Prediction Model

Training

The main objective of my model was to predict whether the user rating was positive or negative based on the user text of a given review. In doing so, I began by adding an additional variable “Positive” in the data frame that was a “1” for a review greater than 3 and a “0” for a review of 3 or less. This allowed me to use logistic regression in the standard fashion, with Tf-Idf, and with N-grams to determine whether each of these prediction methods could most accurately determine the user’s rating based on the user’s text. 75% of the data was used for feature selection and 25% of the data was used for the test set.

Methods

Logistic Regression

For the prediction method, I simply used Logistic regression and compared the three types to determine the most accurate predictor. The logistic regression was used because although the restaurant ratings were numerical, they were categorical and therefore a better method than other approaches.

Results

The Logistic regression with the highest accuracy based on an AUC was N-grams which makes sense given the greater number of features as compared to the standard and Tf-Idf type.

|  |  |
| --- | --- |
| **Logistic Regression** | |
| **Type** | **Accuracy** |
| **Standard** | 71.80% |
| **Tf-Idf** | 61.53% |
| **N-grams** | 72.78% |
| **Average** | 68.70% |