**Introduction**

Yelp and other review websites offer users the ability to submit reviews which include a number and text. Companies operating these websites, market researchers, and academics all may be interested in drawing connections between the text of a review and its numerical rating, in order to gain inference on various questions varying from the operational: “what distinguishes between good reviews and bad reviews?” to the theoretical: “what are the linguistic signifiers of dissatisfaction in the Midwestern United States?”

Using a training sample of approximately 57,000 Yelp reviews from various types of Wisconsin businesses, we aimed to accurately predict star rating, using the text of reviews as our starting point. The model we eventually explored is grounded in two hypotheses. First, we imagined that we could extract a ‘word score’ measuring the sentiment attached to a word by conducting an SLR with star as response for each of the ~48,000 unique words found in the review text we were given, and selecting highly significant results (more on this below). Second, we imagined that both intensifiers (words that emphasize other words, such as ‘really,’ or ‘very’) and the words that follow them could be useful in predicting star rating.

**Background Information**

Our training data came from 57,000 Yelp reviews from Wisconsin, which appeared to originate in multiple industries. After training our model on these reviews, we then validated the model using 38,000 additional Yelp reviews. The original data was not entirely smooth; for example, some reviews lacked a zip code. Yelp ‘star ratings’ are on a 5-point scale, and are restricted to non-zero integer values (one cannot give a 4.5 star rating).

The vast majority of reviews in our training sample were positive: approximately 65% of reviews in our sample awarded either 4 or 5 stars, while only 19% gave 1 or 2 stars. The mean star-rating was 3.76, the median rating was 4, and the mode was 5. We observed that longer reviews were generally associated with lower ratings, before considering other factors, and that some zip codes were significant predictors of star ratings.

**Motivation for Analytics Used**

We first constructed exploratory models using the statistics provided with the data, and struggled to produce a model with adjusted R² > 0.50. These initial models seemed to have low accuracy, given the amount of information contained in the text, but also perhaps reflected the difficulty of analyzing reviews from nail salons, hospitals, and restaurants with the same toolset. We imagined that the words people use to describe positive hospital experiences would be different from those used to describe positive restaurant experiences. With this acknowledged, words on almost every category need to be included in order to improve the accuracy of the final model.

We noticed that our initial models easily distinguished between positive (4 or 5 star) from negative (1, 2, or 3 star) reviews, but struggled to distinguish reviews within each category. As a result, the bulk of our effort consisted in searching for information that could help us distinguish between 1-star reviews and 2-star reviews, and between 4-star and 5-star reviews. We noticed early on, in reading the text of reviews, that both highly positive reviews and highly negative reviews made use of intensifiers (‘very’, ‘really’, etc), but followed them with different words. We also noticed that usually the word modified by the intensifier followed within two words of the intensifier itself.

**Brute force repeated SLR used to measure sentiment: ‘Word Scoring’**

In searching for a numeric representation of the sentiment carried by various words, we first identified every unique word contained in the reviews we were provided (throwing out words like “the” or “in”). Next, we performed an SLR, with star rating as response, and a binary predictor (1 if a review contained the word, 0 if not), using each of these 48,000 unqiue words. We then saved the p-values and coefficients resulting from each of these 48,000 SLRs. This took about 4 hours of (overnight) computer time to run on the training dataset of 57,000 reviews, but we were hopeful that the result could be useful.

After this repeated SLR finished, we then took all words whose SLR produced a p-value < 0.01, and retained their slope coefficients (representing the difference in mean star-rating for reviews containing and not containing the word ‘incredible,’ for example). We used this as our numeric ‘word score,’ a number hopefully capturing some aspect of the sentiment associated with that word.

Finally, for each review in the dataset, we counted how many words were in our list of positive words (p-value < 0.01 and coefficient > 0 in SLR) and negative words, and for each review summed the positive word scores, summed the negative word scores, took the total sum, and transformed each of these measures with sign-adjusted square root and ln(x+1) functions. We imagined that each additional positive or negative word would have a diminishing effect.

**Intensifier Parsing**

We also identified a list of intensifiers in English, and parsed the review texts to find every instance of them, recording the number of intensifiers found in each review. Next, we pulled out all of the words that fell one or two words after an intensifier. Finally, we used the ‘word scores’ from above (the SLR coefficients for words with SLR p-value < 0.01) to assign another numeric score to each review, based solely on the words *following intensifiers* in each review.

**Layman's Interpretation of the Estimates and Inferential Quantities**

Our final model contains 152 predictors, but can be boiled down to a set of three core predictors which perform the bulk of the model’s predictive work. These three predictors produce an adjusted R² of 0.60 on training data, while our final model has adjusted R² of 0.68.

In the model, represents the ratings (number of stars) of each Yelp review, is the square-root of the sum of positive ‘word scores’ (see above) found in the review, is the natural log of the sum of negative ‘word scores’ found in the review, and represents the ratio of negative words to overall words in the review.

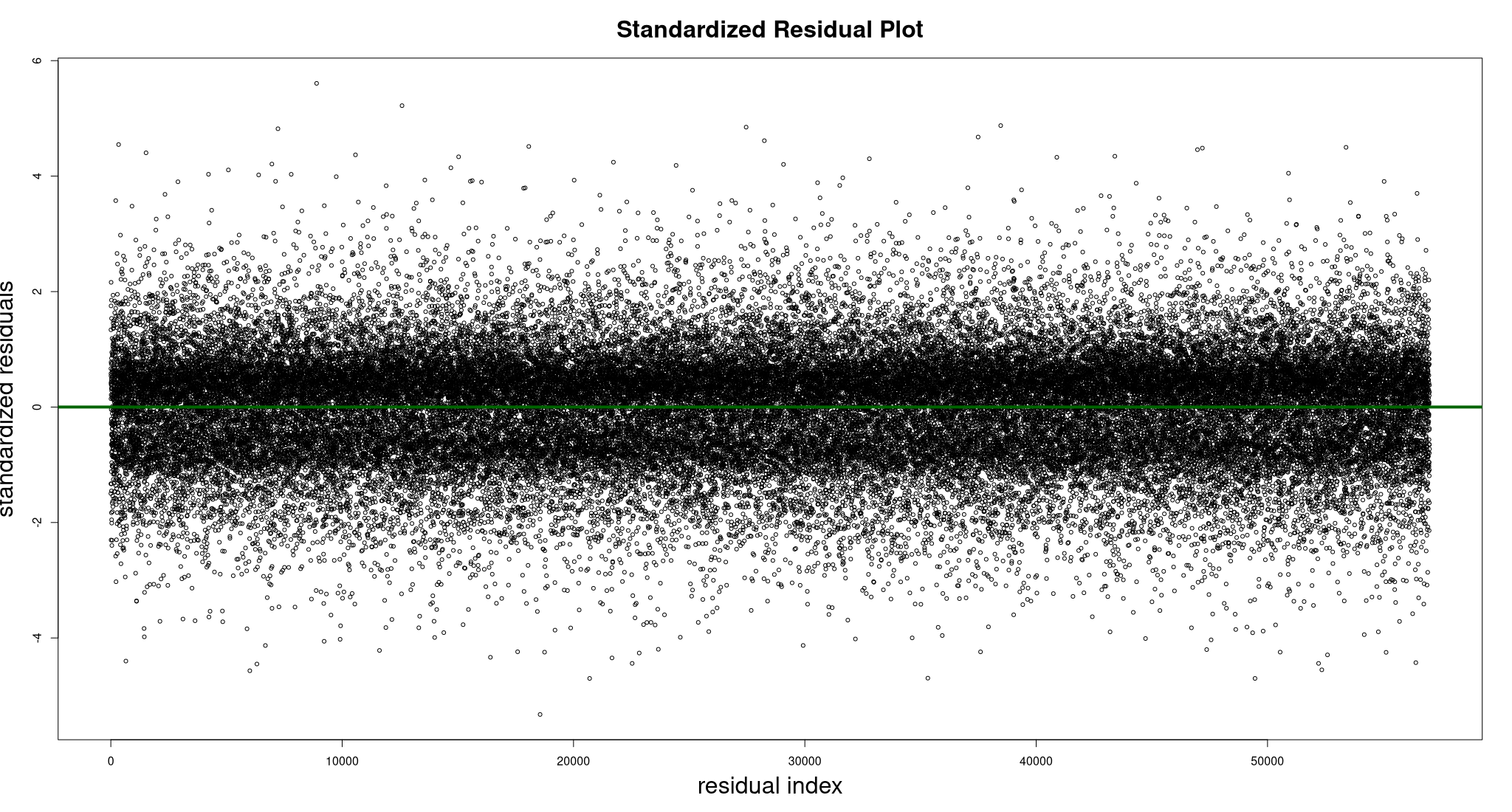
The other 149 predictors include word counts for words we found significant, dummy variables representing zip codes with a significant effect on ratings, once other factors had been considered, and the two-way interactions we found to be significant.

**Model Selection and Assessment**

We initially assembled every predictor with a p-value < 0.05, and then used a ‘backwards’ selection process, using both AIC and BIC as our criteria to arrive at our final model. The backwards AIC process removed only one predictor; backwards BIC removed a number of zip code, word-count, and two-way-interaction predictors, but resulted in a marginally lower adjusted R2 and validation results, so we chose the model resulting from backwards AIC as our best predictive model, suggested for cases where interpretability and parsimony are secondary to raw predictive accuracy (as, for example, in trying to top a Kaggle board).

Below are several model assessment statistics evaluated for a model using only grand mean as predictor, our ‘Lite Model’ (described above, using our 3 best predictors), and the full model we chose.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean-only | Lite Model | Full Model |
| Adjusted R2 | N/A | 0.5988 | 0.6803 |
| Cross-validation | 1.735 | 0.696 | 0.557 |
| BIC | 139,226.2 | 141,188.4 | 129,943.1 |
| AIC | 193,208.3 | 141,143.7 | 128,366.2 |

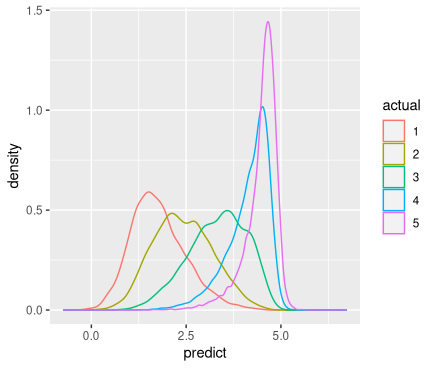


**Model Diagnostics**

Since the measured response was discrete, we did not assess it for normality. We felt it was reasonable to assume normality in the error term suggested (see QQ plot in presentation). In the residual plot above, it appears that there may be a significant band of residuals > 0, though there does not appear to be a curve or a fanning shape to the residuals, so we feel confident believing our assumptions of normality and homoskedasticity have been reasonably met.

**Model Strengths and Weaknesses**

Our model with 152 predictors achieved the highest adjusted R2 value among models we considered (0.6802), and the lowest CV score, BIC, and AIC. The model’s corresponding weakness lies in the number and complexity of its predictors, and perhaps also in the extent of collinearity between different predictors. When we filtered out predictors with high VIF (at least 5), we saw adjusted R2drop significantly (about 0.13) as a result.



As one can observe in the density plot of our model’s predictions, broken out by actual rating (to left), our model still struggles to draw fine lines between 1 and 2-star reviews, or 4-and 5-star reviews, but does a decent job distinguishing between 1 and 3-star reviews, and even 3 and 4-star reviews.

Since the data are mainly collected in Wisconsin, the model may not be as precise when predicting reviews outside Wisconsin, and especially outside the US due to possible differences in word choice.

**Conclusion**

While we remain somewhat uncertain about the efficacy of our brute force ‘word scoring’ technique, our model performed reasonably well on its validation test (it was ranked #2 on the Kaggle scoreboard when we submitted our project). We were interested to notice that positive word scores appeared most significant when square root transformed, while negative word scores appeared most significant when ln(x+1) transformed.

According to our model, the ratio of negative words, and the overall sentiment suggested by the words included in the review together do the best job predicting the actual rating. The rest of the 148 predictors we included only contributed the last 7-10% of the model’s accuracy.

We would be curious to attempt to validate our brute-force SLR method on larger data, as well as data from a different region.

**Contribution**

All three members participated substantially to this project; we wrote the paper and created the presentation as a team. Caleb focused on creating our summary statistics; Jiayi came up with the idea to examine intensifiers and contributed throughout; and Ruochong focused on model selection.