Project: Restaurant Review Rating

Long Ly (1000866434)

Objective:

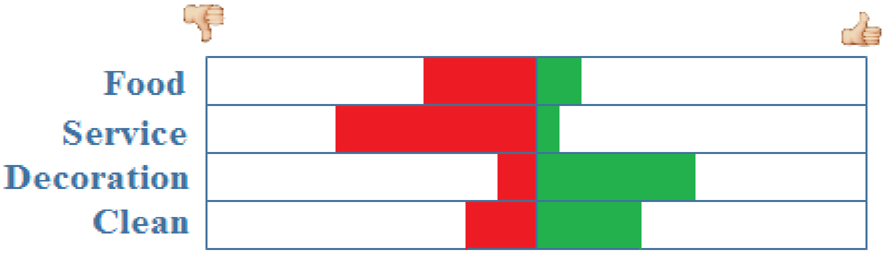
Yelp rating on reviews are sometimes very ambiguous. A 4-star review does not mean the experience of the customer is positive about all of the aspects of the restaurant. For example, the food may be delicious, but the service is really slow and unsatisfied. On the other hand, a 2-star review does not also contain all negative information about food or service at all; it may discuss about the cleanliness or the decoration of the restaurant that make the user did not like about. Therefore, simple rating restaurants based on the average number of star they have received does not reflect the real opinion of the customers about various aspects of the restaurants. In addition, if customers want to learn more details about the restaurant, they have to spend a lot of time reading through all reviews, and some of the reviews may not be totally relevant to the restaurant. Thus, the objective of this project is to provide a more accurate rating system based on users' sentiment in their reviews.

Task:

The business data from Yelp contains different categories, but only businesses that have 'Restaurant' category will be used for the project. In addition, any reviews which is not relevant to the 'Restaurant' category as well as does not have any useful vote will be omitted from the data set. Each review will be cleaned by removing all special character and all text will be changed into lowercase. However, the stop words will not be removed from the reviews and the text will not be stemmed since they may contain important information about users' sentiment. The clean data will be used to train a classifier to identify the category of each review: Food, Service, Cleanliness, and Decoration. The idea is adapted from "Classifying Yelp reviews into relevant categories" of Vaibhav Saini, 2014.

Deliverable:

The project will be delivered as a website similar to Yelp but it will contain features that demonstrate this project. The user will be able to search for the restaurant by name, keyword, or location using the data provided from Yelp. The list of restaurant will be shown and users need to choose the restaurant that they wish to view the rating. The bar chart look like the figure below will show on the screen.



(adapted from "Classifying Yelp reviews into relevant categories" of Vaibhav Saini, 2014)

The red color on the left side represents the percentage of negative rating, and the green color on the right side represents the percentage of positive rating for each category based on users' sentiment. The sample chart above shows that the service of the restaurant is really bad while its decoration is very attractive. By using this chart, Yelp users can easily expect what their experience will be when they are visiting the restaurant.

Challenge:

Each review does not necessarily fall into one particular category. In fact, one review may cover various categories about the restaurant. Therefore, the challenge of the project is to find a way to classify the reviews into multiple categories accurately.

Method and Algorithm:

First, I will tokenize all the words in the review data, and calculate the frequency of each token in the entire corpus. Next, I will calculate the frequency of each token in the entire corpus to determine the most common words and put them into a list. Each word will be categorized manually into Food, Service, Decoration, and Cleanliness. In addition, I will determine the sentiment value of each word (if possible). Using the Naïve Bayesian Classification, I will train the data into 4 different classifiers for each category: Food, Service, Decoration, and Cleanliness. If a review contains multiple sentiment value for the same category, only the highest absolute sentiment value will be taken into calculating the review rating for a particular restaurant.

The rating (both positive and negative) of the restaurant will be using the below formula and represent as bar chart.

$$possitive\ or\ negative\ rating = \frac{|\sum sentiment\ values\,|}{2\ x\ number\ of\ ratings}$$

Initial implementation:

GitHub repository contains the source code to clean the business data as well as the review data from Yelp. Since the output of these data are large, they will be hosted in different location (Google Drive). Among 452413 reviews, I will randomly choose 1000 reviews to build the classifiers for this project.

In addition, the dictionary using to find sentiment value will be "Subjectivity Lexicon" dictionary that has 8222 words (Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005). Even though there are many information in the dictionary to determine the sentiment value, I will only use "Type" and "Prior Polarity" to determine the sentiment value for this project.

Type	Prior Polarity	Sentiment Value
Strong Subjectivity	Positive	+2
Weak Subjectivity	Positive	+1
Strong/Weak Subjectivity	Neutral	0
Weak Subjectivity	Negative	-1
Strong Subjectivity	Negative	-2

Evaluation Plan:

As suggested from “Classifying Yelp reviews into relevant categories” of Vaibhav Saini (2014), I will use Precision and Recall to measure the performance of a classifier. Precision is the percentage of the number of true categories in total prediction while recall is the percentage of reviews that were predicted in their true categories.

Change of Plan:

There are a lot of changes in the project since the proposal. Initially, the proposal was to design a method to extract negative information about the restaurant. However, after carefully researched, the proposal is not feasible since it would require a lot of research in sentiment analysis that will not meet the required time of this project. Therefore, from extracting negative information, the project now is designing a better rating system using negative information from users’ review.

Difficulties Encountered:

I spent a lot of time finding more information about sentiment analysis for each sentence of the review. Since users sometimes discuss various topics about the restaurant, it seems impossible to find a method to extract the correct negative information about the restaurant. Therefore, I decided to slightly change the objective of the project in order to make it deliverable at the end of the semester. In addition, since the project falls into multi-label classification, I also spent a lot of times finding the best method to classify the reviews into multiple categories. However, after discovering that there is no correlation between each category, I’m able to build multiple classifiers for all categories. Finally, another difficulty I faced was to get enough time working on this project. Since three out of four classes I’m taking have semester-long projects, I still cannot find enough time to fully focus on this project yet.

Future Implementation (what left to be done & future plan):

Currently, I have cleaned all the data, but still have not fully implemented the method I mention earlier. Therefore, my next step would be obtain the list of the most common words and categorized them manually. After that, I will follow the method I describe to complete the project. My plan is to finish the project in the next 3 weeks and spend my last week to build the website that contains the feature of this project.

Expected Challenge:

Since I’m doing the project by myself, I will need a lot of time categorizing the word list. In addition, I have never done building a website before, so I may face a lot of problem building it after completing the data processing.

References

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proc. of HLT-EMNLP-2005.*

Saini, Vaibhav, et al. Classifying Yelp reviews into relevant categories. University of California at Irvine, 2014.

Data:

Review Data (clean):
https://drive.google.com/open?id=0Bwry0GwTJUEzT3ZtMTd0cnhSMzA&authuser=0

Restaurant Data (clean):

https://drive.google.com/open?id=0Bwry0GwTJUEzT1E5eXlXYnVWMDg&authuser=0