Sentiment Analysis of Five Best Burger Chains in U.S.

Jiayi Shen, Junxia Zhao, Sixu Li

December 1, 2020

1 Introduction

Yelp is an Internet company founded in 2004 to "help people find great local businesses" by providing a platform for users to write reviews of businesses. From about 1 million overall reviews and ratings in Yelp, we intended to use a machine learning based method to abstract specific information, and to propose data-driven, actionable suggestions to some business owners based on these mined information in order to improve the quality of their food or service. To be more specific, in this project, focusing on five famous fast-food chains—McDonald's, Burger King, Five Guys, Wendy's, and Shake Shack, we used regularized logistic regression classifier to decipher the sentiment tendency of each review by differentiating positive and negative words in reviews, and further give each word a score to understand how positive or negative the words are. Afterwards, we can give those business owners some suggestions on providing better service.

2 Preliminaries

In this section, we are going to introduce the important statistics and model used in our project.

2.1 Information retrieval statistics

In information retrieval, *tf-idf*, short for *term frequency-inverse document frequency*, is a numerical statistic that is intended to measure how important a word is to a document of corpus.[1] For example, to one novel in a collection of novels or to one website in a collection of websites. And the tf-idf is calculated by [2],

$$\mathrm{tfidf}(t, d) = \mathrm{tf}(t, d) \log \frac{N}{|\{d \in D : t \in d\}|} \tag{1}$$

where,

- $\mathrm{tfidf}(t, d)$ is the value of statistic tf-idf for term t in document d;
- $\mathrm{tf}(t, d)$ is the frequency of term t in document d;
- N is the total number of documents;
- $|\{d \in D : t \in d\}|$ is the number of documents containing t.

When measuring the importance of words, instead of only using term frequency, i.e., $\mathrm{tf}(t, d)$, tf-idf adds the $\log \frac{N}{|\{d \in D : t \in d\}|}$ term to decrease the weights for not important but commonly used words and give prominence to the words that are characteristic but not appearing very often in a collection of documents. For example, in the reviews of burger businesses, words "burger" or "coke" would occur many times but providing limited information. In contrast, "mac" is a special word for "McDonald's" but not for other burger business. Therefore, the value of statistic tf-idf for "mac" would larger than words "burger" or "coke".

[1] This definition is given on Wikipedia `https://en.wikipedia.org/wiki/Tf-idf`.

[2] The notations are borrowed from [1].

2.2 Regularized logistic regression

Taking the interpretability into consideration, we decided to use logistic regression classifier. Also, to make the model much simpler and meanwhile deal with multi-collinearity (In the reviews, there exists some words with similar meaning, like "perfect", "nice" and "great". We think this might cause multi-collinearity among the variables), we further chose regularized logistic regression model to fit the data. And the objective function can be written as,

$$\min_{\beta \in \mathbb{R}^n} \quad -\sum_{i=1}^{N}[-\ln(1+e^{\beta^T x_i}) + y_i \beta^T x_i] + \lambda \|\beta\|_1 \tag{2}$$

where [3],

- λ is a hyper-parameter for penalized term;
- β is the vector we want to estimate, indicating the coefficient for each word;
- N is the number of samples (reviews);
- y_i is either 1 or 0, representing whether a review is positive or negative;
- x_i is the feature vector of each review.

By solving this minimization problem, we could obtain the coefficient β_i for each word, which could be used for further statistical inference.

2.3 Polarity score

The same as [1], instead of directly using the coefficients in the logistic regression, we chose to use *polarity score* to measure how much a word contributing to the ratings of reviews. And the polarity score can be calculated as:

$$\text{polarity\_score}(t, c) = \text{coefficient}(t) \times \frac{\text{total\_frequency}(t, c)}{\text{number\_of\_reviews}(c)} \tag{3}$$

- polarity_score(t, c) is the index for measuring how essential word t is among restaurants of type c;
- coefficient(t) is the word coefficient calculated from the regularized logistic regression model;
- total_frequency(t, c) is the total frequency of word t in all reviews of type c restaurants;
- number_of_reviews(c) is the total number of reviews of type c restaurants.

3 Data preprocessing

In this section, we would talk about how we processed the raw data to make them suitable for classification tasks by steps:

(i). **Selecting data:** We first merged the business and reviews data sets by the attribute "business id" and then selected out the data of the five famous fast-food chains using attribute "name". And we finally obtained 6083 pieces of data.

(ii). **Preprocessing reviews:** We first tokenized each review into "bag of words" (unigram). And then we stemmed, removed all the numbers, punctuations and whitespace and filtered out the common stop words in the bag of words. By visualizing the importance of words using statistic tf-idf, we found out some meaningless but recurring words such as "mcdonald", "king" etc. So, we created our own stop words set and also deleted them in the bag of words. And finally, we embedded those left words into a word matrix and weighted each word with its' corresponding tf-idf.

[3] Used the same architecture in [1].

(iii). **Binarizing labels:** We labeled reviews with stars greater or equal to 3 as "positive" while the rest as "negative". And this decision was made based on the fact that the mean of ratings for those five burgers business is about 2.18.

(iv). **Splitting data:** For the classification tasks, we randomly separated the 6083 pieces of data into training and test set with ratio 8 : 2.

4 Experiment results

We first did 10-folds cross-validation on the training set to find the best parameter $\lambda = 0.0037$, and then fitted the final model with this λ again on the training set. We used the test set for prediction, and the final accuracy turns out to be 84.2%, which corresponds to our expectation. Next, we saved the coefficient of each word and used formula (3) to calculate their corresponding polarity scores. And we selected some words with high polarity scores shown in Table 1 and 2.

Chains	Top positive words						
McDonald's	fast	clean	fresh	frie	hot	quick	free
Burger King	fast	fresh	clean	frie	hot	quick	free
Five Guys	frie	fresh	fast	free	clean	flavor	cajun
Wendy's	fast	fresh	clean	frie	hot	quick	free
Shake Shack	shack	frie	fast	clean	fresh	mushroom	flavor

Table 1: Top polarity score by fast-food chains(positive)

Chains	Top negative words						
McDonald's	rude	cold	wrong	slow	sandwich	pounder	ketchup
Burger King	cold	rude	wrong	sandwich	slow	ketchup	min
Five Guys	wrong	meat	cold	bacon	ketchup	price	rude
Wendy's	cold	rude	wrong	sandwich	slow	min	ketchup
Shake Shack	cold	wrong	sandwich	price	meat	rude	bacon

Table 2: Top polarity score by fast-food chains(negative)

5 More trial (Bigram)

Except for splitting each single word, we also tried to tokenize the reviews into bigram and get the most frequently appeared phrase using statistic tf-idf. From those phrases shown in Fig. (1), we could better understand the result we got. For example, in Five Guys, we only know "cajun" and "peanut" are positive words, but from the frequent phrase table, we can see that the popular foods are actually "cajun fries" and "free peanut".

6 Suggestions

From the above results, we could give at least following suggestions:

(i). Note that, "fast", "clean", "fresh", "hot", "quick" and "free" are common positive words for fast-food chains while "slow", "wrong", "cold", "rude" are common negative words. It means that for a fast-food restaurant, clean environment, fresh and hot food, low price with some free food, right, fast and thoughtful service can always satisfy customers, and these things should always be considered when a fast-food restaurant owner wants to improve their services.

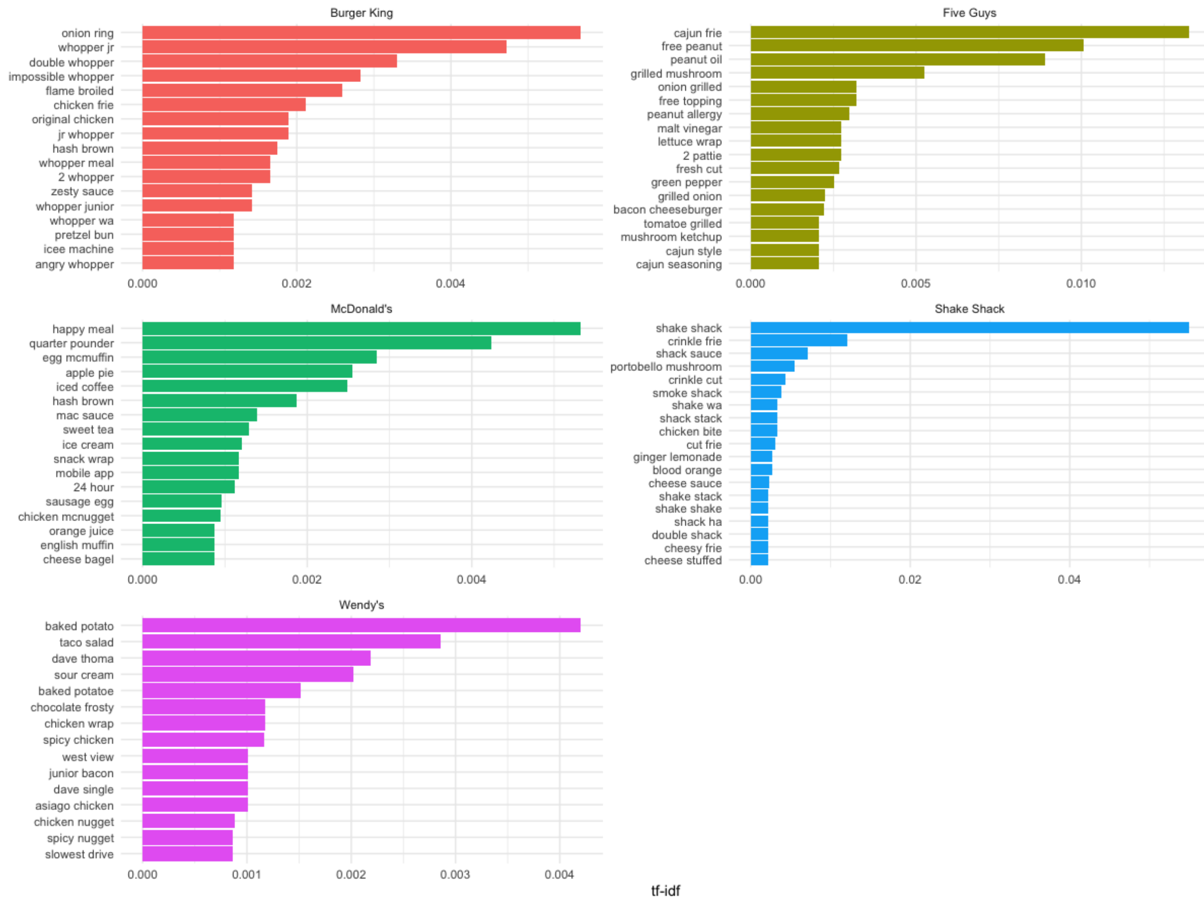

Figure 1: Phrases with large tf-idf

(ii). When it comes to speciality, Shake Shack and Five Guys are more competitive since the word "mushroom" (for product portobello mushroom in Shake Shack) and the word "cajun" (for product cajun fries in Five Guys) all get positive scores, while the word "pounder" (for product quarter pounder in McDonald's) get negative scores. So the owner of the two chains can maintain their speciality while the owner of the other three chains may consider how to ameliorate the speciality to attract more customers.

(iii). Notice that "flavor" is a positive word in Shake Shack and Five Guys, indicating that customers are satisfied with the flavor of these two chains. So the owner of the other three chains may try to improve the taste of the foods.

(iv). Comparing with other four chains, the word "price" has a lower score(negative) in Shake Shack, which means that the owner of Shake Shack may think of offering great buys on some specific food or lowering the entire price to improve the price-performance ratio.

7 Conclusion

In all, in this project, we applied some NLP techniques and machine learning methods on Yelp data set to mine some implicit information among the tons of reviews, which not only helps customers in fast-food chains to choose their favorite cuisine, but also help the business owners to find their advantages and shortages on services or products. Because we only used unigram tokenization in the classification task, we could only find some general words such as "fast", "clean" or "cold" for the fast-food industry. In the future, we could also try to embed bigrams into word matrix for classification which, to our expectations, would provide more special information for each individual business.

References

[1] Yu, B., Zhou, J., Zhang, Y., and Cao, Y. Identifying restaurant features via sentiment analysis on yelp reviews. *arXiv preprint arXiv:1709.08698* (2017).

A Contributions