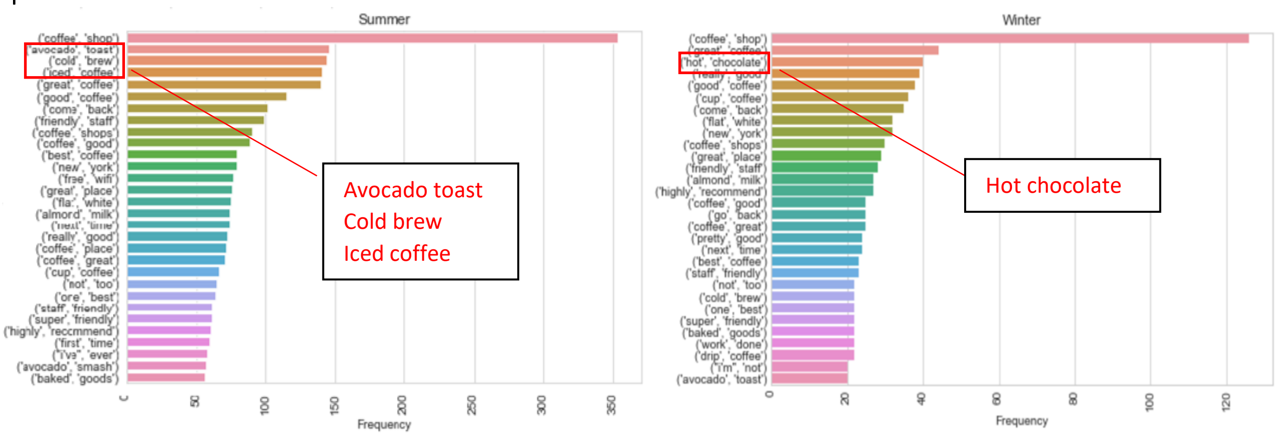


Figure: Top 30 bigrams during the seasons of Summer and Winter

Text Analytics Models

Sentiment Analysis

Sentiment Analysis, also known as opinion mining, is the process of determining whether a text unit is positive or negative. Sentiment Analysis is an active area of study in the field of natural language processing that analyzes people's opinions, sentiments, attitudes and emotions. Since our data consists of opinionated reviews, we need to understand the emotions attached to the words. We utilize sentiment analysis to capture the positive and negative emotion conveyed in each review. Sentiment analysis is different from text mining as it seeks to extract and classify opinion rather than topical information.

We have performed sentiment analysis using both supervised and unsupervised learning methods. The two approaches for Supervised learning are:

a. Multinomial Naive Bayes (NB) Model
b. Support Vector Machine (SVM) Model

For our analysis, we clubbed reviews with individual ratings 1 and 2 as negative reviews and assigned label 0. Reviews with individual ratings 4 and 5 were clubbed together as positive reviews and assigned label 1. Reviews with individual rating 3 were not considered in our analysis.
The data set was then split into 70% training and 30% testing data. GridSearch was used to find the best parameters in both the Naive Bayes and SVM models. The test data was later used on the trained model to calculate performance. For our dataset, Naive Bayes gave better results as compared to SVM model.

Supervised learning: Multinomial Naïve Bayes (NB) Model

Naive Bayes accuracy : 0.6850053937432579

	precision	recall	f1-score	support
0.0	0.68	0.41	0.51	374
1.0	0.69	0.87	0.77	553
micro avg	0.69	0.69	0.69	927
macro avg	0.68	0.64	0.64	927
weighted avg	0.68	0.69	0.66	927

Figure: Classification report for Naïve Bayes

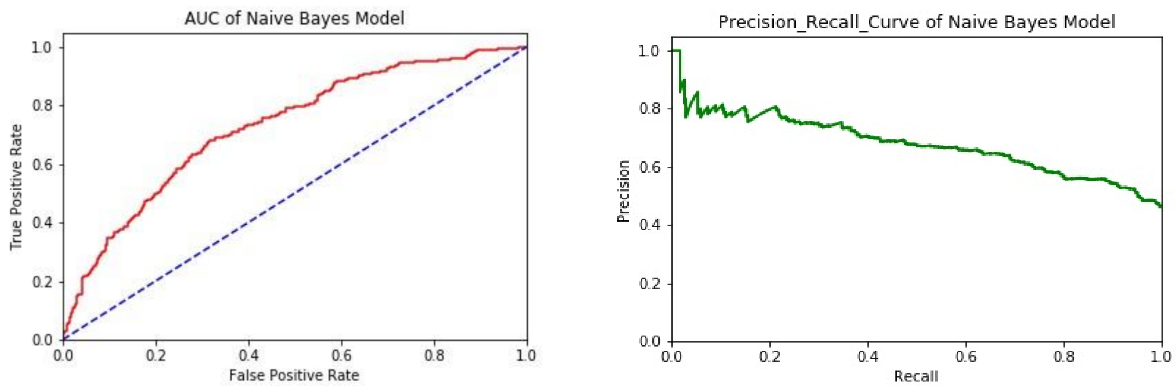


Figure: AUC Curve and Precision-Recall Graphs for Naïve Bayes

Supervised learning: Support Vector Machine (SVM) Model

	precision	recall	f1-score	support
0.0	0.66	0.55	0.60	426
1.0	0.66	0.75	0.71	501
micro avg	0.66	0.66	0.66	927
macro avg	0.66	0.65	0.65	927
weighted avg	0.66	0.66	0.66	927

Figure: Classification report for SVM

Topic Classification

Since classification is supervised learning, a training dataset is needed. We used an external dataset with labels that was obtained from an online source to train and test our models. The labels/topics in the dataset have 5 features consisting of food, service, price, anecdotes/miscellaneous and ambience as shown in the sample set below. The dataset was split into 70:30, and 70 percent was used to train, and 30 percent was used to test.

	A	B	C	D
1	review	label		
2	But the staff was so horrible to us.	service		
3	To be completely fair, the only redeeming factor was the f	food,anecdotes/miscellaneous		
4	The food is uniformly exceptional, with a very capable kitc	food		
5	Where Gabriela personaly greets you and recommends yo	service		
6	For those that go once and don't enjoy it, all I can say is th	anecdotes/miscellaneous		

Figure: Training dataset for topic classification

In our models, we utilized one-vs-rest strategy to implement Multi Label classification. We've used 3 models in the evaluation including Naive Bayes, Support vector machine, and K-nearest neighbor.

Multinomial Naïve Bayes (NB) Model

For the Naïve Bayes, several parameters were tuned to get the best performance from the model. We've tried using stopwords and without using stopwords and also several numbers of min_df in the tf-idf vectorization process.

Also, different alpha values (Laplace smoothing value) were tuned in the Naïve Bayes model. The performance is reported as shown below.

	precision	recall	f1-score	support
ambience	0.95	0.30	0.45	121
anecdotes/miscellaneous	0.78	0.56	0.65	289
food	0.86	0.70	0.77	301
price	0.97	0.41	0.58	80
service	0.88	0.56	0.68	145
micro avg	0.85	0.56	0.67	936
macro avg	0.89	0.50	0.63	936
weighted avg	0.86	0.56	0.66	936
samples avg	0.62	0.59	0.60	936

Figure: Classification report for Naïve Bayes

Support Vector Machine (SVM) Model

For the Support vector machine, several parameters were tuned to get the best performance from the model. Using stopwords and without using stopwords were implemented. Several numbers of min_df and max_df in the tf-idf vectorization process were also implemented.

In the support vector machine model, linear kernel was used. After trying several numbers of C values, the performance of the model is reported as shown below. Support vector machine produced the best performance of 76% compared to the other two models that were evaluated.

	precision	recall	f1-score	support
ambience	0.83	0.56	0.67	121
anecdotes/miscellaneous	0.77	0.69	0.73	289
food	0.86	0.76	0.81	301
price	0.89	0.79	0.83	80
service	0.84	0.74	0.79	145
micro avg	0.83	0.71	0.77	936
macro avg	0.84	0.71	0.77	936
weighted avg	0.83	0.71	0.76	936
samples avg	0.72	0.71	0.70	936

Figure: Classification report for SVM

K- Nearest Neighbor (KNN)Model

In the K-nearest neighbour model, several parameters were tuned to get the best performance from the model. Using stopwords and without using stopwords were also implemented just as previous models, as well as several numbers of min_df and max_df. Several numbers of K values were tuned. The performance of KNN is reported as shown below.

	precision	recall	f1-score	support
ambience	0.61	0.31	0.41	121
anecdotes/miscellaneous	0.71	0.68	0.70	289
food	0.77	0.68	0.72	301
price	0.78	0.44	0.56	80
service	0.79	0.48	0.60	145
micro avg	0.74	0.58	0.65	936
macro avg	0.73	0.52	0.60	936
weighted avg	0.73	0.58	0.64	936
samples avg	0.64	0.62	0.61	936

Figure: Classification report for KNN

In conclusion, after trying those three models mentioned above, the support vector machine model gives the best performance, and it is selected to use in our application described in other section.

Unsupervised learning: VADER Model

For unsupervised sentiment analysis we chose the valence-based approach offered by Vader. Vader analyzes a piece of text to see if any of the words in the text is present in its lexicon. Each word is assigned a numeric value based on its positive or negative sentiment. In valence-based approach, the intensity of the word is also taken into consideration along with the positive/ negative sentiment. For instance, good, better and best will have a positive numeric value in increasing order.

Positive, negative and neutral score is calculated for each text review, to represent the proportion of the text that falls into those categories. The final metric, the compound score, is then calculated from the positive, negative and neutral scores. The lexicon ratings for all words have been standardized in such a way that the compound score lies between -1 and 1. A compound score in the range of -1 to -0.5 would classify a review as negative, -0.5 to 0.5 as neutral, and 0.5 to 1 as positive.

Pie-chart below shows the output of sentiment analysis performed on the entire set of reviews for all cafes in Manhattan.

```
 This coffee shop is so beautiful. I would definitely get back.
Overall sentiment dictionary is :  {'neg': 0.0, 'neu': 0.501, 'pos': 0.499, 'compound': 0.8383}
sentence was rated as  0.0 % Negative
sentence was rated as  50.1 % Neutral
sentence was rated as  49.9 % P
```

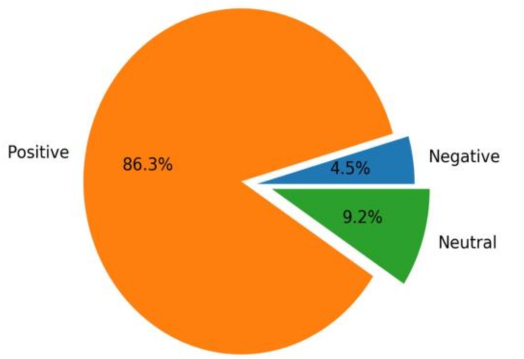


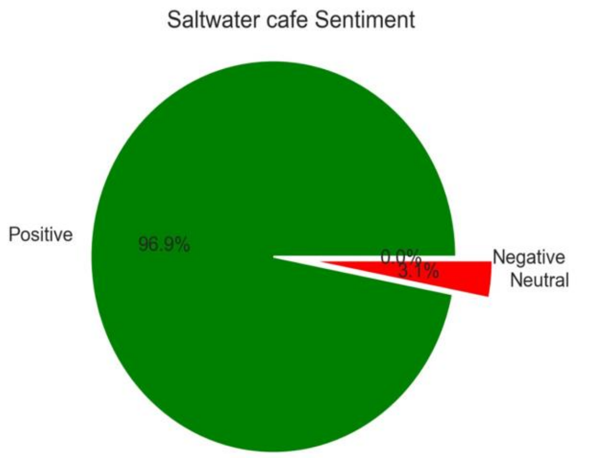
Figure: VADER unsupervised model output for sentiment analysis

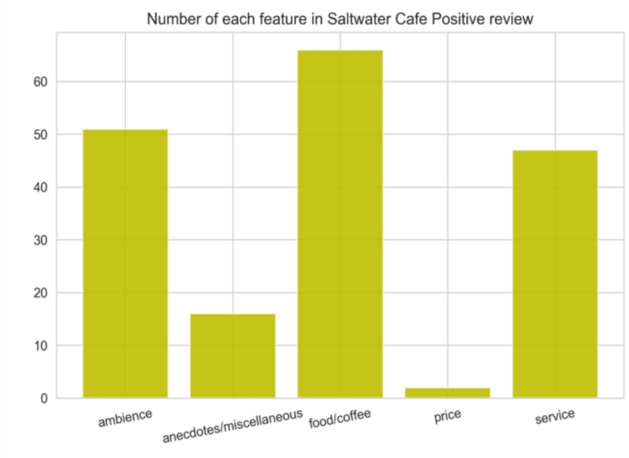
Application of Topic classification (SVM) and Sentiment analysis (Vader)

We applied our SVM and Vader models to analyse reviews for specific cafes to derive business insights from customer reviews.

Saltwater Coffee: Overall rating 5

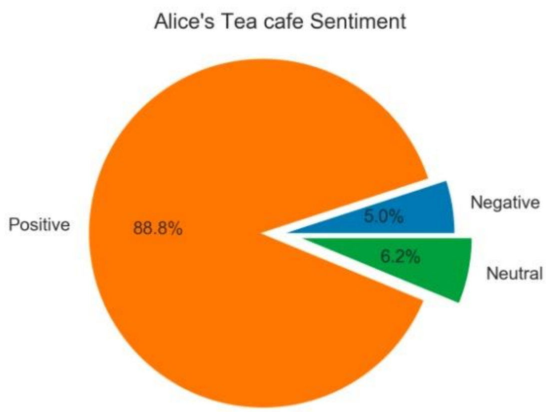
Sentiment Analysis shows that the overall sentiment of the customers was positive (97%) with only 3% neutral reviews. There were no negative reviews for the restaurant. The topic classification on the positive reviews show that customers discussed food/coffee, ambience and service the most in their reviews. The price was rarely discussed. This could be an opportunity for the business to increase the price of the items by a slight margin, without affecting customer feedback and in turn increasing profit.

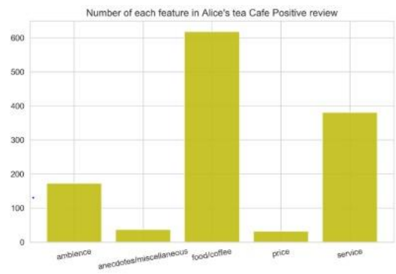


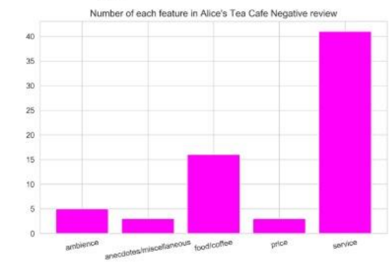


Alice's Tea Cup Chapter II: Overall rating 3.5

Applying the SVM and Vader models to reviews for Alice's Tea Cup revealed different insights. Sentiment Analysis shows that the sentiment was distributed as follows: positive (89%), negative (5%) and neutral (6%). Further analysis of the positive reviews shows that while customers are happy with the taste of food/coffee, there is scope to increase the service and ambience. Analysis of negative reviews show that customers complained a lot about service and then about the food/coffee. So, we can conclude that Alice's Tea Cup should first focus on improving their service, then some improvements on food/coffee to increase customer satisfaction and eventually their rating on Yelp.







Conclusion

In this project, we proposed an innovative method for identifying different features for cafes in Manhattan. Through text mining, sentiment analysis and topic classification, we are able to uncover that the cafe customers' satisfaction levels are determined by taste of food/coffee, service and ambience. Interestingly enough, the price of the coffee/ food is not highly discussed within cafe reviews in Manhattan. This may be attributed to the expectations a customer has when approaching a cafe. Therefore, cafes in Manhattan should emphasize a welcoming, aesthetically pleasing coffee shop with great tasting coffee and food. For our dataset, Naive Bayes proved to be a better method (F1 score 0.69) for sentiment analysis and SVM gave better results (F1 score 0.77) for topic classification.

Future Scope

On the other hand, similar procedures can be used to improve the accuracy of our models using neural networks. Secondly, our models do not handle the indirect and sarcastic comments so studying them in detail in the reviews would give us more specific insights. Also, studying the 3-star rating reviews would make our model more accurate, as they can be the most honest reviews with accurate features. Further to implement more analysis to help cafes improve the menu based on most popular customer choices to improve their finances and serve the customer better.

References

[1] Boya Yu, Jiaxu Zhou, Yi Zhang, Yunong Cao,Center for Urban Science & Progress, New York University "Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews"

[2] Lindsey Wright, Undergraduate Student Honors, East Tennessee State University "Classifying textual fast food restaurant reviews restaurant reviews quantitatively using text mining and supervised machine learning algorithms"

[3] Source:
http://metashare.ilsp.gr:8080/login/?next=/repository/download/479d18c0625011e38685842b2b6a04d72cb57ba6c07743b9879d1a04e72185b8/