To Bar Owners: It's Time to Improve your Yelp Ratings!

Runze You, Yudi Mu, Xinran Miao

November, 2020

Introduction

Yelp is an American company which publishes crowd-sourced reviews about businesses. From Yelp open data with 942,027 entries of reviews for 36,327 entries of businesses in four states in the US (OH, PA, WI and IL), we extracted all the 4743 businesses in the bar category with their 272,346 reviews. With the purpose of helping bar owners improve Yelp ratings, our project follows two threads. First, we analyzed explanatory attributes of businesses on ratings and proposed general advice. Second, we analyzed the customer review texts for aspects that significantly affect customers' judgement and to what extent they like or dislike the food, drink, or service. The statistical methods included exploratory data analysis on reviews, ordinal logistic regression, and latent dirichlet allocation. In the end, we concluded specific suggestions to bar owners.

Data Description and Pre-Processing

Business data

The business data set includes basic information, ratings, review counts, working hours, and Yelp-assigned attributes for all business with missing values. We extracted and simplified the attributes, together with review counts and weekly open hours, to features. After removing either scarcely observed or vague features (Table 1), 8 of them remained (Table 2).

Reasons	Features
Few observations	WiFi, attire, by appointment only, music, best night, outdoor seating, reservations, good for groups, table service, wheel chair accessible, accepts Bitcoin
Vaguely described	Ambience, price range (with levels 2, 3, 4), good for meal, happy hour, alcohol, good for kids

Table 1: Reasons for dropping features from business data set.

Features	Type	Levels / Range
Accepts credit cards	Factor	True, False
Bike parking	Factor	True, False
Business parking	Factor	True, False
Delivery	Factor	True, False
Has TV	Factor	True, False
Noise Level	Factor	Quiet, Average, Loud, Very loud
Weekly hours	Numeric	[3,168]
Review counts	Numeric	[3,1879]

Table 2: Extracted features from business data set.

Review data

Each entry of the review data set includes review texts and Yelp extracted adjectives. While the latter remained unchanged, we dealt with the former by removing invalid symbols, punctuation and stop words as well as converting all characters to lower cases. After that, we tokenized these sentences to words for future exploration. Figure 1 shows the ordered word frequencies, with the top 5 words being *food, good, place, great* and *service*.This sorted list of most frequently appeared 500 words can also be found in wordFreq.txt on Github.

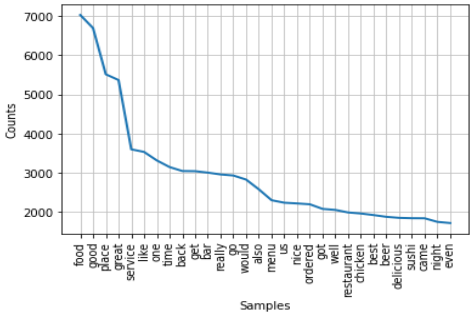


Figure 1: Word counts

Exploratory Data Analysis

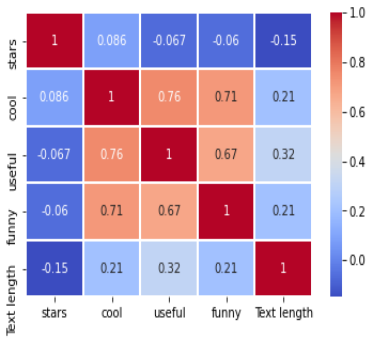


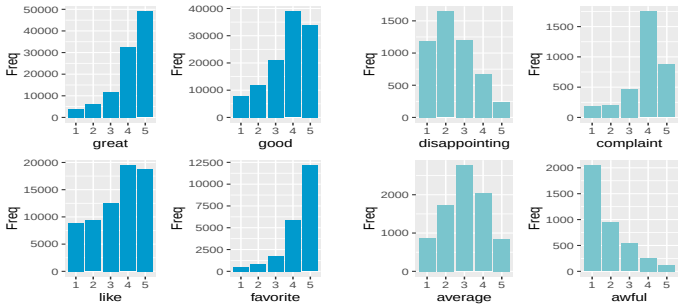
Figure 2: Word counts of pre-processed review texts.

As mentioned before, Yelp provides feedback of each review as three adjectives ("useful", "funny" and "cool") for customers' reference. We treated these three words as variables and plotted the

heat map along with text length versus ratings. Figure 2 suggests negligible relationship between ratings and these variables and hence we excluded them from future discussion.
For the pre-processed review words with frequencies ranked within top 500, we chose the most important ones and grouped them into the following topics.

1. Feeling

The result for feelings is the same as common sense that positive feeling related to higher star ratings while negative feeling related to lower star ratings. However, "complaint" has a relative high star rating. This might be that even people complaint some thing about the bar, they are more like suggestion.

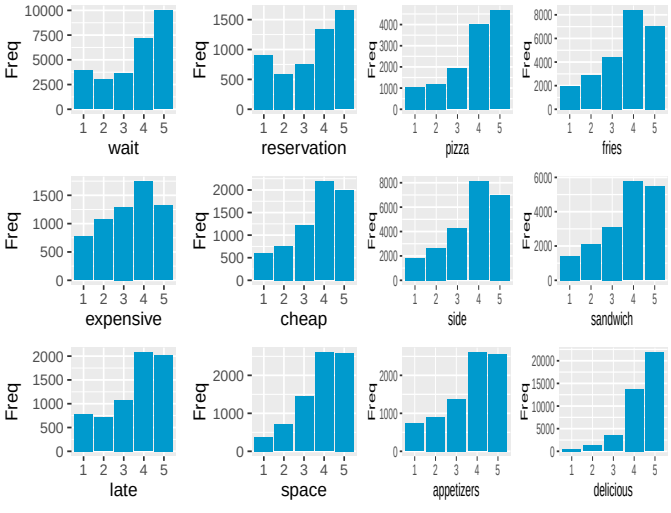


(a) positive words (b) negative words

Figure 3: Feeling

2. Service and Food

From the second panel in Figure 4, customers are not averse to reservation, with a possible reason that reservations save time. From the second row of panels, the price is not a determined aspect. But if the price is lower, the 5 stars tends to appear more frequently. Moreover, although people are fond of most common food, bars can still improve themselves by serving better side food and fries.



(a) service (b) food

Figure 4: Service and Food

3. Wine

Drinks usually relate to positive feedback. Among the drinks, coffee and beer are most popular (Figure 5). For the containers, bottles are preferred compared with the draft. It was a counter-intuitive that some common drinks including Vodka, Gin and Rum weren't found. We suspected that was because people are used to them in bars so they don't give feedback on them.

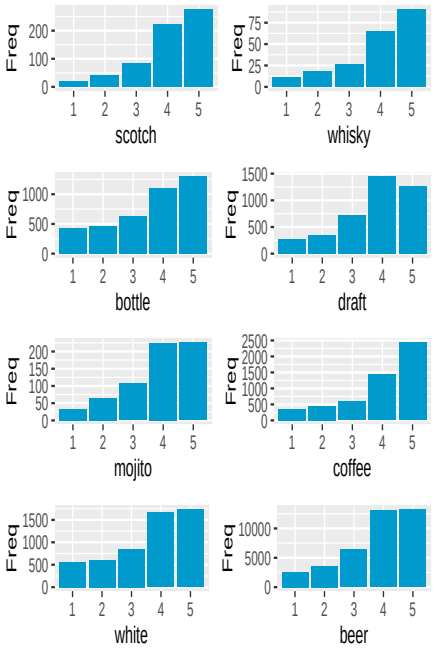


Figure 5: Wine

Part1: General Findings About Bar Market on Yelp

Generalized Linear Model

In order to predict ratings via eight variables extracted from the business data set (Table 2), we fitted a generalized linear model with probit link function and selected variables by AIC criteria. As displayed in formula (1), the probability of rating stars less than j has a linear relationship with different variables after a non-decreasing transformation $probit$. With the purpose of increasing ratings, conversely, it's desired to increase one variable if the coefficient is negative and vice versa. Therefore, it's better to have more review counts, less working hours, provide bike parking, don't have TV, and keep a relatively quiet (at least not too loud) environment. The conclusion about TV and working hours seem to be strange, but it makes sense if we consider them as a reason for making noises.

$$\begin{aligned}\Psi(P(\text{Stars} \le j)) =& I_j - 0.001\ \text{ReviewCount} - 0.13\ \text{BikeParking} \\ &+ 0.01\ \text{WeeklyHours} + 0.37\ \text{HasTV} \\ &+ 0.19\ \text{Loud} - 0.24\ \text{Quiet} + 0.70\ \text{Very Loud},\end{aligned}$$

where j taking values from 1.5 to 5 with step 0.5, I_j's are coefficients that increase as j increases, and Ψ is the probit link function that is chosen by plotting the simulated p-values for a goodness-of-fit test across link functions (figures here)
After fitting this model, we checked the assumptions of homoscedasticity and normality by plotting the probability level

residuals [1] and normal quantile-quantile plot for them. From Figure 6, the assumptions weren't violated.

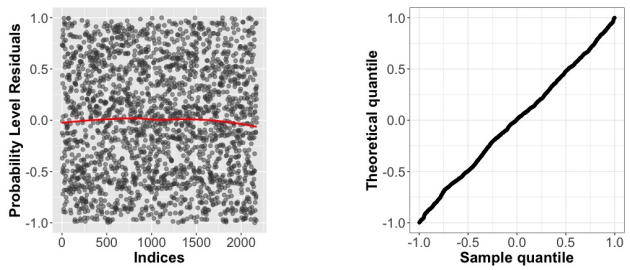


Figure 6: Model Diagnostic for homoscedasticity (left) and normality (right).

Sentiment Score and Latent Dirichlet Allocation

Using the positive and negatives words tokenized from the preprocessed review texts, we calculated sentiment scores via formula (1) for each review (check review sentiment.txt on Github). We regarded reviews with positive scores and positive ones and the rest as negative ones, on each of which we fitted Latent Dirichlet Allocation (LDA) models separately. Figure 7 & 8 shows the results of two subjects, where the number of subjects was chosen after a bunch of trials.

$$\frac{\#positive\ words - \#negative\ words}{length\ of\ sentence} \quad (1)$$

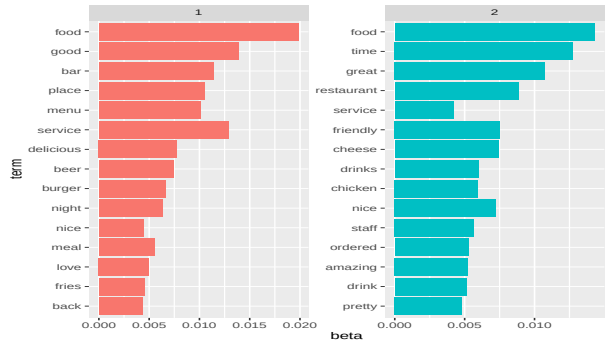


Figure 7: LDA on positive reviews

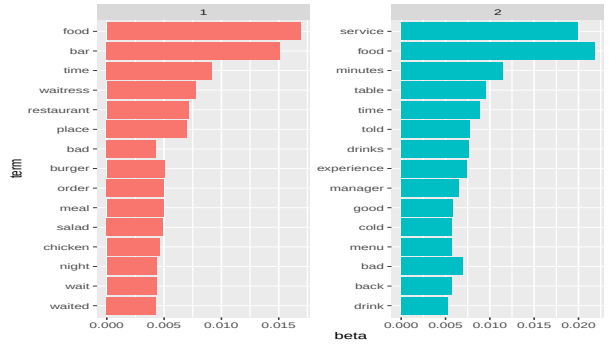


Figure 8: LDA on negative reviews

Findings:
Figure 7 & 8 showed first 15 words in each subject. The boundary in two subjects are not so distinctive especially in the first few words, but we could still roughly define subject 1 as Food Service and subject 2 as Experience. For food, many people pay attention to the menus, within which they will comment more on burgers than drinks. Good burgers and fries will attract more positive reviews, while bad chicken will lead to negative reviews. Except for food, short waiting time, good service and friendly atmosphere are also important.
Recommendations:
For bar owners, it's recommended to provide attractive and clear menus. In the sense that most bars also serve as restaurants, good food, especially delicious burgers, fries and chicken, will play an important role. It's also a great idea to offer reservations for customers to reduce wating time.

Information Gain and Decision Tree

In the learning process of decision tree algorithm, information gain is an important index of feature selection. It is defined as how much information a feature can bring to the classification system - the more information a feature brings, the more important it is, and thus the greater the corresponding information gain. We used the sentiment scores as an important feature to calculate entropy, and then made the decision tree. Figure 9 shows part of it and the whole plot can be found on [Github].

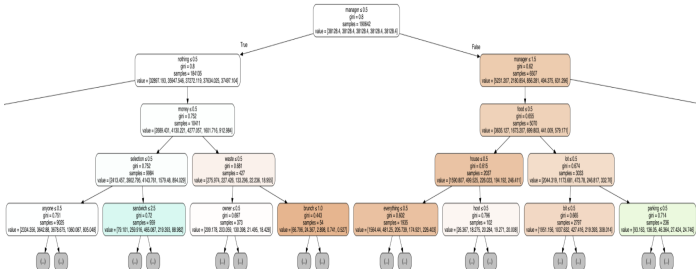


Figure 9: Part of decision tree

Findings:
From Figure 9, *manager*, together with *waste, price, money* and parking are determined words. Compared with LDA where some food such as sandwiches and salads plays an role, price and parking along with some other aspects are influential with respect to information gain.
Recommendations:
Food and service mean a lot to bars. Because most of the business in the bar category are also restaurants. Some common food like burgers and fries are important, while those side food for drinkers are also important. And besides food, service of course is a significant factor. Price and parking also show their influence on star rating, even though they are not significant in the linear model.

Part2: Specific Businesses Plan Recommendation - Interpreted with Word Count

In order to offer specific recommendations to business owners, we counted the numbers of closest tendency words, either positive or negative, of the selected words, each of which represented a specific aspect. In detail, we did the following steps.

1. Selected the most frequently appeared words, especially nouns, and divided them into four aspects we focused on.

    - **specific food**: cheese, chicken, burger, sushi, fries, sauce, salad, sandwich, pizza

- **flavour**: food, bar, menu, beer, restaurant, dinner, drink, meal
- **price**: ordered, order, table
- **service**: place, service, time, night, staff, people, experience, atmosphere, wait, server, area, minutes, lunch

2. Found the appearance numbers of nearest positive words or negative words based on the data-set [positive and negative words collection]. Besides, in order to eliminate potential influence of sentence structure, we deleted the function words in each sentence.
    - For example, there is a review "the shrimp tacos and house fries are my standbys. the fries are sometimes good and sometimes great", and after deleted function words, the sentence becomes "shrimp', 'tacos', 'house', 'fries', 'standbys', 'fries', 'good', 'great'". Hence, the count of the aspect 'fries' adds 1.
3. Conducted the Chi-square test on 2×2 contingency table, which consisted of positive and negative words counts for total reviews of a specific aspect. An example is shown in 3
4. Based on p-value of chi-square test and the positive or negative proportions, we believed that the attitudes towards this aspect of certain merchants was different from general case, and if the proportion of negative count is greater, we would get the conclusion that this aspect needs improved.

Specific Aspect	♯ of Positive	# of Negative
Cheese in "The Old Fashioned"	1481	576
Cheese in total reviews	29766	12597

Table 3: example table

Limitations

- For the generalized linear model to explore the influence of business attributes on ratings, the assumption of linearity may be violated since the AIC value is large (more than 1000).
- For Latent Dirichlet allocation model, we are not sure which is the best number of subjects, in the project, we tried three-subject at first, but the result was not satisfying while two-subject was okay but this was our subjective judgement.
- Although the word count result roughly demonstrate the weakness in a relatively specific selected aspect for business merchant, the real attitude revealed in the reviews may be much more complicated than the word count model, which only counts the nearest tendency word. Besides, for the shop who only has few reviews, this kind of test became meaningless. Last but not least, a weighted reviews count combined with the number of people who think it is useful may seem more reasonable.

Conclusion

- It's better to have more review counts, less working hours, provide bike parking, don't have TV,and keep a relatively quiet (at least not too loud) environment.
- Food and service are even more important than drink. Most alcohol drinks have positive reviews, coffee is also important. If owners want to improve, they should at first improve their menus, burgers and fires and also improve atmosphere and service especially reduce waiting time.
- For each merchant or shop, our word count model tells which aspect should be reinforced compared with the whole market.

Contributions

Yudi Mu

- Wrote github codes and summary for data cleaning on review dataset, EDA, sentiment score and LDA, information gain and decision tree.
- Shiny App second tab: LDA model result.
- Presented EDA, sentiment score and LDA, information gain and decision tree.

Xinran Miao

- Wrote github codes and summary for pre-processing the business dataset, constructing the generalized linear model and its diagnostic.
- Wrote Shiny App codes for the first tab (generalized linear model).
- Presented the introduction & data cleaning and the generalized linear model in the presentation.

Runze You

- Wrote github codes and summary for pre-processing the reviews dataset, constructing the word count model.
- Wrote Shiny App codes for the second tab (word count model for business reviews).
- Presented the word count model and the usage of our shiny app in the presentation.

References

[1] Brandon M Greenwell, Andrew J McCarthy, Bradley C Boehmke, and Dungang Liu. Residuals and diagnostics for binary and ordinal regression models: An introduction to the sure package. *R J.*, 10(1):381, 2018.