Yelp Text Mining

Linchen Deng, Kenny Jin, Runxin Gao

UW-Madison

ldeng33@wisc.edu, jjin59@wisc.edu, rgao35@wisc.edu

***Abstract* —**Yelp is a popular website which allows users to post free-form reviews and recommendations of many kinds of entertainment. Meanwhile, users can rate businesses by stars from 1 to 5. In this paper, we focus on ratings and review texts of restaurants in Wisconsin and try to predict ratings based on review texts by using the datasets Yelp released in 2015. We employed a machine learning based method to analyze natural language and to extract features that influence the rating. Then we built our prediction model. Our model generated a 0.82306 RMSE for the test dataset.

I. INTRODUCTION

Advances in the Internet have transformed commercial activity that people do, never more so than today considering its substantial impacts in customer behaviors. Many people have the experience that they search in the Internet for top businesses in food, shopping, nightlife, and entertainment in order to pursue the most enjoyable experience. Yelp, a popular review website serves as a guide for the public to fill these needs. They allow users to rate businesses, to leave own reviews, and to give recommendations for other users.

Yelp is important for both customers and business owners. A survey conducted in 2017 showed that 85 percent of customers regarded these online reviews as valuable as personal recommendations ("The enormous influence of online reviews", 2018). As a result, positive ratings and reviews are helpful for business owners. According to BrightLocal, a consulting firm, "Customers are likely to spend 31 percent more at businesses with excellent reviews." Even one-star improvement in the overall rating can bring 5 to 9 percent of revenue to a small business's income. On the other hand, a single piece of negative review can discourage 22 percent of respondents to purchase on a business ("The enormous influence of online reviews", 2018).

As increasing number of people highly depend on the reviews for food hunting, it is reasonable to regard the review as a powerful tool to evaluate restaurants. In our analysis, we aim to find out how different review contents lead to distinct star ratings. When some words and phrases indicate sentiments, we try to extract use them as indexes for the rating star. Inspired by the idea of natural language processing (NLP), we decide to use a machine learning method to extract and to select features. Then we fit a multiple linear regression model and apply it for prediction.

This paper is organized into five parts: Part I gives a background of the Yelp website and an overview of our methods. Part II gives a general idea of the dataset and detailed explanations of our methods. Part III gives the prediction results of the test dataset. Part IV discusses the strength and limitations of our models. Part V concludes about our project and future works.

II. DATA AND METHODS

A. Data Cleaning

We are provided three datasets. The training dataset contains reviews from 48860 users rating stars, restaurant ID, restaurant type, review date, restaurant names. The test and validation datasets consist of all features the training dataset has except stars.

The reviews are in a free format (See Figure 1). They may contain symbols, extra punctuations, numbers, and even emojis. As a result, it is inevitable to normalize the texts. We decapitalized all words, removed all punctuations and symbols, and stemmed every single word by using Porter Stemming Algorithm which replaces the word by its root (Shakarad, 2011).

> Ponkey Kong was here, and Ponkey Kong *LOVED* it! No frills kind of place with great congee. And the prices??? YES PLZ. For less than five bucks I had congee with a generous portion of pork liver and preserved egg, a large red bean bun (sweet and moist) and coffee. Long live Sweets Asian Bakery!

Figure 1: an example of review text

B. Data Processing

Among 1015049 extracted words, 3000 words whose total frequencies are the higher than others were selected as features. In addition, features including "useful", "funny", "sentiment", and "nword" from the original dataset were combined with 3000 word features. Since nword appeared to be heavy tailed, log transformation was performed to ensure approximate symmetrical distribution (See Figure 2).

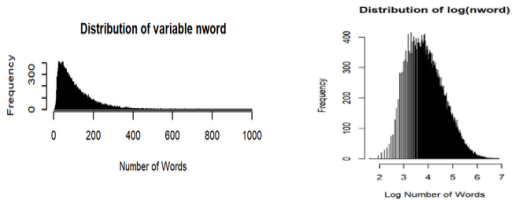

Figure 2. Distribution of variable "nword" before and after the transformation

After feature selection, we noticed some frequent words selected seemed not to be important intuitively in predicting star ratings (See Figure 3). In order to penalize such common words and give higher weights on unique words that appeared less often, we introduced a method called Term Frequency and Inverse Document Frequency (TF-IDF) (Koujalagi, 2015). Details of the functions can be found in our R codes. After TF-IDF projection, common words were weighted lower in individual documents and unique words were given higher weights (See Figure 4). We then proceeded to model fitting.

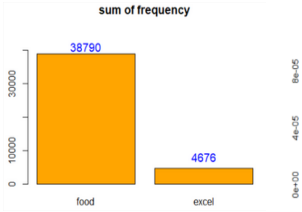

Figure 3. Total frequency of word "food" and "excel" showed

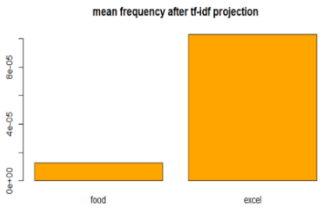

Figure 4. Mean frequency after TF-IDF projection. Common words were assigned less weights than unique words.

C. Model Fitting

After the phrase extraction and TF-IDF transformation, we fitted a linear regression model for the transformed data. The motivation of using this linear model is that we tried other models such as logistic regression and lasso regression models, but they did not work very well and usually produce a relatively higher error rate. Therefore, a multiple linear regression model is chosen. The dependent variable in this model is the possible star rating of a specific text, whereas the predictors are the numbers of the words in the specific text, after the TF-IDF transformation. We used 3004 predictors to fit this model, however R seems to automatically remove 3 of them because these 3 variables are perfectly collinear with some other predictors. Therefore, 3001 predictors (including "useful", "funny",

"sentiment", and "nword") are actually in this model (See Figure 5).

```
Call:
lm(formula = stars ~ ., data = yelp.tokens.tfidf.df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1019 -0.4823  0.0251  0.5099  4.3029

Coefficients: (3 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.440e+01  5.656e+00   2.547 0.010870 *
useful              9.430e-03  2.334e-03   4.039 5.37e-05 ***
funny              -2.190e-02  3.519e-03  -6.223 4.93e-10 ***
nword              -2.672e-05  4.525e-05  -0.590 0.554946
sentiment           2.716e-01  6.490e-03  41.846  < 2e-16 ***
hous               -6.120e+00  4.217e+00  -1.451 0.146778
good               -3.514e+01  1.584e+01  -2.219 0.026518 *
menu               -1.414e+01  7.573e+00  -1.867 0.061856 .

Residual standard error: 0.7761 on 45858 degrees of freedom
Multiple R-squared:  0.6633,    Adjusted R-squared:  0.6413
F-statistic: 30.11 on 3001 and 45858 DF,  p-value: < 2.2e-16
```

Figure 5. The excerpts of the linear model summary

Looking at the summary of this model, we can see that some variables (useful, funny) are significant as they have small p-values. The model is overall significant, as the p-value of this model is also very small.

The interpretation of the model coefficients is a bit tricky, since we transformed the predictors using TF-IDF. For example, to interpret the coefficient of "good", we have to say that if the transformed value of "good" increase by 0.1, the predicted value of stars will decrease by 3.514. This interpretation is not intuitive because of the TF-IDF transformation, so this might be a weakness of the model.

We check the linear model assumptions using the diagnostic plots (see Figure 6). We observe that the residuals seem to be normally distributed from the normal QQ plot. From the Residuals vs Leverage plot, we see that there are a few outliers in the data. The residual plot and the scale-location plot contains strange shapes, and the linearity assumption and the homoscedasticity assumption seem to be violated. We speculate, however, that the shape in these 2 plots are caused by the fact that the response variable "stars" is regarded as a continuous variable, rather than a categorical variable with only 5 categories.

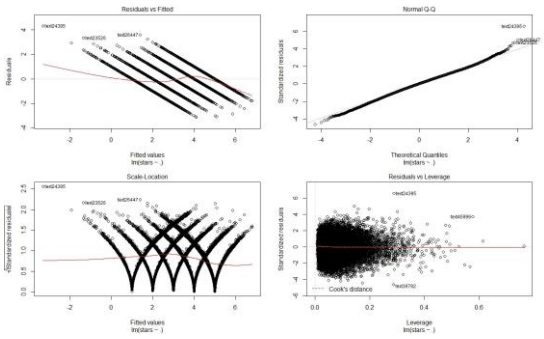

Figure 6. The diagnostic plots of the model

III. RESULTS

Corresponding our expectation, the multiple linear regression model suggested words with strong sentiment polarity are strongly correlated with star ratings. These words include: sweet, love, good, tasty, really good, worst, tough, horrible, poor, and gross. In addition to these emotionally polar words, we also discovered that non-adjective phrases like "even though" tend to imply unsatisfied experience. More subtly, phrases such as "last time" implies worse experience while "next time" implies active expectation for the next visit (see Figure 7). These findings gave us insights about the subtleness of language which could be explored more in the future.

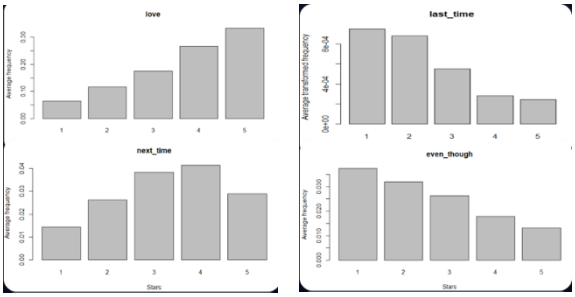

Figure 7. Words and phrases frequency that correlate with star ratings.

IV. DISCUSSION

In this project, we have utilized machine learning based methods to extract significant words from restaurants reviews and built a multiple linear model using mostly words extracted.

Using high-dimensional training data including unigrams and bigrams, we have improved the prediction power by lowering RMSE from 0.97638 to 0.90358. Addition of the four important features from the original dataset led to lowering RMSE of 0.84408.

Another advantage of our method was introducing TF-IDF methods, which decreased our model prediction RMSE to 0.81212. However, the drawback with such method was that our linear model coefficients were difficult to interpret without extra efforts to transform them back to original scale, so we can only observe the correlation through bar plots.

Inevitably, the technological trade-off in our model is that the process is computational heavy and laborious. Due to limited time, we could not restrict our features to fewer significant ones to reach good result, so there might be overfitting problems caused by this. Similarly, our model may not work on small datasets if the sample size was strictly less than 3000.

Regarding prediction results, we did not consider the range of ratings is from 1 to 5. Otherwise, we should have rounded predicted values lower than 1 up to 1 and rounded down any predictions higher than 5 down to 5. This would minimize our prediction error as well.

V. CONCLUSION

We used multiple linear regression to construct our prediction model and achieved a decent error rate (0.82306 RMSE) in the prediction of the testing data set. We demonstrated that with some carefully chosen data extraction and transformation methods, we can largely improve the prediction accuracy of the linear regression model.

All members in our group have actively contributed to the project. Jin and Deng are mainly responsible for the code while Deng and Gao are mainly responsible for the PowerPoint and the summary report.

Peer Rating:

	Average ratings within group
Linchen Deng	5
Kenny Jin	5
Runxin Gao	5

References:

Koujalagi, A. (2015). Determine word relevance in document queries using TF-IDF. *International Journal of Scientific Research*, 284-285. doi: 10.15373/22778179.

Shakarad, G. (2011). A prospective study of stemming algorithms for web text mining. *Ganpat University Journal of Engineering & Technology*, 28-34. Retrieved from https://www.researchgate.net/publication/50359272_A_Prospective_Study_of_Stemming_Algorithms_for_Web_Text_Mining

The enormous influence of online reviews. (2018, May 2). Retrieved from https://theweek.com/articles/770712/enormous-influence-online-reviews