Yelp Customer Review Bias Analysis through Linear Mixed Effect Models with Natural Language Sentiment Polarity Scores

Alice Ni and Lavanya Sunder
Department of Statistical Science
Duke University
Durham, North Carolina

Yelp is a primary source of restaurant reviews and information nationwide, with restaurants often depending on positive Yelp reviews for sustained business. However, there is speculation that the star rating in a Yelp review is influenced by bias and factors outside of the actual dining experience of a restaurant. In this paper we attempt to un- Yelp the reviews to find the true score (instead of a rating of one to five) of a given review. We attempt to do this by clustering customer reviews of restaurants by sentiment magnitude, as a proxy for clustering restaurants by similar overall quality. This procedure involves calculating the positive and negative sentiment of reviews, by using multiple lexicons and boosting our approach. We then created multiple linear mixed effects models on the clusters in order to isolate the biasing from number of restaurant factors.

Usage: For users to unbias reviews on restaurants as well as for Yelp to find interesting trends within user reviews.

I. INTRODUCTION

Yelp is one of the primary food recommendation sources, used by over 80 million unique users a month to find local restaurants. However, many people believe that there is inherent bias in a Yelp review, that not all Yelp ratings which are "equal" are indicative of the same restaurant caliber. Many also believe that users will tend to rate a restaurant either higher or lower than they would another restaurant that they liked equally, due to some of the restaurant's attributes (the type of cuisine, the number of ratings, the price, etc.).

In order to try to normalize for this bias, we try to "unbias" the Yelp review for a user. Our approach in doing this was to look deeper into the polarity of the review, which we believed indicated the true sentiment of the review better than just the the user-given star rating. We first cluster the reviews by polarity scores to create groups of similar sentiment. We then fit a mixed effect model within each cluster to see the factors that influence the difference in user-given star rating for similar polarity-scored reviews. We treat this difference in the coeffiecients of the fixed effects in the mixed effect model as the bias due to those covariates. Because we find bias at a review level, to find the bias in one restaurant, we treat the conglomeration of all the differences in the reviews for one restaurant as the bias in that restaurant.

II. DATA

We used the data from the Yelp Dataset challenge, specifically the Review and Business Datasets. While the business dataset originally had 61184 records and the reviews dataset originally had 1569264 reviews, we ended up subsetting the data to a working business dataset of 7323 restaurants and a corresponding reviews dataset of 704214 records.

A. Variables

1. Restaurant Dataset

In the end, we decided to include 16 variables from the restaurant dataset in our model. These variables included attributes of the restaurant (delivery, takeout, reservations, credit card, price), as well as meals of the day the restaurant served (breakfast, brunch, lunch, dinner, dessert, latenight). Other variables used were a transformation of some of the variables in the restaurant dataset and are found in Table 1 below.

Table 1: Transformed Variables

Variable from Transformation	Column Transformed On	Transformation
Main Category	categories	Chose the first informative category
Restaurant Weekday hours	hours	Added all the hours the restaurant was open on Weekdays (MTWThF)
Restaurant Weekend hours	hours	Added all the hours the restaurant was open on Weekends (SaSu)
Median Income	address	Found zip code then found median income from the zip code
Mean Income	address	Found zip code then found median income from the zip code
Population	address	Found zip code then found population from the zip code

2. Reviews Dataset

The variables used from the reviews dataset were the business id (used to link the review dataset and the

restaurant dataset), the user-given star rating (used as the response variable in our Mixed Effects Model), and the text (used to find sentiment and create clusters of similar reviews).

B. Subsetting the data

As stated before, we subsetted the data through a series of conditions. These conditions are listed below.

1. The business had to be a restaurant. We determined this by seeing if "Restaurants" was in the categories attribute of the business.
2. The restaurant had to be a currently open restaurant.
3. The restaurant had to have a value for any of the attributes we used in the model.
4. The restaurant had to have values for hours open.
5. The restaurant had to have one informative category (must have had a category that wasn't just "Restaurants" or "Food")
6. The business had to have a zip code. The restaurants that didn't have zip codes were primarily food trucks.

The reviews were subsetted based on the subsetted restaurant dataset.

III. METHODS

We planned to find bias within reviews through two main processes: Sentiment analysis and mixed effects models. We clustered our reviews by a weighted sentiment score, then within each cluster, fit a mixed effect model to see the relationship between user-given star ratings and variables of interest, accounting for the random variability given for a particular restaurant. In this process, we assumed that restaurant quality could be isolated through sentiment analysis of the user reviews and that fixed-effect variables have biasing effects on star rating.

A. Sentiment Analysis

In order to find bias, the first step was to cluster reviews by their true sentimental meaning. By doing this, we could further examine how factors affected user ratings within similarly sentiment reviews; in other words, what accounts for the difference in user-given ratings between reviews which have the same polarity. We used a weighted average of three different methods to come up with our final polarity score, then clustered them into five clusters, each with around 140,000 reviews. In all three methods we used essentially the same preprocessing. For each review, the text was broken down into words using a Corpus, stemmed, had stop words removed, had whitespace stripped, and had punctuation removed. Then, we gave the review a score by the number of "positive words" found in each review over the number of "negative words". The variability in our three methods comes from the different lexicons utilized.

1. Lexicon 1 & Lexicon 2

The method used with Lexicon 1 and Lexicon 2 was a simple ratio of the amount of positive words in a review over the number of negative words in a review. [1][2] Thus, the polarity score of a review with words x_i, using Lexicon 1 and 2, is equal to:

$$\sum(\mathbb{1}(x_i \in P))/\sum(\mathbb{1}(x_i \in N))$$

where:

$$P : \{\text{All positive words in the lexicon}\}$$

$$N : \{\text{All negative words in the lexicon}\}$$

The varying lexicons provided fairly different scores. Lexicon 1 contained 2006 positive words and 4783 negative words, while Lexicon 2 contained 2718 positive words and 4910 negative words. The correlation between Lexicon 1 score and review rating was $r \approx .38$ and the correlation between Lexicon 2 score and review rating was $r \approx .26$.

2. Polarity with qdap

The score that this method returned was weighted the greatest for the final weighted average of sentiment. This is due to the fact that it seemed the most comprehensive of all of the methods. For each polarized word found in the sentiment dictionary, the value of that word, w, found in the sentiment dictionary is additionally weighted by the surrounding words. Essentially, for each word, a context cluster (x_i^T) is pulled from around the word to be used as valence shifters. The words in this context cluster are tagged as neutral (x_i^0), negator (x_i^N), amplifier (x_i^A), or de-amplifier (x_i^D). [3]. The polarity score of that context cluster is then equal to:

$$\sum((1 + c * (x_i^A - x_i^D)) * w(-1)^{(\sum x_i^N)})$$

where

$$x_i^A = (w_{neg} * x_i^A)$$

$$x_i^D = max(x_i^{D'}, -1)$$

$$x_i^{D'} = \sum(-w_{neg} * x_i^A + x_i^D)$$

$$w_{neg} = (\sum x_i^N) \mod 2$$

The overall polarity score then given to the word, C, then becomes:

$$C = x_i^T / \sqrt{n}$$

where n is the number of words in the context cluster. [3] For each review, the review becomes an average of the polarity scores for each context cluster surrounding each polarizing word in the review. The correlation between qdap score and review rating was $r \approx .45$.

3. Normalization and Weighted Average

The three scores produced from Lexicon 1, Lexicon 2, and qdap were normalized in order to average the scores. The normalizing equation used was:

$$m_{ij} = x_{ij} - \min(\boldsymbol{x_j}) / \max(\boldsymbol{x_j}) - \min(\boldsymbol{x_j})$$

where $j \in \{1, 2, 3\}$ represents the different scoring method used. Then, a weighted average sentiment score w_i was created by:

$$w_i = (1 * m_{i1} + 1 * m_{i2} + 3 * m_{i3})/5$$

where $j = 1$ represents scores from Lexicon 1, $j = 2$ represents scores from Lexicon 2, and $j = 3$ represents the qdap score.

B. Mixed Effects Model

We essentially used mixed effect models in order to give meaningful estimates of our covariate coefficients in the presence of effects that added random variation (specific restaurants).

The mixed effects model used is a linear regression model including both fixed effects and random effects. Fixed effects are variables of interest that have levels which is often repeatable. Random effects, on the other hand, are thought of as a small, random selection from a much larger set and is a source of random variation.

In a normal linear regression model, the response variable Y is multivariate normal, however in a mixed effects model, the response variable (in our case, the number of stars given to a review by the user), conditioned on our random effects, is multivariate normal, with the form,

$$(Y|\boldsymbol{\Gamma} = \boldsymbol{\gamma}) = N(\boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\epsilon}, \sigma^2 \boldsymbol{W^{-1}})$$

where $\boldsymbol{X}$ is the matrix of the covariates, $\boldsymbol{W}$ is a diagonal matrix of known prior weights, $\boldsymbol{\beta}$ is a coefficient vector, Z is the model matrix for the random effects variable $\boldsymbol{\Gamma}$, whose value we're fixing at $\boldsymbol{\gamma}$. [4]
The distribution of $\boldsymbol{\Gamma}$ is a multivariate normal, with mean zero and a parameterized $\boldsymbol{\Sigma}$, the variance-covariance matrix of the random effects.

The coefficient estimates for the fixed effects in $\boldsymbol{\beta}$ can be interpreted the same way as an ANOVA, regression, or ANCOVA, but realize that coefficient represents the mean over all subjects (or restaurants in our case), and each individual subject (restaurant) has their own personal value for that coefficient. Random effect estimates in $\boldsymbol{Z}$ are variances, interpreted as the magnitude of the variability of "personal" coefficients from the mean fixed effects coefficient. [5]

To find the maximum likelihood estimates of the covariates of interest for our project, we utilize the lmer package in R, fitting a model using both fixed effects and random effects to our data to obtain our coefficient vector β. [4] Model selection was accomplished using step-wise AIC regression.

IV. RESULTS

After completing our sentiment analysis and running different mixed effect models for each cluster, our analysis produced a number of interesting results. First, our sentiment analysis step was somewhat validated by the fact that each of the five models (M_1,M_2...M_5) had varying significant variables, as well as significant differences in coefficient values.

Table 2: Significant Variables in Each Model

Variable	M_1	M_2	M_3	M_4	M_5
Intercept	X	X	X	X	X
Review Count	X	X	X	X	X
Weekday Hours		X	X	X	X
Weekend Hours	X	X	X	X	X
Credit Card	X	X			
Price	X	X	X	X	X
Delivery	X	X			X
Reservations	X	X	X		
Dessert				X	X
Latenight					X
Lunch	X	X	X	X	X
Dinner	X	X		X	X
Breakfast				X	X
Brunch				X	X
Main Cat*	X	X	X	X	X
Median Inc	X	X	X	X	X
Mean Inc					X
Population		X	X		X

This table shows all of the significant variables in the 5 linear mixed effect models. Note*:Main Cat actually represents more than 30 dummy variables for restaurant categories, and significant categories differ between models.

As Table 2 shows, while there were some variables that were significant across the board (Review Count, Median Income, etc.), most variables were only significant for certain models. Additionally, though the table does not indicate this, the Main Categories that were significant differed between models (for instance, Southern is significant for all models but M_1).

Our analysis also uncovered a number of interesting relationships between our covariates and review bias. We defined review bias as:

$$b_i = m_i - r_i$$

where r_i is the original rating of the review, and m_i is the model-produced rating of the review. We found that, in general, restaurants with a larger number of reviews experienced less overall bias. Additionally, as Table 3 shows, the number of hours that restaurants were open on the weekend was significant and had a negative coefficient in all 5 models, indicating that the more a restaurant is open on the weekend, the more artificially inflated its ratings are.

Table 3: Comparative View of Significant Variables

Variable	M_1	M_2	M_3	M_4	M_5
Intercept	3.532	4.076	4.223	4.612	4.768
Median Income	$-2.37*10^{-6}$	$-7.00*10^{-7}$	$-5.85*10^{-8}$	$5.61*10^{-7}$	$3.04*10^{-6}$
Review Count	$6.91*10^{-4}$	$4.98*10^{-4}$	$3.06*10^{-4}$	$2.09*10^{-4}$	$1.30*10^{-4}$
Price	$-.124$	$-.0777$	$-.0366$	$-.0211$	$-.0154$
Weekend Hours	$-1.48*10^{-2}$	$-9.50*10^{-3}$	$-7.76*10^{-3}$	$-7.05*10^{-3}$	$-5.77*10^{-3}$

This table shows the estimates of some particularly significant variables in the 5 linear mixed effect models.

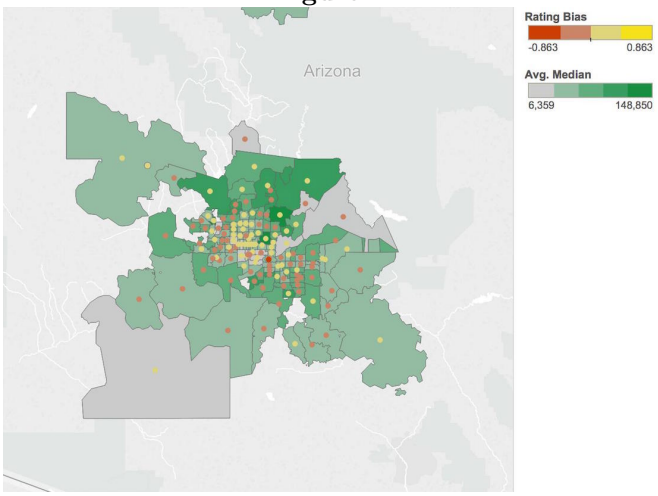
Like Weekend Hours, we found that Price had a negative coefficient in all five models (even considering a 95% CI). However, the impact of price is more severe on reviews in more "negative" clusters than on more "positive" clusters. The interpretation is such that Price will have a negative effect regardless, but the magnitude of this effect will be more severe if the rating itself is more severe.

We also found that, median income had a positive coefficient in M_4 and M_5, and a negative coefficient in M_1, M_2, and M_3 (even considering a 95% CI). This implies, as shown in Figure 1, that restaurants in affluent neighborhoods experience a negative bias in reviews.

With regards to "meal" offerings (i.e. Lunch, Dinner, etc.) and other details (Delivery, Reservations, Credit Card), the results were not particularly interesting, particularly given the fact that different models had vastly different significant variables.

Finally, with regards to Main Categories, the categories themselves were significant. However, it is difficult to interpret the coefficients, as the number of dummy variables Main Categories produced was > 30 for all models.

Figure 1



The above map shows the data available from the greater metropolitan area of Phoenix. The shading is average median income by zip code, and the points are calculated bias from restaurants in that area. As shown, there appears to be an positive bias (i.e. the new rating is smaller) from larger zip codes, and a negative bias (i.e. the new rating is larger) from more affluent neighborhoods.

V. CONCLUSION

Our examination of Yelp customer reviews, and factors influencing ratings bias, largely showed that there is in fact a measurable bias in Yelp ratings. Particularly, the Median Income of the surrounding area, Price rating of the restaurant, and number of hours the restaurant is open on the Weekend were significant in determining the bias in customer rating.

A. Limitations

We encountered a number of limitations throughout the data analysis. Firstly, we encountered issues with processing power, when dealing with the size of the data set. For instance, we had to abandon one Lexicon scoring method (not mentioned above) because the processing time was too big.

Additionally, we dealt with issues during our natural language processing. Dealing with sarcasm and other language deviations was difficult, and there was no real solution found during the sentiment analysis process. We also simply had trouble finding a sentiment analysis process that was as accurate as we hoped; the Lexicon and

qdap methods, while effective, could certainly be improved.

Finally, there were some limitations during model selection, particularly when dealing with Main Category. While there were some categories identified as non-significant during the step-wise AIC process, overall the process of dealing with Main Category was not optimal.

B. Future Work

For future analyses, we would like to work on making our sentiment analysis process even more accurate, in order to better form the initial clusters for the mixed effect models. Additionally, we would like to come up with a better way to deal with the Main Categories, like potentially creating (or working with Yelp to create) larger Main Categories so that interpretation and analysis is easier.

We would also ideally like to create a user-interface for Yelp (or Yelp's customers) to "un-Yelp" customer reviews directly from the website. This could provide customers with another rating that is perhaps more accurately representative of how good a restaurant is.

[1] B. Liu and M. Hu, "Opinion Mining, Sentiment Analysis, and Opinion Spam Detection," (2004).
[2] T. Wilson, J. Wiebe, , and P. Hoffmann, University of Pittsburgh , 1 (2005).
[3] T. W. Rinker, *qdap: Quantitative Discourse Analysis Package*, University at Buffalo/SUNY, Buffalo, New York (2013), 2.2.4.
[4] D. Bates, M. Mächler, B. Bolker, and S. Walker, Journal of Statistical Software **67**, 1 (2015).
[5] H. J. Seltman, "Experimental Design and Analysis," (2015).