What matters: words that lead to positive/negative reviews

Huanci Wang, Shuxuan Liu

New York University

Introduction

According to Investopia, a comparative advantage gives a company the ability to sell goods and services at a lower price than its competitors and realize stronger sales margins. Through an investigation of Yelp's review data set, we hope to find the comparative advantages and disadvantages through numerous consumers' reviews. By classifying the positive and negative reviews, we find the words that contribute the most to positive and negative reviews. Consequently, we are able to learn the aspects of a business that are important to customers.

Problem definition and Algorithm

1. Task

Using techniques in NLP, we find which nouns and adjectives have the most weights in a positive/negative reviews, thus finding the comparative advantages and disadvantages of businesses.

Similar to the input in Perceptron algorithm, the input X is an array of word matrices of only the nouns and adjectives in each review, using 1s and 0s in specific indices to represent whether each word in the word dictionary is contained in the review. The input y is 0s and 1s indicating whether the review is positive or negative. In the data, we classify reviews with at least three stars rating as positive reviews and others as negative reviews.

The output of this project is the weights of each word in determining whether the review is positive or negative. We then select the words that have the most weights and analyze their business value.

2. Algorithm

SVM for classification: In the project, we applied SVM to differentiate between positive and negative reviews. Our hypothesis was that adjectives and nouns about food, service and environment would contribute most in classifying reviews. We tested different kinds of kernels by sklearn and found linear kernel to yield the best validation reuslt as well as testing result.

Initialize: Choose $\boldsymbol{w}_1 = 0, t = 0$.
1. For iter $= 1, 2, \cdots, 20$
2. For $j = 1, 2, \cdots, |\text{data}|$
3. $t = t + 1; \quad \eta_t = \frac{1}{t\lambda}$
5. If $y_j(\boldsymbol{w}_t \cdot \boldsymbol{x}_j) < 1$
6. $\boldsymbol{w}_{t+1} = (1 - \eta_t\lambda)\boldsymbol{w}_t + \eta_t y_j \boldsymbol{x}_j$
7. Else
8. $\boldsymbol{w}_{t+1} = (1 - \eta_t\lambda)\boldsymbol{w}_t$
8. Output: $\boldsymbol{w}_{t+1}$

Fig. 1: SVM algorithm

Perceptron: perceptron algorithm does not yield a high testing score.

Decision Tree: Since all of our data are dummy variables, we believe this will lead to over-fitting so we did not choose decision tree.

Experimental Evaluation

1. Data

We use the 'review.json' dataset. The original dataset is too big (5GB), so we randomly chose a subset of it (51MB) for the project.

We first deleted from our data reviews that are marked as 'not useful' (useful < 3) since those review should contain little information. Then, we applied lemmatization in NLP to clean the data and ensure we don't use the same words but different forms as separated features. We then applied pos_tag in NLTK to choose only nouns and adjectives in our word list since we consider them richest in business information and do not want meaningless interference from words such as 'and', 'like','was'.

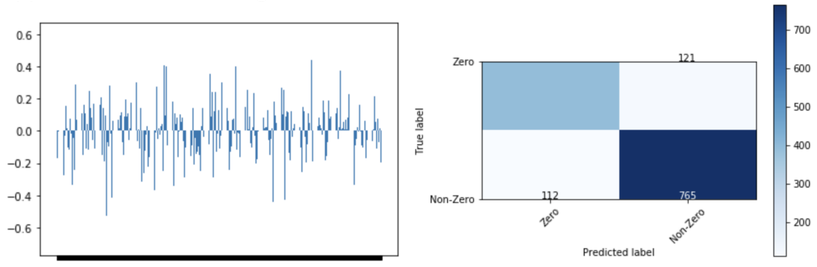
2. Methodology

Initially, we applied the Perceptron algorithm to learn the words that classify the positive and negative reviews. As an alternative, we also tried SVM with different kernels as a classifier. We used K-fold cross validation to test our model's performance.

In Perceptron and SVM with kernels other than the linear kernel, we were only able to achieve a 80% accuracy on the training data. Therefore, we chose SVM with linear kernel as it was able to achieve a 100% accuracy on the training data and 88% accuracy on the validation data set.

3. results

We have a total of 2453 words. The graph below on the left demonstrates the weights SVM outputs for each word. We also tested the weights on our testing data set. The confusion matrix is given below on the right. We found that words 'comfortable', 'fresh', 'friendly', 'class', 'efficient', 'fun', 'price', 'reasonable', 'size', 'family', 'new' contributes most to positive reviews and words 'rude', 'trash', 'unprofessional', 'old', 'bland', 'hassle', 'yelling', 'soggy' appear most often in negative reviews.



4.discussion

We get plenty business insights from our result. For instance, 'fresh' gets a high positive weight of 0.4795, indicating that customers value fresh ingredients which is seemingly hard to notice for customers. This could be a good reminder for restaurant managers. Among the negative words, many of then such as 'unprofessional', 'rude', 'hassle', 'seated' and 'yelling' indicate bad services which shows customers greatly emphasize service quality. Moreover, 'new' and 'old' appear in positive and negative reviews respectively, which could be an indicator of people's preference on novelty. This could be further assured by doing phrase analysis.

Conclusions

In this project, we extracted from the reviews what aspects of a business are most important to customers. From the outcome, we learned that fresh ingredients (food), good services and novelty are key factors to positive customers' reviews. Therefore, businesses can place more emphasis on food ,service and novelty to raise their comparative advantages over other competitors.

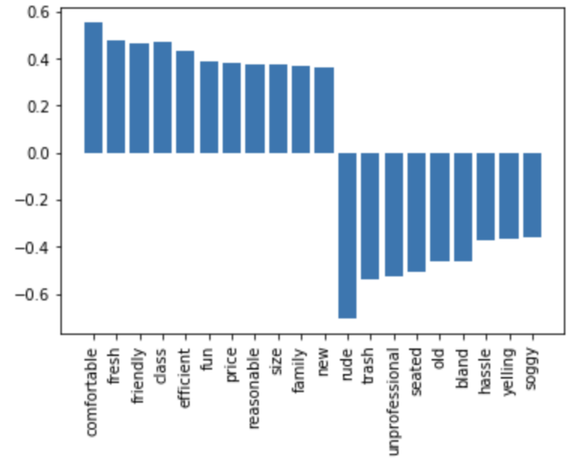


Fig. 2: Meaningful positive/negative words

Possible next steps:

Do bigram analysis (phrases analysis) especially on adjectives+noun phrases to see the more accurate aspects that attract/repel customers.

Learn what words contribute most to an "useful" review. This will further complete our study of important comparative advantages of a restaurant for customers.

The review data we have for each restaurant is relatively limited therefore we cannot learn the comparative advantages and disadvantages for specific restaurants. However, attempting to do this in the future will add great business value to our study as well as the restaurants' performances.

Bibliography

"Improving Restaurants by Extracting Subtopics from Yelp Reviews." James Huang, Stephanie Rogers, Eunkwang Joo. University of California, Berkeley.

"Clustered Layout Word Cloud for User Generated Review." Ji Wang, Jian Zhao, Sheng Guo, Chris North. Virginia Tech and University of Toronto.