Yelp Dataset Analysis

STAT 140XP

Annaka Schone, Caitlin Ree, Damien Ha, Joyce Xu, Meiyi Ye, Sylvia Deng, Wenhong Sun

March 2024

Department of Statistics and Data Science

Abstract

Yelp is a platform that crowdsources reviews about businesses, with the majority being restaurants. In this study, our objective was to gain insight into the following questions:

What factors contributed to the overall star rating of Chinese restaurants? Which factor has the greatest effect?

Do a restaurant's Yelp reviews accurately reflect its star rating?

Contents

1. Introduction

Yelp is a platform that crowdsources reviews about businesses, with the majority being restaurants. In an era where online reviews can make or break a restaurant, understanding the relationship between customer feedback and ratings on platforms like Yelp is crucial for both consumers and business owners. With a focus on Chinese food restaurants, this study seeks to explore the complex relationship of Yelp reviews and restaurant ratings. By analyzing an extensive Yelp dataset, we aim to answer two questions: what factors significantly influence the star ratings of restaurants, and do reviews accurately reflect a restaurant's star rating?

1.1 Dataset Overview

The Yelp dataset being examined in this study was assembled for the **Yelp Dataset Challenge** in 2018. It contains Yelp business data, reviews, and user data. For our analysis, we reduced the data down to only Chinese food restaurants. This reduces unwanted variability but keeps the dataset relatively large, allowing us to perform an accurate analysis.

The variables we focused on in this analysis include the following:

- Stars given by each review
- Average stars given by each user
- Average stars for each business
- Total number of reviews for each business
- Sentiment (positive or negative) expressed in each review
- Overall sentiment orientation of each business
- Percentage of reviews with positive sentiments

1.2 Data Cleaning and Merging

Before conducting our analysis, a crucial step was to prepare the dataset for examination.

Given the large size of the Yelp dataset, it was essential to load the data in manageable chunks. We loaded three files: business.json, review.json, and user.json into R to transform the raw data into csv, a workable format.

Then we merged data from these files based on common columns in order to create a comprehensive view of each restaurant, encompassing user reviews, business information, and user data.

We cleaned the resultant dataframe to remove any inaccuracies or irrelevant information. Additionally, we selected a subset of the data under the Chinese food restaurants category and our variables of interest. This refined dataset was then ready for exploratory data analysis and further statistical modeling.

2. Exploratory Data Analysis

In the process of exploratory data analysis, we wanted to see the variability and most common values of the features in the dataset that pertained to the questions we wanted to answer. This process also involved some initial work in determining the sentiment of each review, which will be discussed further in the "Methods" section.

The variables of interest were:

stars-review: Stars given by user for that particular review.

average-stars: Average stars given by that user.

stars-business: Average stars for that business.

review-count: Number of reviews for that business.

sentiment: Sentiment for that particular review.

business-sentiment: Majority sentiment for that business.

sentiment-percentage: Percentage of positive review sentiments

We aimed to find the average sentiment for the reviews in our dataset so that we could see how the sentiment changes based on number of reviews, star rating, and the other factors listed above.

2.1 Figures and tables

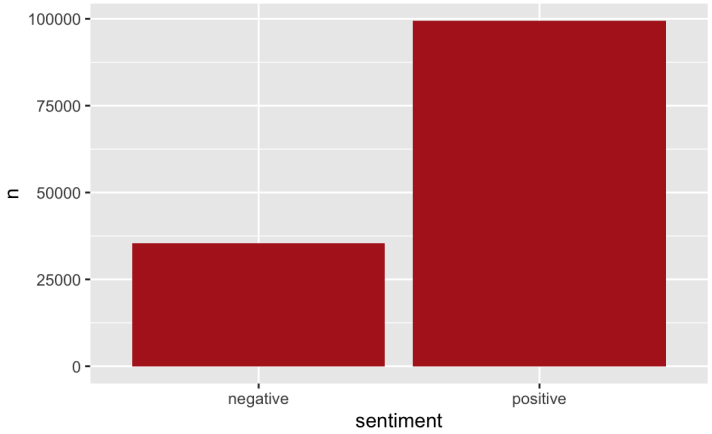

Figure 2.1: caption for word sentiment bar chart.

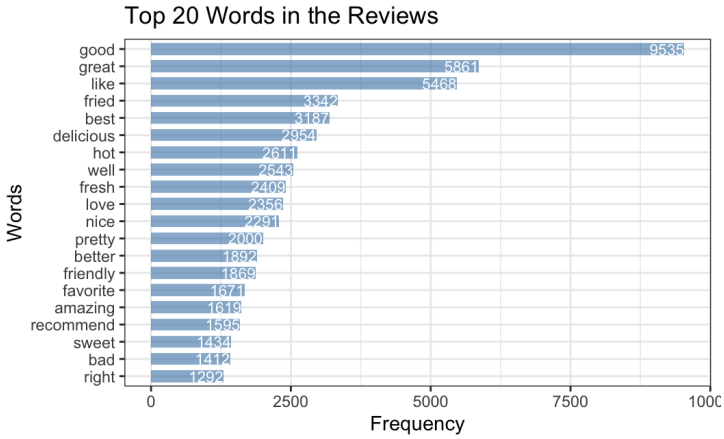

Figure 2.2: caption for top 20 words graph

Figure 2.3: caption for wordcloud

3. Methods and Analysis

To examine whether a Chinese restaurant's Yelp reviews accurately reflect its star rating, we performed a sentiment analysis on the reviews in our data. Each review was evaluated using a sentiment lexicon, which allowed us to categorize it as either "positive" or "negative." This step allows us to quantify the emotional tone of the reviews, and we incorporated this sentiment data into our dataset as a predictor variable.

Building on the sentiment analysis, we want to further investigate which aspects of a Chinese restaurant have the greatest effect on its Yelp star rating. To achieve this, we employed multiple linear regression and logistic regression models.

The linear regression model was designed to compare the original average star ratings of the businesses with the sentiment metrics we derived from the reviews, such as the overall sentiment and the percentage of positive reviews. It aimed to highlight the predictive power of review sentiments on the star ratings.

Concurrently, the logistic regression analysis was utilized to delve deeper into the categorical outcomes of star ratings, examining how various aspects, including review sentiment, review count, and other business-specific features, impact the likelihood of a restaurant achieving a certain star rating.

4. Conclusion

Generally speaking, a Chinese restaurant's star rating provides an accurate reflection of its review sentiment on Yelp.

The variables with the largest overall impact on the sentiment of Chinese restaurants include the star ratings on each particular review and the number of reviews per business.

Further research could include running this analysis on another subset of restaurants and businesses in the Yelp dataset. From there, we would gain further insight into whether our conclusions about star rating and the relevant factors contributing to review sentiment apply to businesses on Yelp in general.

A. Appendix

R Code Generating the Linear and Logistic Regression Models

```
##
## Call:
## lm(formula = stars_business ~ business_sentiment + sentiment_percentage,
##     data = yelp_dataset_final_perc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58778 -0.29719  0.07715  0.31904  2.42215
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.07785    0.01850 112.298   <2e-16 ***
## business_sentimentpositive -0.02417    0.02081  -1.162    0.245
## sentiment_percentage        2.03410    0.03313  61.401   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4888 on 16003 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.3295, Adjusted R-squared:  0.3295
## F-statistic:  3933 on 2 and 16003 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = stars_business ~ stars_review + average_stars +
##     state + review_count + sentiment + business_sentiment + sentiment_percentage,
##     data = yelp_dataset_final_perc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.39589 -0.24846  0.01247  0.28313  2.03499
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.928e+00  3.096e-02  62.258  < 2e-16 ***
## stars_review                  1.056e-01  3.418e-03  30.903  < 2e-16 ***
## average_stars                 6.613e-03  5.207e-03   1.270 0.204074
## stateAZ                       8.029e-02  2.642e-02   3.039 0.002378 **
## stateCA                       1.040e-01  3.079e-02   3.377 0.000734 ***
## stateDE                      -9.356e-02  4.212e-02  -2.221 0.026344 *
## stateFL                       3.009e-02  2.466e-02   1.220 0.222437
## stateID                      -1.071e-01  3.596e-02  -2.977 0.002914 **
## stateIL                      -1.991e-01  5.263e-02  -3.783 0.000156 ***
## stateIN                       1.216e-02  2.651e-02   0.459 0.646403
## stateLA                       1.301e-01  2.841e-02   4.578 4.72e-06 ***
## stateMO                       6.014e-02  2.537e-02   2.371 0.017757 *
## stateNJ                       5.863e-02  2.722e-02   2.154 0.031233 *
## stateNV                       2.387e-02  2.542e-02   0.939 0.347727
## statePA                       3.570e-02  2.328e-02   1.533 0.125241
## stateTN                      -6.521e-02  2.690e-02  -2.424 0.015371 *
## review_count                  1.313e-04  3.445e-06  38.121  < 2e-16 ***
## sentimentpositive            -1.906e-01  1.114e-02 -17.110  < 2e-16 ***
## business_sentimentpositive -5.797e-02  1.906e-02  -3.041 0.002362 **
## sentiment_percentage          1.799e+00  3.223e-02  55.823  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.442 on 15862 degrees of freedom
##   (132 observations deleted due to missingness)
## Multiple R-squared:  0.4523, Adjusted R-squared:  0.4516
## F-statistic: 689.3 on 19 and 15862 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## glm(formula = binary ~ stars_review + average_stars + state +
##     review_count + sentiment, family = "binomial", data = yelp_dataset_fina
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.6240284  0.2128889  -2.931  0.00338 **
## stars_review       0.2113596  0.0306044   6.906 4.98e-12 ***
## average_stars      0.0932458  0.0424371   2.197  0.02800 *
## stateAZ           -0.6344344  0.2231646  -2.843  0.00447 **
## stateCA           -1.4146165  0.2681666  -5.275 1.33e-07 ***
## stateDE           -0.1395539  0.3559806  -0.392  0.69504
## stateFL           -0.3902056  0.2007986  -1.943  0.05198 .
## stateID           -0.5804108  0.2862261  -2.028  0.04258 *
## stateIL            1.1888855  0.6319747   1.881  0.05994 .
## stateIN           -0.2298430  0.2299165  -1.000  0.31747
## stateLA           -0.1959170  0.2645303  -0.741  0.45892
## stateMO           -1.0312063  0.2003593  -5.147 2.65e-07 ***
## stateNJ           -0.1744060  0.2235313  -0.780  0.43526
## stateNV           -0.9646351  0.2345429  -4.113 3.91e-05 ***
## statePA           -0.7446485  0.1855902  -4.012 6.01e-05 ***
## stateTN           -1.3417298  0.2065399  -6.496 8.24e-11 ***
## review_count       0.0149429  0.0006237  23.959  < 2e-16 ***
## sentimentpositive  1.6210484  0.0877576  18.472  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```