Leanna Hue, Casey Poon
Professor Hu
CS 396 - Introduction to the Data Science Pipeline
01/3/2020

Part 2 Summary

One application of machine learning we applied was gathering the average sentiment based on tips.json for each unique business id at different meal times. Our initial approach was to apply a bag of words technique by getting the count and map to positive words in each review in tips.json and then averaging them for each business id. However, this proved to be successful due to the fact the reviews with more words skewed higher in general sentiment. Our second approach to extract average sentiment was to use VADER sentiment analysis to extract a polarity score for positive sentiment (https://github.com/cjhutto/vaderSentiment). We chose this approach because the VADER's sentiment analysis model was specifically trained on social media text dataset, and Yelp's reviews and comments are very similar to how people post social media. Although sentiment analysis is considered an unsupervised learning problem, we were able to evaluate the VADER model's accuracy by examining average star rating for each business. Typically reviews with higher star ratings would have more positive sentiments in their reviews so we looked to see if there was a positive correlation/relationship between a restaurant's star rating and average positive sentiment. In order to evaluate this we decided to examine the Pearson and Spearman correlation coefficients between average sentiment and star rating. In the subset examining only Fast Food, we discovered that Spearman and Pearson correlation

coefficients were .41 and .39 respectively. From this we were able to conclude that although there is some sort of positive relationship, there isn't a very strong correlation. However, further analysis or preprocessing might be needed to determine whether or not we could use the VADER model for our project.

The goal of our project is to explore the relationship between food cuisines (Fast Food, American (Traditional), American (New)) and general consumer rating for a business. Although yelp does include tip rating and business rating in the dataset, examining the actual text content of reviews might help us gain more insight on the relationship we're trying to explore.

Initially we limited the amount of text processing we applied to the dataset. We considered \ removing punctuation in our text processing; however, after reading VADER's documentation we realized removing punctuation would actually decrease the accuracy of the 'positivity score' since the model includes punctuation and variations of words to further sway its perceived sentiment. After seeing the relationship between star rating and average sentiment, some future text processing we might include is the correction of spelling mistakes and lemmatization to hopefully reduce the noise in the sentiment scoring.

Aside from using sentiment analysis, we also engineer a dataset that extracted boolean values that were mapped to 0 or 1 for attributes such whether or not a business was 'Romantic', 'Classy', 'Hipster' etc. Each business entry is then mapped to a classification of cuisine from original business.json. We plan on using this dataset of

vector representation for businesses for future analysis and machine learning to potentially experiment with supervised learning models.