A Sentiment Analysis of Yelp Reviews

1. The Project

1.1 The Problems Motivating the Project

For restaurants to thrive as businesses it is important to know whether their customers have a positive or negative experience. There are a number of ways restaurant owners might acquire this knowledge, but a readily available method is to read reviews that customers write about their experiences. A central problem with this method, however, is that it involves conflicting incentives for restaurant owners. On the one hand, reviews are a valuable source of information about customer experiences, so restaurant owners should want more customers to write more reviews. On the other hand, reading reviews can be very time-consuming, and the more reviews there are, the more time it takes to gain value from them. Furthermore, it can be difficult to gain an overall sense of customer experiences simply by reading a set of reviews. For example, reading seven shorter positive reviews in combination with a single extremely long and negative review might skew the impression someone gains of overall customer experience.

1.2 The Goal of the Project

The goal of this project is to provide a more efficient means of discriminating between positive and negative reviews. In particular, the goal is to construct models that can reliably classify reviews as positive or negative based on the content of the reviews. The models will be trained from reviews that have labels of positive or negative, which are derived from the star ratings associated with the reviews. Having models that can do this not only removes the problem of needing to read through each review just to get some sense of customer experiences, but it also provides a way to get a clearer view of overall customer experience. For example, after classifying a set of one-hundred reviews, even before reading any of the reviews, a restaurant owner could easily figure out whether a majority of the reviews are positive or negative, and if so, how much of a majority.

2. The Current State of the Project

Up to this point, I have acquired and cleaned the data, explored the data, and applied some statistical analysis to the data. In order to build the models, I am using a dataset of approximately 6.5 million Yelp reviews.[1] To prepare the data for this project, I completed the

[1] The dataset was acquired here: https://www.yelp.com/dataset.

following four steps.[2] First, since the project is focused only on restaurant reviews, I removed reviews for businesses that are not restaurants. Second, I isolated the columns of the dataset that were most important for the present project: 'business_id', which is a unique identifier for each restaurant, 'stars', which includes the star ratings associated with the reviews, and 'text', which includes the text of the reviews. Third, I removed rows in the dataset that were either duplicates or empty. Finally, upon finding that some of the reviews were not written in English, and since analyzing non-English reviews is beyond the scope of the project, I removed any non-English reviews from the dataset.

After performing these cleaning and preparation steps, I explored a number of interesting features of the data.[3] In particular, I analyzed properties of the star rating data, the review data, and then the interactions between those two sets of data. Regarding the star rating data, I investigated the total number of each star rating, the proportion of each star rating out of all ratings, the average star rating for each restaurant, and the distribution of ratings for some specific restaurants that have multiple reviews. For the review data, I explored the number of reviews that each restaurant has received and the distribution of lengths for the reviews. Finally, in order to explore how the star rating data interacts with the review data, I identified the relationships between the different star ratings and lengths of reviews, and then I created word clouds to depict words that occur frequently in the reviews at each star rating level.

One of the more interesting findings from this exploration was that the average review length (measured in terms of mean character count) differed across each of the star ratings. Since this feature could potentially be used in the classification models, I decided to use hypothesis testing to determine whether these differences in average length were statistically significant. First, I performed a one way analysis of variance (ANOVA) to test whether there were any statistically significant differences between the mean lengths of reviews for the different star ratings. The null hypothesis was that the mean review lengths are the same for each star rating, and the alternative hypothesis was that at least one of the mean review lengths differs from another. The test produced an F-statistic of 36.22 with a p-value approximating 0, which provided strong reason to reject the null hypothesis and conclude that at least one of the mean review lengths differs from another.

Star Rating	Mean Length
1.0	669
2.0	734
3.0	702
4.0	616
5.0	469

[2] The notebook in which I clean the data can be found here: https://github.com/walkerpage/springboard/blob/master/Capstone_Projects/Capstone_Project_1/Code/Capstone_Project_1_Data_Wrangling.ipynb

[3] The notebooks in which I explore an analyze the data can be found at these two links: https://github.com/walkerpage/springboard/blob/master/Capstone_Projects/Capstone_Project_1/Code/Capstone_Project_1_Data_Storytelling.ipynb, and https://github.com/walkerpage/springboard/blob/master/Capstone_Projects/Capstone_Project_1/Code/Capstone_Project_1_Statistical_Data_Analysis.ipynb

Since the ANOVA test can only tell *that* one of the mean lengths is significantly different from another mean length, and not *which one* is significantly different, I subsequently conducted several t-tests (using the Bonferroni correction to control for familywise error) between pairs of star ratings to determine which differences in mean length were statistically significant. In particular, because when I train the models, I will associate star ratings of 1, 2, or 3 with negative sentiment, and star ratings of 4 or 5 with positive sentiment, the most important comparisons for my purposes was between these two groups of ratings. Accordingly, a performed t-tests of the difference in mean review length between the following six pairs of star ratings: 1 & 4, 1 & 5, 2 & 4, 2 & 5, 3 & 4, 3 & 5. The t-tests revealed that the differences between nearly all of these pairs were statistically significant. The only difference that was not statistically significant was between the 1 and 4 star reviews. The analysis suggests, therefore, that I may be able to use the lengths of reviews to help classify them as positive or negative when I develop the models.

3. Next Steps

The next step of the project will be to begin constructing the classification models. I am approaching this project as a supervised classification problem, where the labels are derived from the number of stars associated with reviews. I will divide the Yelp dataset into training and test datasets and then develop baseline models using logistic regression and Multinomial Naive Bayes. After constructing these models, I will evaluate whether constructing other models might improve upon the baseline. Finally, I will write up a final report on the project and create a slide deck for presenting the project.