Yelp Sentiment Analysis

Final project for Data 765 Introduction to Computational Social Science at Queens College

By Adriana Sham

Abstract

Yelp is an application to provide the platform for customers to write reviews and provide a star-rating. This project uses Support Vector Machine and Natural Language Processing to understand the sentiment of the given reviews from the dataset. Identifying the most significant words of the reviews and run our data in algorithm that feature select and ship a model that is useful toward the goal of gaining insights of a good restaurant or a restaurant that needs improvement, in order to have higher star rating.

Introduction

Sentiment analysis is taking sentences, documents and running it in a machine learning algorithm such as natural language processing (NLP) and using the translated data from NPL; feeding the data into a sentiment analysis classifier, in this case, Support Vector Classifier, in order to answer real world questions. In this project, reviews from yelp users are used and extracting the user's feeling about the restaurant.

The main purpose of this project is mainly analyzing customers' reviews and figure out reasons why customers like or dislike the restaurant.

Analytic Approach | Methods and Data

Data

The data is a raw download from Kaggle.com/data/yelp which is a subset of the yelp dataset; to obtain a complete yelp dataset you can refer to yelp.com/dataset. The Kaggle data set and the yelp dataset are subsets of yelps businesses, reviews, and user data for personal, educational, and academic purposes. The whole dataset contains:

- 8,021,122 reviews
- 209,393 businesses
- 200,000 pictures
- 10 metropolitan areas
- 1,320,761 tips
- 1,968,703 users
- 1.4 million business attributes like hours, parking, availability, and ambience.

This project focuses on the average star the restaurant is rated and the text (review) and star given by the user, therefore it only uses two table business and review table. In total, there are 5,261,668 user reviews, information on 209,393 business. We will focus on two tables which are business and review table.

Attributes of business table are as following:

- business_id: string, ID of the business
- idname: string, name of the business
- address: string, the full address of the business
- city: string, the city
- state: string, 2 character state code
- postal_code: string, the postal code
- latitude: float, latitude of the business
- longitude: float, longitude of the business
- stars: float, average rating of the business
- review_count: integer, number of reviews received
- is_open: integer, 1 if the business is open, 0 otherwise

- attributes: object, business attributes to values.
- categories: array of strings, multiple categories of the business
- hours: object, hours which business is open

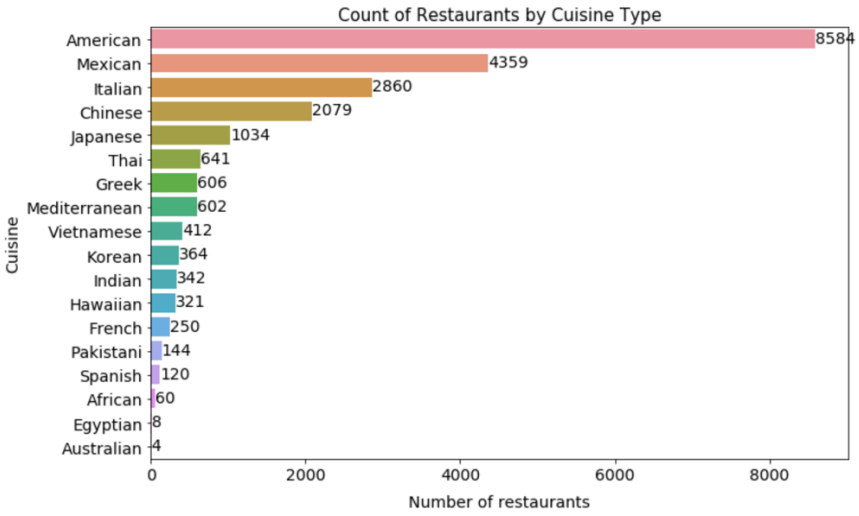
Attributes of review table are as following:

- review_id: string, ID of the review
- user_id: string, ID of the user
- business_id: string, ID of the business
- stars: integer, ratings of the business
- date: string, review date
- text: string, review from the user
- useful: integer, number of users who vote a review as useful
- funny: integer, number of users who vote a review as funny
- cool: integer, number of users who vote a review as cool

The population of interest is the entire restaurant business in the United States. However, this dataset partly represents the population of interest because this is a subset of the yelp dataset from Kaggle and some restaurants are not registered in the yelp system. Moreover, certain restaurants have more reviews than other restaurants, in other words some restaurants have larger observations than other which can perform better in the model.

Exploratory Statistics

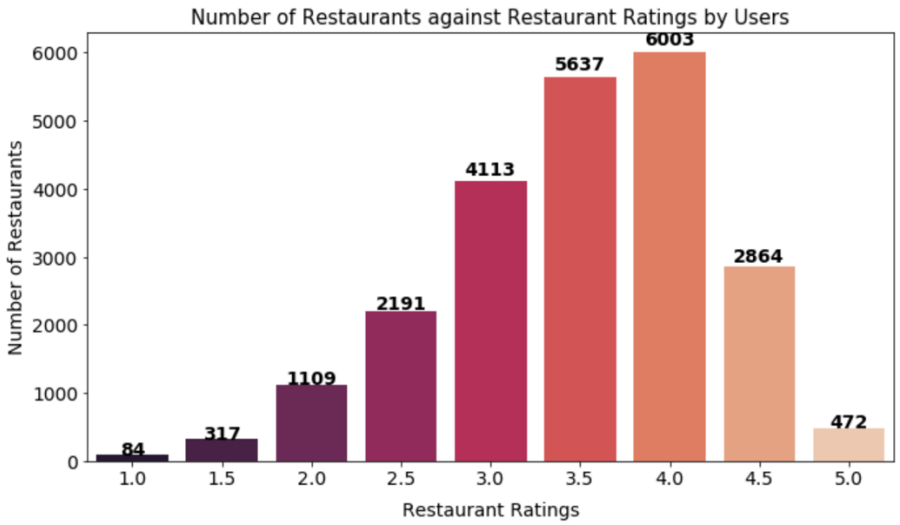
Before jumping into our machine learning, we explored and familiarize ourselves with these datasets through graphs. For given dataset, we tried to find patterns and potential problems which could be useful information in the machine learning process. One main area we explored on are restaurants categories, and star ratings' distribution across the United States.

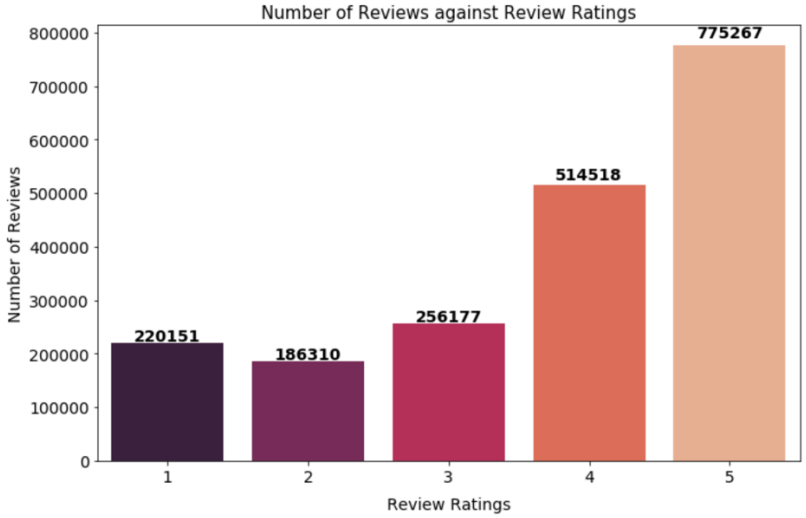


Each color indicates a type of cuisine. Grouped by the cuisine type and the number of restaurants is there by cuisine. The number of restaurants in each for each cuisine is shown at the end of each

color bar. For example, the majority of the restaurants are American. In fact, we noticed that most of the restaurants are American, in comparison to other cuisines.

In order to comprehend key insights of what makes a good restaurant, it is necessary to have enough data point in this case restaurants and also reviews. Therefore, we picked top 5 most common restaurant cuisines in the US (Chinese, Japanese, American, Italian, Mexican) to make further analysis.





Above two graphs, the first one is number of restaurants against restaurant ratings by users, and the second graph is number of reviews against review ratings. We noticed most restaurants are most likely have good reviews.

To explore more in into good restaurants and restaurant that may need more improvements, we will discuss more in the natural language processing section. Presenting a more precise model and analysis on key words about restaurants.

Sentiment Analysis

In order to identify restaurants features, sentiment analysis is used. The overall rating for each restaurant in Yelp is only useful to convey the general experience, but there is not enough information to give a judgement whether the restaurant's service, food quality and environment is good or not. We are going to run Support Vector Machine, a machine learning model, in order to further understand the restaurant essential features.

Text pre-processing

Before performing machine learning, we need to use natural language processing to preprocess and clean text data. Converting the text into vector format by splitting a review into individual words and returning a word list.

1. To clean the text, we first convert all words into lower case and taking out all kind of punctuation. Then, the function get_data is used to extract the rows from a specific cuisine and making a new table on its own.
2. filter_data is used to filter out all irrelevant words that are not useful, utilizing the NLTK library to remove stop words such as "if," "but," "or" and so on. The function also filters out all words that are not in the list of positive and negative words, leaving only positive and negative words on the table. The reason using positive and negative words only is that words with very rare appearance in the text are not useful for model improvement and it only adds noise to our model.

Machine Learning Model

Because we would like to know the importance of words on creating positive or negative reviews in a restaurant. Therefore, we consider SVM is an efficient model to perform text classification. The Linear SVM creates a hyperplane by using supported vectors to maximize the distance between the two classes, in this case, positive and negative.

Previously in text processing, we converted each review into a word list, we use support vector machine to differentiate positive and negative words in the text, utilizing the review ratings which converted into the model's label. (If the stars ≥ 3, the review is 'positive' and if stars < 3, the review is 'negative').

A data frame is created to mainly store three columns,

1. the score (weigh) of each word from the user yelp review which are calculated by the SVM model.
2. The frequency of each word in the entire user review of a specific category

3. the polarity score which is a value that is calculated in order to evaluate the sentiment of the word; the score of each word was first multiplied by its frequency, then it normalizes the score by dividing each word count by the total number of reviews this word appears in for the specific cuisine.

The weights obtained from svm.coef_ indicates how satisfied or discontented the user is and the polarity score we calculated shows how much a word contributes to the score of all restaurants of a specific cuisine.

For example, the score of American restaurants is lowered by 0.5240 on average due to 'overpriced' while 'cold' is lowered by 0.2651on average. Then we might have an assumption that 'overpriced' displeased customers a lot more than 'cold', thus 'overpriced' is a more negative characteristic of American restaurants.

In order to find specific words that were used to indicate customers' concerns for the restaurant adjectives that only describes the polarity of sentiment such as "well", "good", "sorry" were neglected. Then, the top positive and negative words are extracted learning key features of each cuisine.

Results

Positive words

	1	2	3	4	5	6	7	8	9	10
American	delicious	friendly	recommend	fresh	fun	incredible	perfection	outstanding	fast	attentive
Mexican	delicious	fresh	recommend	authentic	fast	clean	incredible	bomb	outstanding	friendly
Italian	delicious	fresh	recommend	authentic	perfection	incredible	reasonable	outstanding	friendly	fabulous
Japanese	delicious	fresh	friendly	recommend	authentic	hot	reasonable	incredible	clean	fast
Chinese	delicious	fresh	friendly	recommend	authentic	hot	reasonable	incredible	clean	fast

Extracting key positive words from user reviews, we discovered that for cuisines, 'delicious' ranks first and 'fresh' second among all positive words which tells us that customers value more the flavor and freshness of the food than other aspects of the restaurants, such as service, ambience or price.

Different characteristics are also shown for different restaurant categories. American restaurants have positive reviews mainly for their 'friendly' and 'attentive' service, both are in the top ten positive words. Similar for the Mexican, Japanese and Chinese restaurants, service is important since 'friendly' and 'fast' service have a high ranking.

We observed that the key features for Mexican and Italian restaurants are 'delicious', 'fresh' and 'authentic' food which means customers have high expectations on the food that the restaurant provides. Japanese and Chinese have positive reviews not only by their food, but also mainly for their friendly service, especially for Korean restaurants, since attentive ranks third.

Negative words

	1	2	3	4	5	6	7	8	9	10
American	worst	bland	slow	rude	hard	cold	awful	overpriced	poor	mediocre
Mexican	bland	worst	rude	mediocre	cold	slow	poor	overpriced	awful	disgusting
Italian	bland	worst	rude	mediocre	cold	slow	poor	overpriced	awful	expensive
Japanese	bland	worst	rude	mediocre	cold	dirty	poor	overpriced	awful	disgusting
Chinese	bland	worst	rude	mediocre	cold	dirty	poor	overpriced	awful	disgusting

From the negative word list, we could observe that bland is one of the main problems for the cuisines which means customers expect food to be spicy. And It is likely to have the low score because the food is cold and overpricing. Slow service is the main negative characteristic for American, Mexican and Italian restaurants. Rude is one of the main problems for American cuisine. The low score of Japanese and Chinese restaurants are due to the dirty environment.

This analysis may help to extract specific features from set of reviews. Therefore, restaurant owners can extract information to improve their business with the received Yelp reviews. The obtained reviews may help the business growth and having an insight of why customers like or dislike their restaurants, such as, obtaining great reviews are primarily due to fresh food, or perhaps unsatisfied reviews are caused by overpricing.

Discussion and Conclusion

In this project, our goal is to help restaurant owners realize the effect on good/bad Yelp reviews. Owners can have a better overview about features they are good at and needs to improve on for better star rating such as pricing, food, ambiance and services quality.

Reference:

1. https://medium.com/tensorist/classifying-yelp-reviews-using-nltk-and-scikit-learn-c58e71e962d9
2. https://www.yelp.com/dataset/documentation/main
3. https://github.com/shekhargulati/sentiment-analysis-python
4. https://towardsdatascience.com/fine-grained-sentiment-analysis-in-python-part-1-2697bb111ed4
5. https://realpython.com/sentiment-analysis-python/#using-machine-learning-classifiers-to-predict-sentiment
6. https://monkeylearn.com/sentiment-analysis/
7. https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76
8. https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089