Predicting Star Ratings in Yelp Reviews based on the Frequency of Positive and Negative Sentiment Words in Reviews

I. Introduction

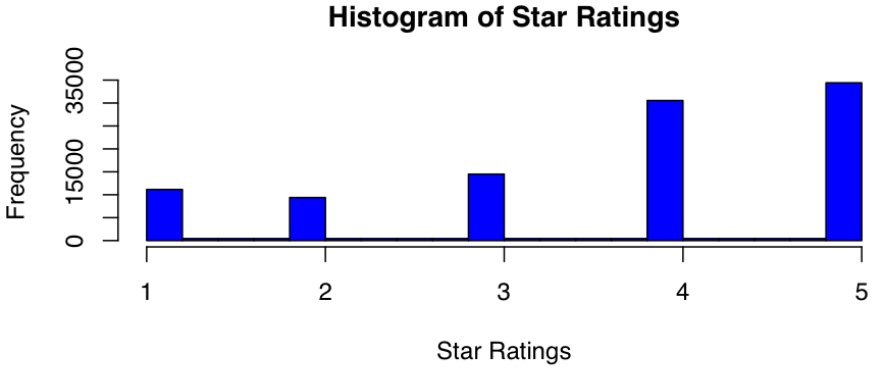
Sentiment analysis has been gaining increasing acceptance as research conducted in the field has resulted in the development of widely-cited sentiment lexicons. To better understand the linguistic basis for social review systems, we decided to use two different sentiment lexicons to address the question: How well does the frequency of occurrence of positive sentiment words and negative sentiment words within a Yelp review predict the star rating of that review? First, we used the widely-cited list of positive and negative sentiment words in English compiled by Minqing Hu and Bing Liu (around 6800 words), which can be downloaded along with their research papers from http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. Further research uncovered another frequently-cited sentiment lexicon developed by Finn Arup Nielsen, which assigns positive and negative valences to words, and can be found at http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010.

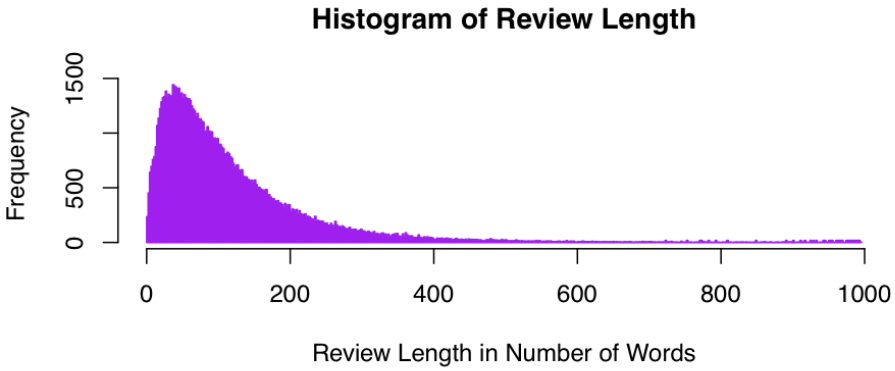
We performed exploratory data analysis with both sentiment lexicons to determine the number of positive and negative words in reviews. Word counts were normalized by dividing by the number of words in a review so frequencies weren't unduly biased based on review length. Then we developed linear regression models using the frequency of positive and negative words detected in a review based on the two sentiment lexicons. Summary statistics were examined to determine the best model fit, concluding that the best model uses all four of the Hu & Liu and Nielsen normalized positive and negative word counts. For the final composite model, each of the four predictors was found to make a significant contribution.

II. Methods and Data

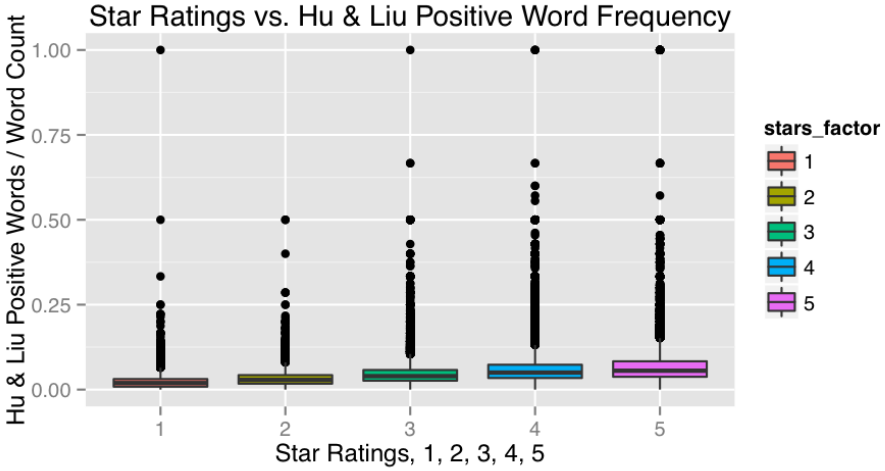
A. Data Processing First, the jsonlite, stringi, ggplot2 and qdap libraries were loaded and the jsonlite streaming function was used to read the Yelp review dataset (in JSON format) into a data frame. A new data frame was constructed with star ratings and review text along with columns for review word count, positive and negative word counts and normalized positive and negative word counts (divided by review word count) for the Hu & Liu and Nielsen lists. The Hu & Liu positive and negative word lists were read in as text files, and the Nielsen word list was divided into a positive word list for words with valences between 1 and 5, and a negative word list for words with valences between -5 and -1.

B. Exploratory Data Analysis To get a feel for the data, we computed the mean number of stars to be 3.67687 and plotted a histogram of star ratings. Next we computed the mean length of reviews to be 116.98306 and plotted a histogram of the review length. Performing exploratory data analysis to directly interrogate the question of interest, we then counted the frequency of occurrence of positive and negative words for the Hu & Liu and Nielsen sentiment lexicons and made boxplots showing star ratings vs. positive and negative word frequencies in reviews.

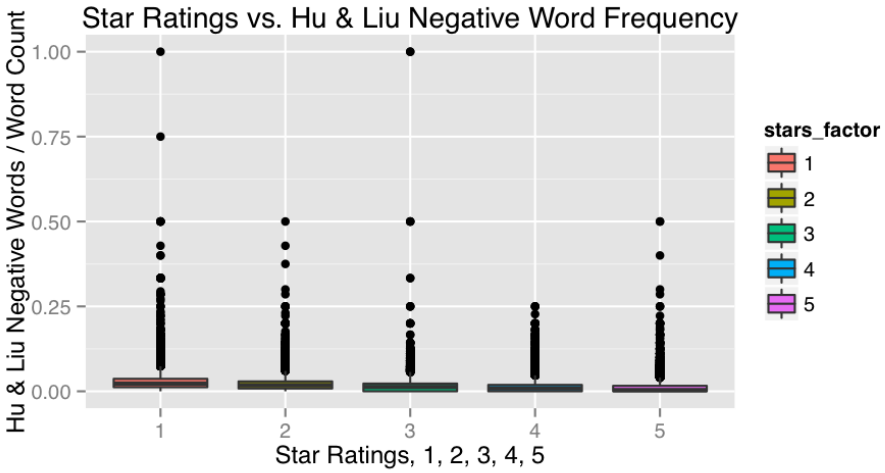




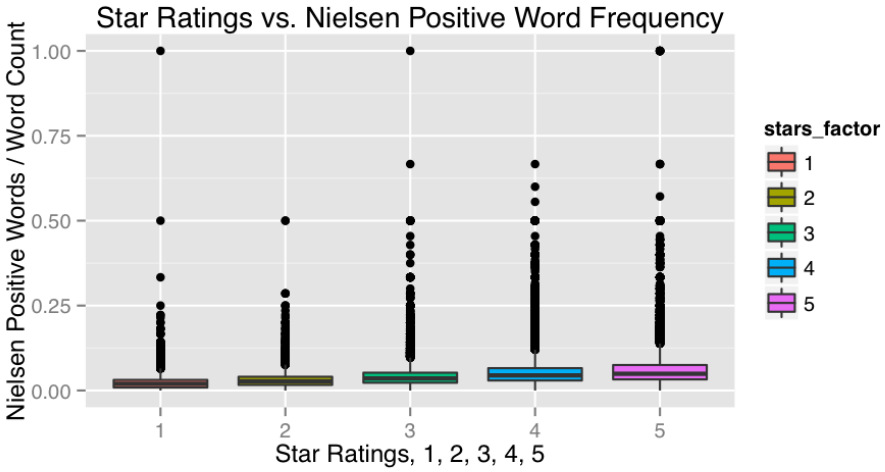
An interesting challenge arose in the process of counting the positive and negative sentiment words. Each sentiment lexicon contains sentiment words expressed as different parts of speech, such as “dirt” and “dirty,” and “dangerous” and “dangerously.” In addition, shorter unrelated words might be found inside longer words, such as “anger” inside “danger” or “sly” inside “dangerously.” Therefore, in one instance we found a two-word review “dangerously dirty!” which was initially computed to have six negative words in a two-word review. To address this problem, we pasted a blank space on either end of each word in the sentiment lexicons as well as on either end of each word in the review text before searching for matches. We also normalized the frequency by dividing by the number of words in the review in order to prevent longer reviews from having undue influence.



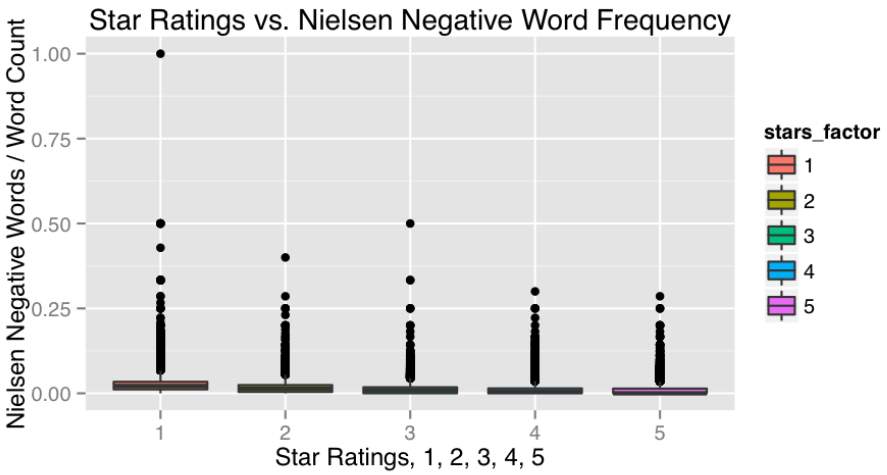
Looking at the boxplot with positive sentiment words from the Hu & Liu sentiment lexicon, we see an upward trend in the frequency of positive sentiment words as the star rating increases. Observing instances of outlier values of 1.0, we did further exploration of the data and found some one-word reviews, for example “awesome” and “good”, which were associated with a star rating of 5, and “great”, which, in one case, was associated with a star rating of 1. Since these reviews consisted of one word and the single word in each of these reviews was considered to be positive, the normalized frequency of positive words in these reviews would be equal to 1.0.



Looking at the boxplot with negative sentiment words from the Hu & Liu sentiment lexicon, we see a downward trend in the frequency of negative sentiment words as the star rating increases. Further analysis of the upward trend in the frequency of positive sentiment words and the downward trend in the frequency of negative sentiment words as the star rating increases will be performed in the process of constructing and analyzing linear models.



Again, we see an upward trend in the frequency of positive sentiment words as the star rating increases in the boxplot showing positive sentiment words from the Nielsen lexicon.



Lastly, we see a downward trend in the frequency of negative sentiment words as the star rating increases in the boxplot showing negative sentiment words from the Nielsen lexicon. We will do further analysis to determine whether these observations based on the boxplots are significant.

C. Model Construction To determine how well the frequency of occurrence of positive and negative sentiment words in a Yelp review predicts a review's star rating, we used linear modeling. We constructed four separate linear models with each of the Hu & Liu and Nielsen normalized positive and negative word counts as the sole predictor and star ratings as the outcome:

- lm(formula = stars ~ hl_norm_pos, data = df2)
- lm(formula = stars ~ hl_norm_neg, data = df2)
- lm(formula = stars ~ n_norm_pos, data = df2)
- lm(formula = stars ~ n_norm_neg, data = df2)

Then we constructed a new model using both the Hu & Liu positive and negative word counts as predictors and star ratings as outcome and a second combined model using both the Nielsen positive and negative word counts as predictors and star ratings as outcome:

- lm(formula = stars ~ hl_norm_pos + hl_norm_neg, data = df2)
- lm(formula = stars ~ n_norm_pos + n_norm_neg, data = df2)

The final model we constructed uses all four Hu & Liu and Nielsen positive and negative word counts as predictors and star ratings as outcome:

- lm(formula = stars ~ hl_norm_pos + hl_norm_neg + n_norm_pos + n_norm_neg, data = df2)

D. Linear Model Summary Statistics

In the Results section, we will consider the adjusted R-squared value of the predictors to assess how well the frequency of occurrence of positive and negative sentiment words predicts the star rating. For the final model, we perform an analysis of variance to compute a p-value to determine if we can reject the null hypothesis that the predictor variables do not contribute to the model fit. In addition, we examine a normal Q-Q plot of residuals for the final model.

III. Results

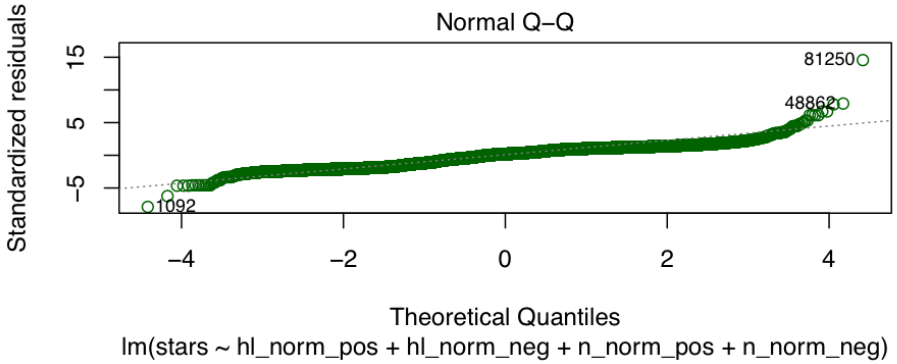
The adjusted R-squared value indicates how much of the variability in star ratings can be explained using the predictor variable(s) in that model. We consider the adjusted R-squared value rather than the R-squared value since the adjusted R-squared value takes into account the number of predictor variables.

A. Adjusted R-squared value The adjusted R-squared value of the model using only the Hu & Liu positive word frequency equals: 0.1, indicating that 10% of the variability in star ratings can be explained using the Hu & Liu positive word frequency. The adjusted R-squared value of the model using only the Hu & Liu negative word frequency equals: 0.07, indicating that 7% of the variability in star ratings can be explained using the Hu and Liu negative word frequency. The adjusted R-squared value of the model using only the Nielsen positive word frequency equals: 0.08, indicating that 8% of the variability in star ratings can be explained using the Nielsen positive word frequency. The adjusted R-squared value of the model using only the Nielsen negative word frequency equals: 0.08, indicating that 8% of the variability in star ratings can be explained using the Nielsen negative word frequency. The adjusted R-squared value of the model using both the Hu & Liu positive and negative word frequencies equals: 0.15, indicating that 15% of the variability in star ratings can be explained using a combination of the Hu & Liu positive and negative word frequencies. The adjusted R-squared value of the model using both the Nielsen positive and negative word frequencies equals: 0.14, indicating that 14% of the variability in star ratings can be explained using a combination of the Nielsen positive and negative word frequencies.

Addressing the question of how well the frequency of occurrence of positive and negative sentiment words in a Yelp review predicts the star rating of the review, the adjusted R-squared value of the final model using both the Hu & Liu and the Nielsen positive and negative word frequencies equals: 0.16 indicating that 16% of the variability in star ratings can in fact be explained using a combination of all four predictor variables: the Hu & Liu and the Nielsen positive and negative word frequencies.

B. Analysis of variance to compute p-value of final model We use analysis of variance and compute the p-value of the final model to be 0, which is very low, indicating that we can reject the null hypothesis that the final model predictor variables do not contribute to the model fit.

C. Residuals with Normal Q-Q Plot Lastly, a Normal Q-Q (quantile-quantile) plot of Residuals was examined for the final model, with all four of Hu & Liu and Nielsen normalized positive and negative word counts as predictors. With the points generally falling on or close to the line in the Normal Q-Q plot, we see that the residuals are approximately following a normal distribution. This indicates that information not being captured by the final model is more or less normally distributed random noise.



IV. Discussion

It is interesting to note that the model with positive and negative word counts based on the Hu & Liu sentiment lexicon has an adjusted R-squared value of 0.15 which is higher than the adjusted R-squared value of 0.14 for the model with positive and negative word counts based on the Nielsen sentiment lexicon. This means that 15% of the variability in star ratings can be explained using the Hu & Liu positive and negative word frequency predictors, while 14% of the variability in star ratings can be explained using the Nielsen positive and negative word frequency predictors. However, since there were differences between the sentiment lexicons, the best model was based on using a combination of the frequency of positive and negative sentiment words determined by the Hu & Liu lexicon and the Nielsen lexicon, which has an adjusted R-squared value of 0.16, indicating that 16% of the variability in star ratings can be explained using a combination of Hu & Liui positive and negative word frequency and the Nielsen positive and negative word frequency. Furthermore, the very low p-value of 0, indicates that we can reject the null hypothesis that the final model predictor variables do not contribute to the model fit. Therefore, the final linear model using all four predictor variables of the Hu & Liu positive and negative word frequencies and the Nielsen positive and negative word frequencies demonstrates how well the frequency of positive and negative sentiment words in reviews can predict their star ratings, answering the question posed in this project.