Yelp Review Analysis for Steakhouses

Yike Wang Yuhan Zhou Zhongling Liao
wang2557@wisc.edu zhou453@wisc.edu zliao42@wisc.edu

I. Introduction

The steakhouse is a very popular type of restaurants in not only the United States, but also many other countries. The number of reviews for a steakhouse on Yelp is often decent and thus provides an objective evaluation. Our project aims to extract information from the steakhouse reviews, create models to specify how the evaluations influence the rating and give analytic advice to this industry and individual businesses. We firstly utilized several NLP models (LDA, LSA, NMF) to extract main topics from reviews and got the intuition of what customers care mostly. Then we developed a linear model to reveal the influence of every aspect on ratings. Based on these, we proposed advice to this industry on how to improve their performance and to serve customers better. Also, by comparing individual businesses performances on the specified aspects, we give specific advice to each steakhouse to improve their rating on Yelp, which will be demonstrated in our Shinyapp.

II. Data Pre-processing

Selecting on Yelp review and Business data, we picked out the reviews and business data for steakhouses (excluding the Japanese steakhouses that largely differ from American steakhouses), which has 28409 reviews and 314 businesses. From the review and business data, we extracted the information of rating, business ID,review text, name, address and state.

We classified the reviews into ”good reviews”(star upon 3) and ”bad reviews”(below 3). Then we use NLP to clean the text: do tokenization, strip the stopwords and neutral words (such as would), strip the punctuation and reduced the words to prototype.

To extract meanings from the review texts, we now perform topic analysis using extensive models like LSA, LDA and NMF.

By LSA, we got the strengths of topics in reviews, shown in Figure 1. There’s a ’elbow’ around Topic 3, thus including 4 topics in our analysis is decent. Though efficient, the results of LSA are hard to interpret as the components may be arbitrarily positive/negative.

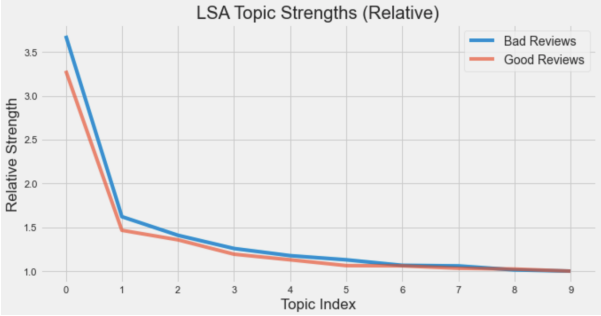

Fig. 1. Topic Strengths

We then tried NMF and LDA, LDA is a probabilistic model and NMF is a matrix factorization and multivariate analysis technique. Both give well interpretable results: listing top topics and words in positive and negative reviews. But NMF seems to work better in short text topic modeling, so we use NMF to get topics weight and normalized the sum of positive/negative to 0-1. For each business, we calculate the mean as topic strength and count review numbers for regression. Summarized topics are listed in findings.

III. EDA

After we get clean data of steakhouse, we merge positive and negative weights. There’re one business with no negative review and two with no positive reviews, so we dropped them for regression. To see the linear relationship of topics and ratings, we add text counts to our data frame, and weight positive/ negative topics by positive/ negative text counts.

Before regression, we used EDA to explore the distribution of data. We draw a correlation

map of stars and 8 topics in Figure 2 and histograms of each topic weights in Figure 3.

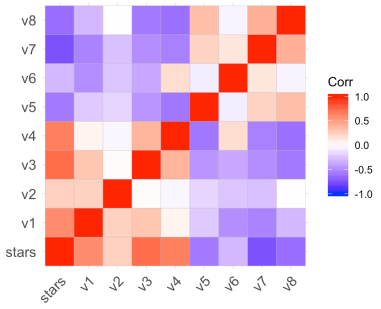

Fig. 2. Correlation map of stars and topics

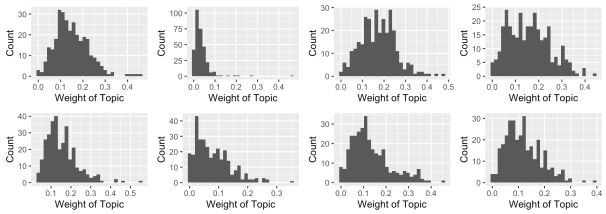

Fig. 3. Histograms of topics 1-8 by row

From Fig.2, we could tell that v3 and v7 are highly correlated with stars (≈0.7). And there exists relatively high correlation among positive and negative topics, which inspires model selection and multi-collinearity check.

From Fig.3, we could see that most of these histograms are symmetric or slightly left skewed, and there might exist some outliers with a large topic weight. In particular, there seems to be an outlier for the business whose topic 2 has a weight over 0.4.

To detect any outliers, we also draw bivariate boxplots to see which business has weight far from the majority. Index 142 with an excessive topic 2 weight. We also dropped index 262, which shows as an outlier in qqplot and residual plots as and an influential point. Then we refit regression.

IV. Linear Regression Model

After summarizing topics and get the normalized weights, we build linear regression model. Since steakhouse businesses are located in four states, so we add states in interactions in the regression model to explore the state effect.

A. Model Selection

We first fit our full model using all topics and interaction with states, then use all possible subsets and stepwise for selection. we also tried lasso on all topics. Topics are denoted as v1 - v8. We also tried Lasso for resolving multicollinearity. Compare different models by cross-validation, results were listed in Table 1.

By comparison, we could see that with interaction terms, the third model has best prediction accuracy—with the highest R^2, lowest RMSE and low MAE. It's MAE is 1.8% larger than the first model, but the formula is much simpler. So we chose it as our final model.

B. Model Specification and Interpretation

$$stars = 1.79176 + 3.01248v1 + 2.17228v2 + 2.97643v3 + 3.00886v4 - 0.52453v7$$

The coefficients of v1-v4 are positive and v7 is negative, which are identified with our previous analysis. Therefore, better bar and place, better cheese curds and decoration, better service and atmosphere, better steaks quality temperature control would help steakhouse business increase their ratings. For example, if the strength of service and atmosphere increase by 0.1, the rating would increase by 0.298.

C. Statistical Analysis

The model has an adjusted $R^2 = 0.8866$, which implies a good fit. The proportion of variation in ratings explained is 0.8866.

To check the significance of variables, we could perform overall and partial F test. From the ANOVA table, all variables in the model

Criteria	Model	R^2	RMSE	MAE
Ad R^2, C_p	stars∼v1+v2+v3+v4+v2:oh+v7:oh+v7:il+v8:oh	0.874	0.196	0.161
BIC	stars∼v2+v5+v6+v7+v8	0.851	0.215	0.172
stepwise bic	stars∼v1+v2+v3+v4+v7	0.877	0.196	0.164
Lasso	stars∼v1+v2+v3+v4+v5+v6+v7	0.852	0.216	0.174

TABLE I
Candidate Models and Prediction Accuracy

are significant with p-value=0.05. Specifically, to see the significance of v7 in the final model, we suppose $H_0 : \beta_{v7} = 0 \quad vs. \quad H_1 : \beta_{v7} \neq 0$, the F statistics=5.7632 with p-value=0.01697$<$0.05, so we reject the null hypothesis and say v7 is significant with α=0.05.

We could also get the confidence interval of coefficients. β_{v7}=-0.525, with 95% CI (-0.954,-0.095)and thus we are 95% confident that the population parameter is between -0.954 and -0.095. Also, since the CI not includes 0, we reject the null hypothesis.

D. Model Diagnostics

Now we checked model assumptions. Since all coefficients have VIF $<$ 2, there's no multicollinearity detected.

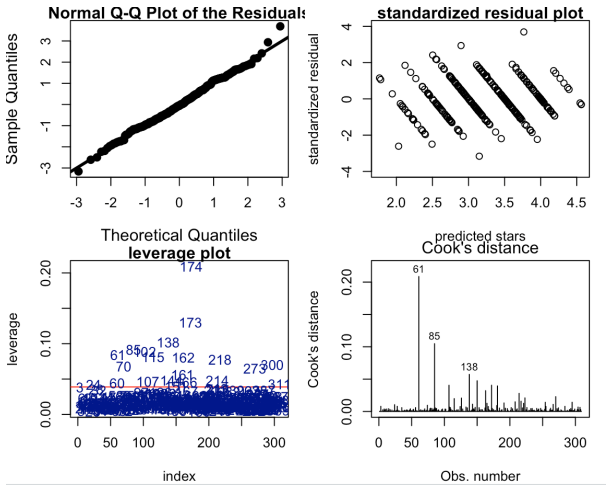

Fig. 4. Diagnostics plots

From Figure 5, QQ plot showed the normality is plausible, as points were distributed closely to the qq line, and residual plot showed the linearity and homoskedasticity are also reasonable, since points look randomly scattered and no non-linear trends were detected.

The leverage plot shows several high leverage points, like 174,173,138,85,102, etc. There's not influential points as Cook distances are lower than 1, we also consider DFFITS, DFBETAS here.

Given the results of diagnostics and model validation, our model meets the requirement for accuracy, simplicity and robustness.

V. Findings from the Model

A. Details of the topics

From the topic modeling, we get summarized topics:

Positive: 1. Bar/Good place/Happy hour, 2. Cheese curds/Decoration, 3. Service and atmosphere, 4. Quality of steaks

Negative: 1. Service and service time, 2. Place/Bar, 3.Temperature of Steaks, 4. Quality of other sides and snacks

As we can see, both of positive and negative topics are focused on the following aspects:

1) Alcohol, event setting(happy hour/Bar)
2) Taste of steak(quality & temperature)
3) Service; 4) Quality of food;

The concluded aspects are very reasonable and corresponds to common sense. However, it should be noticed that if you want your restaurant to excel among others, steak quality may not be your priority. Instead, focus more on the things like service and bar could help.

B. Reason behind the variable selection for regression model

However, there's no v5, v6, v8 in regression model. Considering that, there are relationships between all topics, for they describe these 4 aspects collectively. EDA also shows this. The information behind are somehow overlapped.

After looking into the details, we fount it reasonable to pick these 5. For they have enough information about the 4 aspects, and the variable of temperature can depicts deliciousness of steak, for both quality of steak meat and the temperature are very important for the taste.

VI. Recommendations for Businesses

Bar and wine settings really worth your attention, for it's the top1 influential positive topic and top2 influential negative topic. Other than this, you should always focus on service, it's the top3 influential positive topic and the most influential negative topic.

A. Recommendations for wine & other alcohol

For bar and event arrangement, this is the most significant positive topic from the model, therefore we advise the steakhouses to put

the utmost attention to this. Wine is the best companion for steak with no doubt. More than this, happy hour is valued too. Hence, we could conclude that people value almost all kinds of alcohol provided in the restaurant.

Restaurant shouldn't only focus on the collection of wine, but also the collection of whisky maybe even of beer. A good collection of alcohol and a well established bar would considerably improve your rate. Furthermore, appropriate promotion, for instance, happy hour or some discount on drinks may attract customers to your restaurant and improve your profit by selling drinks. Improving your alcohol collection is a win-win strategy and we strongly recommend you to take this.

B. Recommendations for Food

For the food, firstly, temperature and quality of steak are highly valued, hence it's necessary to practise cooking style on steak instead of simply caring about meat quality. As we all know, the quality of meat and the cooking temperature can significantly influence the taste and everyone loves juicy steaks.

Meanwhile, for sides and other dishes, quality of cheese is emphasized. It should be noticed that, the data provided by Yelp only consists restaurants from several regions and among them, many are from Wisconsin. Since Wisconsin is famous for its cheese, this result might comes from the data bias. But indeed, cheese is a good companion of steak, it would be wise to improve the quality of other foods from your restaurant, especially for cheese.

And if the targeted restaurant is not located in Wisconsin, this won't be a problem. For, sides and dishes still need to be improved and the types and quality of the sides and dishes do require improvement, if you want to impress costumers with your high quality food.

C. Recommendations for Service

For the service, long waiting time is criticised a lot. Therefore, we would still advice the restaurants to arrange schedule and working process to minimize the waiting time between each dishes. Meanwhile, atmosphere and decorations are also viewed as important factors, so we would advise to improve the decorations of the restaurant, please don't hesitate spending money on this, it's really important.

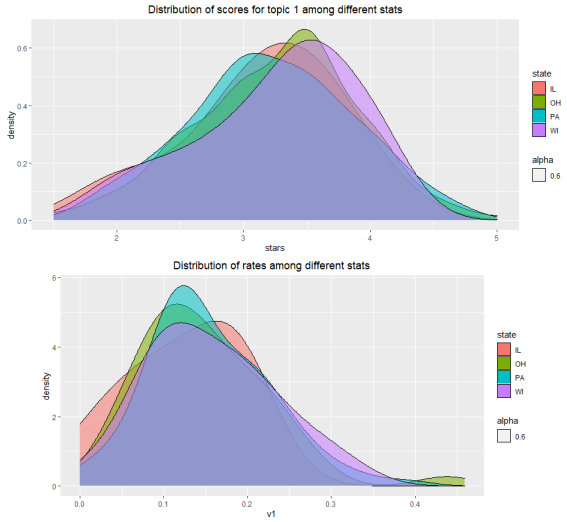

Fig. 5. Examples for the differences in distribution

D. Summary of the recommendations

It should be noticed that, the distribution of stars and the scores varies in different states. For instance, just as the plot shows, the mean of the topic 1 score of Pennsylvania is higher than the one in Ohio. Since the distribution of each aspects varies among states, what we should bear in mind is that situations may vary in different ethnic community, states and region. Hence, the recommendations provided is not absolute.

Different region has different diet habits, so for the much more detailed reference, please use our app for the tailored recommendations.

VII. Conclusion and Discussion

Overall, we perform topic modeling and MLR to investigate the factors impacting ratings in steakhouses, obtain analytic insights and give constructive recommendations for steakhouse business. Ratings could always be affected by many factors, so concerted efforts are needed to increase business ratings.

However, there're some steakhouses with few text counts, and some even receive 0 positive/negative reviews. So if more reviews were given, we could get more precise topic weights and model to reveal the relationship.

VIII. Contributions

A. Summary

1) YW wrote III,IV; edited the summary overally
2) YZ wrote V, VI, VII; edited the summary overally.
3) ZL wrote I, II; edited the summary overally.

B. Presentation

1) YW explain Regression model and Finding from the model parts; edited slides of these parts.
2) YZ present data pre-processing and topic selection parts; edited slides of these parts.
3) ZL present recommendation and conclusion part; demonstrate Shinyapp; write the slides overally.

C. Github/Code

1) YW wrote EDA and regression code; edited and tided up NLP codes. (part3_organized* and bus_regression*)
2) YZ wrote NLP models code; decided topic selection. (2.Model_Buiding*)
3) ZL wrote data-preprocessing and part of topic selection (part1_dataprocessing*); tided up all codes and files.

D. Shiny app

1) YW edited and provided general feedbacks.
2) YZ helped debugging and provided general feedbacks.
3) ZL created and wrote the code.