Nick Videtti

nvidetti@syr.edu

Assignment 4

Multinomial Naive Bayes for Text Data

Pandas DataFrames, SciKit Learn’s CountVectorizer, SciKitLearn’s MultinomialNB

5/4/2023

| Restaurant Reviews and Their Validity | |
| --- | --- |
| **Introduction** | Restaurants are a big part of most social lives. They are often where friends go to meet up after a long week of work, where families go to experience something new, where couples go for a romantic date night, or where a bachelor gets his favorite meal for takeout on Tuesday nights. Regardless of the person or lifestyle, food is an integral piece of American culture, and the restaurant is at the forefront of food culture. This makes a restaurant’s reputation as valuable as any factor towards its success. Having a strong and consistent customer base with a large number of “regulars” will make any restaurant successful, regardless of even something as important as the restaurant’s quality.  All that said, new customers need to come from somewhere, and the reputation of a restaurant often drives that through reviews. It is extremely common to look up a restaurant online before deciding to try it out for the first time, and usually an exceptional rating with mostly exceptional reviews are what it takes to convince prospective customers to go to that new restaurant rather than choosing another tried and tested option.  Sadly, with the continued advancement of technology and the “good old fashioned” phenomenon of people not being truthful, many reviews are fake and/or false. This can unfortunately make a good restaurant fail over time by bringing down its prestige and hurting its reputation.  Putting all of the pieces together, this study will try to find exactly how to identify reviews by whether they are positive or negative and whether they are truthful or fallacy. It will also be attempted to identify both positive/negative sentiment and truthfulness at the same time. 92 restaurant reviews were supplied with their corresponding sentiment (positive or negative) and truth (true or false) and will be used to conduct some experiments. Part of these reviews will be used to make predictions based on the words that are in each review, and the remaining reviews will be used to see how accurate those predictions are. |
| **Analysis** | These analyses will be done using SciKit Learn’s Multinomial Naive Bayes classifier (sklearn.naive\_bayes.MultinomialNB()), but numeric data are needed in order to do that. This will be addressed by converting each review into a count vector using SciKit Learn’s Count Vectorizer (CountVectorizer() from sklearn.feature\_extraction.text).  Before that can be done, the first step is to pull in the data, which comes in the form of a CSV file. This file contains three fields: one for truth called “lie” (either t or f), one for sentiment (either n or p), and the text for the review itself. These data will be read into a Pandas DataFrame. After that, four more columns will be created: lie\_sentiment, words, cleaned\_words, and cleaned\_review. “lie\_sentiment” will be “lie” concatenated with “sentiment” (will take values of either “tp”, “tn”, “fp”, or “fn”). “words” will be a list of the words from the review, generated using NLTK’s word tokenizer. “cleaned\_words” will be “words” with all word phrases converted to lowercase. Also, each list in “cleaned\_words” will contain only word phrases with exclusively letters and will exclude any word phrases that fall into NLTK’s English stopwords. Lastly, “cleaned\_review” will be the cleaned words with each list being combined into a single string. This will be the version of the review that is used for the remainder of these analyses.  Once the “master” DataFrame just described is created, the next step is to prepare data for the Multinomial Naive Bayes classifier. This will require subsets of the “master” DataFrame for each analysis, one for “lie”, one for “sentiment”, and one for “lie\_sentiment”. Each of these will contain its corresponding label with the cleaned reviews from the “cleaned\_review” column. For each of these subsets, they will need to be converted into count vector DataFrames using a combination of Pandas and SciKit Learn’s Count Vectorizer.  The last step before finally running the models will be to split each of the 3 count vector DataFrames into 3 separate DataFrames: one for training data, one for testing data (without labels), and one with the testing data labels that can be used to test model accuracy.  The three models will then be trained and confusion matrices will be generated, which will be used to evaluate the accuracy of the models. |
| **results** | The results of the data preprocessing, count vectorization, and multinomial naive bayes models were as follows.  **“Master” DataFrame - First 10 rows and First 4 columns**    **“Master” DataFrame - First 10 rows of “words”**    **“Master” DataFrame - First 10 rows of “cleaned\_words”**    **“Master” DataFrame - First 10 rows of “cleaned\_review”**    **“lie” DataFrame - First 10 rows**    **“sentiment” DataFrame - First 10 rows**    **“lie\_sentiment” DataFrame - First 10 rows**    **“lie” Model - Sample of Training and Testing Data**    **“sentiment” Model - Sample of Training and Testing Data**    **“lie\_sentiment” Model - Sample of Training and Testing Data**    **Model Confusion Matrices with Accuracy** |
| **conclusions** | The results of this specific study suggest that it is actually quite difficult to tell whether a review is truthful or not based on the words in the reviews. It was however much easier to tell whether the review was positive or negative. As far as predicting both at the same time, it was somewhere in between. The actual accuracy of these three sets of predictions were as follows.   | **Truth or Fallacy** | 10 out of 28 correct | 35.7% accurate | | --- | --- | --- | | **Positive or Negative** | 23 out of 28 correct | 82.1% accurate | | **Both Concurrently** | 15 out of 28 correct | 53.6% accurate |   With an accuracy of 35.7%, these results suggest that it is likely to be more accurate flipping a coin to predict whether a review is truthful or not rather than using the method used in this study. However, the tables were turned when predicting whether a review is positive or negative, and the method used in this study resulted in 82.1% correct predictions. It is intuitive that predicting both factors concurrently has an accuracy roughly half way between the difference of the two independent accuracies, but it is still only at 53.7%. While it might make sense at first to bring back the coin flipping analogy, there are four different outcomes in predicting both concurrently rather than the binary outcomes of the other two sets of predictions. That being said, this suggests that the method used to predict both concurrently is actually much better than guessing, rolling a four-sided die, drawing a card from a deck and using the “suit”, or any other method of randomly choosing one item from a group of four. |