Timothy Chu
Sam Lakes
CPE582

Restaurant Project Report

Approaches:

Exercise 1: Binary Score of Paragraph

Ultimately, we ended up using 1 main feature with the Naive Bayes Classifier, which was the word count for all the good and bad words. The good words were found by finding the most common words in all of the "good" reviews (score of 4-5) and the bad words were found by finding the most common words in all of the "bad" reviews (score of 1-3). The feature then indicated "more_good_words" if the count of good words exceeded bad words or "more_bad_words" if the count of bad words exceeded good words. This feature gave us around a 6% accuracy over the baseline of 50%, which would have been random chance of predicting the paragraph to be good or bad with zero knowledge. There was around a 4 to 1 chance that a "more_bad_words" indicated a bad review and around a 3 to 1 chance that a "more_good_words" indicated a good review.

Many other features were tested as well, but with no effects to the accuracy. One notable feature was attempting to utilize NLTK's corpus of movie reviews and trying to draw parallels between the corpus and our data. By finding the most common words in the good reviews and the most common words in the bad reviews, we expected to find an overwhelming number of words that would have described a good or bad movie exclusively. However, similar words were intermixed between the good and bad words. Words like "good," "great," and "bad" were found in both good and bad reviews. Note that stopwords were removed from this selection. We also tried looking specifically at adjectives, but this also had no noticeable effects on the accuracy.

Another big feature that we tried to use was the senti_synset corpus provided by NLTK. We tried counting the positive and negative scores of the words in the paragraph. This saw no change in accuracy above what we already had with the word counting of good and bad words. We tried restricting the scope to adjectives for this approach as well, but there was no difference.

Overall, many variations of the described approaches did not see much of a difference. We believe this to be due to the fact that those features held no significant correlation. In doing this exercise, it was surprising to see that many patterns that we might expect to exist did not exist at all.

Average Accuracy = 0.572
Average Precision for Bad Rating = 0.5265
Average Recall for Bad Rating = 0.51275
Average F-measure for Bad Rating = 0.52
Average Precision for Good Rating = 0.6008
Average Recall for Good Rating = 0.6752
Average F-measure for Good Rating = 0.6186
Average RMSE = 0.638

Exercise 2: Interesting Features

For this exercise, we picked out things that we thought would have signified a significant correlation in the data sets. We chose to examine the ratio of adjectives to nouns, common words in good and bad reviews, and similar words in the reviews with overall ratings of 4 and 2.

In examining the ratio of adjectives to nouns, we hypothesized that bad reviews would undoubtedly be more expressive. Thus, we expected more adjectives. However, when looking at the class data, the ratios were pretty much similar. For good reviews, the ratio was around 6 nouns to 1 adjective. For bad reviews, the ratio was around 6.5 nouns to 1 adjective. Looking at this data, we concluded that there was no significant difference in expressivity for good or bad reviews.

When finding the unique common words in good and bad reviews, we expected to see words that had obvious connotations. For example, we thought we'd see things like "great," "incredible," "amazing," and "fantastic" for the good reviews, and "terrible," "disgusting," "gross," and "dislike." However, the actual results showed less promising answers. Instead, we found that many types of foods were grouped in the two categories (like "burritos," "chicken," "mediterranean," "grill," "sandwich," etc). We thought that this perhaps showed a correlation between what types of food people thought were good/bad, or that our data set was simply too small. We tried to extend this to NLTK's corpus of movie reviews (which were excluded from the submission due to timing issues), but even with those, there was no obviously connotated words.

Lastly, when looking at similar words in the reviews with overall ratings of 4 and 2, we hypothesized that the words would represent two polar opinions. However, this exploration found that this was not true.

Exercise 3: Predicting Overall Rating of Review

For this exercise, our main feature in successfully determining the overall rating of the review was to use the average of the Food, Service, and Venue Scores. This feature gave us around a 45% accuracy over the baseline of 20%, which would have been random chance of predicting the review to have an overall score of 1 to 5. There was around a 12 to 1 chance that an averaged score of 3 indicated an overall score of 3 compared to a 4, a 4 to 1 chance that an averaged score of 5 indicated an overall score of 5 compared to a 4, and a 2 to 1 chance that an averaged score of 4 indicated an overall score of 4 compared to a 3.

We tried other techniques to try and improve the accuracy, but those chosen features didn't significantly improve the results. One thing we tried to do was using the classifier from exercise 1 to predict the scores of the paragraphs, and using those score to predict whether it would be a good (4-5) or bad review (1-3). Admittedly due to our lower accuracy for exercise 1, the results saw no major correlations. We also tried many of the techniques used in exercise 1, but applying them to the whole review rather than paragraphs. We thought that with more data, it would have potentially saw more use, but it surprisingly also had no results.

Average Accuracy = 0.69
Average RMSE = 0.532

Exercise 4: Authorship Attribution

For this we ended up trying a variety of different features, however none of them seemed to be useful in increasing the accuracy. We tried looking at average sentence length, top 30 words, top 30 bigrams (For these two features we tried to see if the words in the top 30 for features were in the top features of the corpus, which in this case was all of the words/bigrams inside in the reviews), average paragraph length, and lexical diversity.

The best accuracy we could get was 25% with the provided example data and a 15% with the class data. It seemed that lexical diversity was the only feature that had a say in the accuracy unfortunately. Everything else we tried either had no effect or lowered the accuracy to a 3.7%. Notably, this 15% accuracy for the class data is still better than the baseline of 3.3%, which would have been randomly selecting a student from those in our class.

Average Accuracy = 0.146
Average RMSE = 0.916

	Adam Calabrigo	Aditya Budhwar	Alanna Buss	Brandon Cooper	Brandon Livitski	Christian Durst	Cody Hunt	Daniel Kauffman	Gavin Scott	Ivan Pachev	Jeffrey McGovern	Jeremy Kerfs	Joel Dentici	Jonathan Sleep	Justin Postigo	Kishan Patel	Logan Williams	Michael Williams	Miguel Aguilar	Nicole Martin	Ryan Gelston	Ryan Smith	Sage Maxwell	Samuel Lakes	Sean Bayley	Timothy Chu	Tobias Bleisch	Vivian Fong
Adam Calabrigo	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3.6%
Aditya Budhwar	.	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Alanna Buss	.	.	<.>	.	.	.	3.6%	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Brandon Cooper	.	.	.	<.>	.	.	.	.	3.6%	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Brandon Livitski	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3.6%
Christian Durst	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3.6%
Cody Hunt	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3.6%
Daniel Kauffman	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3.6%
Gavin Scott	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3.6%
Ivan Pachev	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3.6%	.
Jeffrey McGovern	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3.6%	.
Jeremy Kerfs	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3.6%
Joel Dentici	.	.	.	.	.	.	3.6%	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Jonathan Sleep	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Justin Postigo	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	.	3.6%	.
Kishan Patel	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.	3.6%	.
Logan Williams	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	3.6%	.
Michael Williams	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.	.	.
Miguel Aguilar	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	3.6%	.
Nicole Martin	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	.	.	.
Ryan Gelston	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	3.6%	.
Ryan Smith	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	.	3.6%
Sage Maxwell	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	.	3.6%
Samuel Lakes	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	.	3.6%
Sean Bayley	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	.	3.6%
Timothy Chu	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<.>	.	3.6%
Tobias Bleisch	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<3.6%>	.
Vivian Fong	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	<3.6%>

(row = reference; col = test)

Confusion Matrix:

We have a bit of a weird phenomena for when we do a confusion matrix on the authorship prediction. Most of the predictions seem to be people later in the alphabet. If you look at the matrix you can see Vivian Fong is predicted a lot, as are Tobias, Sam, and a few others. A lot of this has to do with the fact that we had trouble finding relevant features.

Since we did not find enough relevant features for our classifier to look at when predicting the author of a review, it is likely that the classifier ended up with predicting the same people over and over again because they were at the top of the list for the categories of the features. Our most relevant feature was lexical diversity, so it may have been that many people fell into the same categories. And when predicting, the classifier chose the same person over and over again when looking into that specific category.