Rohini Shrivastava

IST 736

Homework 7

**Introduction:**

Restaurants and other dining places have been around for centuries, with the oldest record being in 12th century China. People were able to go to various shops to try different types of food. Slowly over time other countries such as Japan and France also had small shops that would serve food. In France, usually these shops would not have a menu, but instead provide only one type of food that the local chef would prepare. The English word “restaurant” come from the French verb *restaurer*, which means “to restore oneself”.

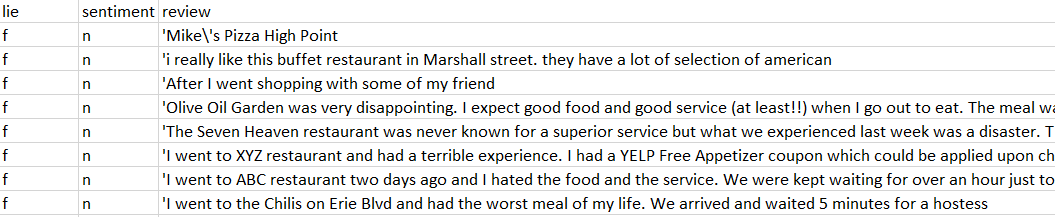
Since restaurants have become so popular, there needs to be a way to distinguish which places are worthwhile visiting. In the 19th century the first restaurant guidebook was published in France. This encourages fellow “foodies” to visit the best local restaurants and encouraged them to travel as well.  Almanacs have slowly faded with the rise of newspapers. Food critics were able to reach a broader audience through these newspapers. With the increased use of technology, it is now easy to go online and find details about restaurants. Websites such as yelp allow users to write about their experiences at these restaurants.

Yelp allows users to rate the entire experience on a scale of one to five stars, with one star being a negative review. It can further be broken down into experience of wait staff and experience of eating the food. Users are able to describe their entire experience. They can discuss how to food tasted to how long the wait was. These websites also give information on estimated wait time based off trends. These types of websites have changed the review industry and allow those who are not specialized in food to speak about how they feel going to a restaurant in an anonymous way. By allowing virtually anyone to see and write the reviews, it is very important for these restaurants to keep their high quality of food and service.

**Analysis:**

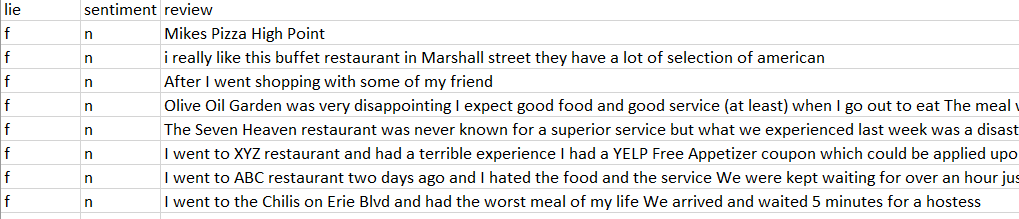
About The Data:

The set of data was provided as a Comma Separated Value (CSV) file and as a Text Separated Data (TSV). The file contained information about restaurants and their reviews (*Figure 1).* This csv contained the sentiment (whether its positive or negative), whether it was true, and the reviews associated with that restaurant.



*Figure 1: Restaurant Reviews Before Pre-Processing*

For the data, a script was run to create a new text file and then clean up the data. The punctuation, such as “.”,”?”,”’”,”/” etc, were removed (*Figure 2).*



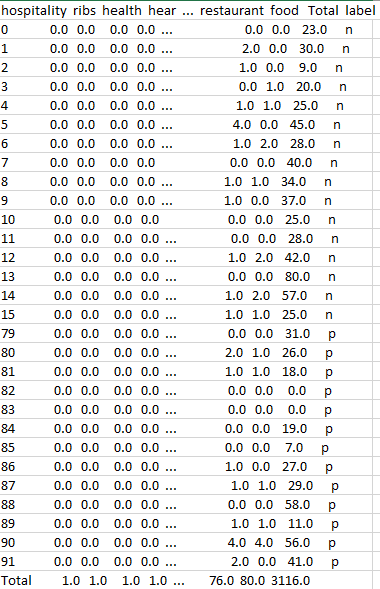
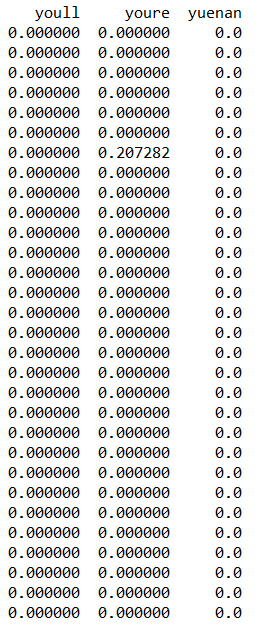
*Figure 2: Restaurant Reviews After Pre-Processing*

CountVectorizer/Tf-IDF:

SKLearn is a python programming library that is used for classification, regression, and clustering algorithms. It uses both supervised and unsupervised algorithms. Inside there is a module called CountVectorizer. CountVectorizer converts documents into a matrix of tokens and their counts. The Tf-IDFVectorizer lists the frequencies of each word in the overall document.

CountVectorizer was used to create a matrix of the TSV file. The file were imported through the OS library, which looks at the operating system interface, and then cleaned up. The CountVectorizer function was used on the files, and then was converted into a data frame. For the Bernoulli model, CountVectorizer had to be set to a binary value.

The data frame had labels of the TSV as the columns. Once CountVectorizer had been run, a total was then calculated for all the columns and all the rows. Each row contained the number of times the word from the column was present in the review. The column total calculated the total of words used in each review, and the row total calculated the number of times each word was present in the entire corpus. Then columns were sorted based on row totals in increasing sequence (*Figure 3)*. The restaurant reviews had 92 rows and 1256 columns.

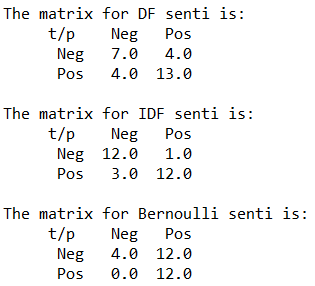
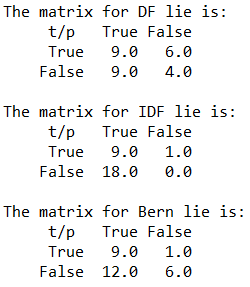


*Figure 3: Sample of Vectorized Data with CountVectorizer and TfIDFVectorizer*

NBM/Bernoulli:

The data was run against the Naive Bayes’ Multinomial and Bernoulli model. This model is a popular way to analyze categorical data, specifically text data. It is used to classify the data. Bernoulli Naive Bayes only takes in binary values.

Once the data had been vectorized, a test and training set was created. The labels were then removed as the multinomial and Bernoulli model cannot run with the label there. Once the models were run, the output was compared against the actual results. A confusion matrix was created for the sentiment (*Figure 4)* and for the authenticity (*Figure 5).*

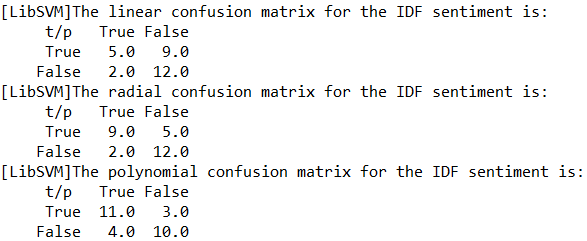
 

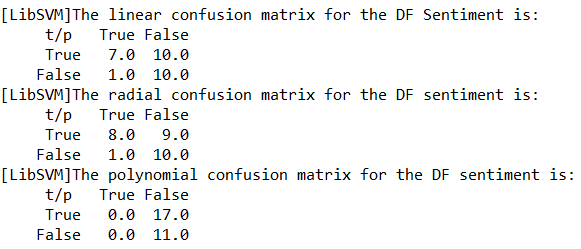
*Figure 4: Confusion Matrix for Sentiment Figure 5: Confusion Matrix for Lies*

Support Vector Machines:

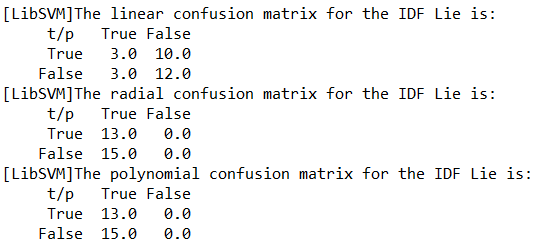
Support Vector Machine (SVM) are supervised learning models that analyze data for classification and regression. It looks at different types of data such as text and image data. The data set is split into test and training sets and run with different kernels. The polynomial, linear, and radial kernels were used.

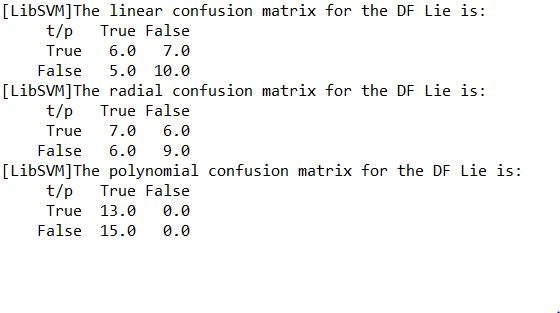
After the data was vectorized and separated into test and training sets, the labels were removed and output were compared with the predicted and actual results. A confusion matrix was created for the sentiment (Figure 6) and the authenticity (Figure 7).





*Figure 6: Confusion Matrix of SVM for Sentiment*





*Figure 7: Confusion Matrix of SVM for Lies*

**Results:**

All the reviews combined contained 3116 words, not including stop-words. The top used words for restaurant reviews were “food” and “restaurant”. The least used words were “hospitality”, “ribs” and “health”. These top words follow in line with what is typically expected to hear about a food service. “Food” may be relating to what was eaten and the quality of it, and “restaurant” referring to the place the customer ate.

The multinomial model was more accurate when determining the sentiment over the authenticity. When the model was ran for sentiment, the model was 71.40% accurate. When it was run for authenticity it was 46.42% accurate. The multinomial model with the TD-IDF vectorization was 84.2% accurate with the authenticity being 32.14%. The last model, the Bernoulli model was 58.43% accurate for sentiment and 52.19% for authenticity (*Figure 8/Figure 9)*. For sentiment, the IDF model had the highest prediction, but for authenticity the Bernoulli Model had the highest. The Bernoulli model stayed around the same percentage when getting the accuracy of sentiment and authenticity. The other two models saw an average of 20% decrease in accuracy. It was harder for the model to determine authenticity based on the words of the text document, showing that machines cannot accurately determine whether a review is true or not just based off the words written.

*Figure 8: Accuracy of the Sentiment Models*

*Figure 9: Accuracy of the Sentiment Models*

When running the SVM models, a similar trend was found of the sentiment accuracy being higher than finding the authenticity of the reviews (*Figure 10)*. The sentiment IDF vectorizer had the highest average accuracy for all three SVM models at 78.3%. The least accurate vectorizer was the IDF for authenticity at 41.2% (*Figure 11)*. This is similar to what was found with the Naïve Bayes models. The multinomial model for sentiment using IDF was the highest, with authenticity using IDF as the lowest. The multinomial Naïve Bayes model for sentiment was 5.9% greater than the average SVM. For authenticity, the multinomial model was about 9% less than the SVM.

*Figure 10: Accuracy per Data for Each Model*

*Figure 11: Overall Average Accuracy for each Vectorizer*

In the three models, the linear and radial models had almost the same overall accuracy. The polynomial model was the least accurate for both sentiment and lies (*Figure 12).* Both linear and radial models were 75% accurate on sentiment. However, the linear SVM was 4% more accurate than the radial SVM.

*Figure 12: Average accuracy per model*

Overall, the linear SVM is the most accurate SVM model. The IDF Vectorizer was better than the DF Vectorizer, but not by much. Both the Naïve Bayes and SVM model were about the same in average accuracy, but the multinomial model was more accurate than both the linear and radial SVM models.

**Conclusion:**

It is very important to know the authenticity of the review. These unauthentic reviews can cause the restaurant to lose their credibility. It can also cause rival restaurants to gain more traction in their business. The opinion of the review also can impact the number of customers visiting the shop. Negative reviews can cause people to avoid the store, while positive reviews can increase the number of customers that may visit.

92 reviews were collected about various restaurants. These reviews were both positive and negative, and included information about the authenticity of the review. In terms of looking at just the most common words, they were mostly neutral. Words like “restaurant”, “food”, and “dining” were common. These are all neutral words that do not really showcase what the overall sentiment of all the reviews are.

Using the data, three different models were run to determine if the sentiment or authenticity could be determined. The models were more accurate when predicting the sentiment of the review rather than the authenticity. While predicting the sentiment is vital, determining the authenticity is equally, if not more, important. The model needs to be fine tuned to be more accurate and not showing as many false-negatives.