Ayush Malakar

Mike Wright

Text Mining

Final Project

**Introduction**

People have always had opinions and feedback on everything ranging from experiences to the purchase of goods and services. There is an entire evolution as to the way people have expressed feedback and opinions ranging from word of mouth, newspaper, radio and with the latest invention in the past 30 years now the internet. With instant access to millions of products and services it is as important now as it was ever to get feedback on what the consumer is really paying for. Over the last 20 years Amazon has transformed from selling books to becoming the world’s largest online retailer. Amazon’s sales have only increased in the last three years especially with many people isolating during the pandemic. This has called for an increase in customer reviews, as gone are the days of going to a brick-and-mortar store to see and feel products in person.

**Statement of Work**

For this project the analyst’s goal is to identify sentiment from Amazon reviews. Both analysts plan on accomplishing this task by a combination of using modeling techniques learned throughout this course, programming techniques learned in other courses, and use of open-source code repositories. The analysts will use two sets of data that will be outlined that will be used as a training and test data set. The analysts will use the training data set to train multiple models to conduct a Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Logistic Regression. The purpose of conducting this analysis is as stated earlier to identify sentiment of Amazon products, to identify the best and worst rated products based on sentiment.

**Data**

The data used for analysis was collected from a Kaggle Repository. The repository consists of twoe separate data sets that are both in a .csv format. The first data set to be used by the analysts is Datafiniti\_Amazon\_Consumer\_Reviews\_of\_Amazon\_Products\_May19.csv, this data set consists of 28,332 rows and 24 columns and is used as the training dataset. The other dataset is called 1429\_1.csv and consists of 34,660 rows and 21 columns this dataset is used as the test dataset. For the purposes of the analysis the only variables that will be measured consist of: ‘rating’, ‘asins’, ‘text(reviews)’, and ‘name(product)’.

**Data Cleaning and Pre-Processing**

Since this data came from a Kaggle repository the data was obtained in a ‘clean’ state. That is, it did not require the analyst to perform any transformations or make any changes to the data upon receiving it. The only cleaning conducted was the removal of special characters, application of stopwords, and the lower casing of words after the data was tokenized.

**Exploratory Data Analysis**

Chart, bar chart

Description automatically generatedDuring initial Exploratory Data Analysis (EDA) it was discovered that the training data had more rows than the test data this included: ‘dateAdded’, ‘dateUpdated’, ‘primaryCategories’, and ‘manufacturerNumber’. It is also identified columns such as reviews.id, and reviewsnumbHelp had NAN values. Since none of these variables were used to measure the data, no changes were made, these were more passive observations in the data. To identify sentiment a sentiment column was created known as ‘senti’. To measure sentiment the ‘reviews.rating’ variable was measured, this variable measures star ratings of products on a scale of 1-5. The decision was made that star ratings of 4 or greater was to be interpreted as positive and star ratings of 3 or less was to be interpreted as negative. When identifying how many positive and negative reviews it was discovered that there was a total of 25,545 positive reviews and 2,787 negative reviews. Based on these reviews it was very clear that more than 90% of the data in the Training data is positive.

Chart, bar chart

Description automatically generatedAfter analyzing the Training data, the Test data was analyzed to identify how many positive and negative reviews were in that dataset. The Test data set had a total of 34,660 reviews so it was slightly larger than the Training set. Although slightly larger the Test data almost mirrored the results of the Training data. Of the ~35,000 reviews 32,316 reviews were positive and 2,344 were negative. Moving forward with training the models to properly identify sentiment based on reviews between both data sets, it is known that they have overwhelmingly positive reviews.

**Results**

**Feature Extractor for NLTK Naïve Bayes Classifier**

Prior to conducting any of the analysis outlined above a Feature Extractor is conducted using the NLTK package and Naïve Bayes modeling technique. The purpose of conducting this is to identify the “most informative” features. That is the words with the most weight within the dataset. When interpreting the results of the model it returned a 65.2% accuracy. The following results were obtained: ‘worst’, ‘dies’, ‘ruined’, ‘acid’, and ‘terrible’. Although the model isn’t extremely confident, the results are very interesting and that is because there are less than 3,000 negative reviews in the data and all the top 5 results can be interpreted as negative sentiment.

**Multinomial Naïve Bayes**

The first analysis conducted is the Multinomial Naïve Bayes. This is a probabilistic model that leverages the Bayes theorem to determine the sentiment of the data. When preparing the data for this model Tokens were obtained from the data and all the Tokens were lowercased. This is done to add more structure to the data to ensure all the words were in the same format. Next, the NLTK StopWords were applied. The purpose of applying stopwords is to “filter out the fluff”, the purpose of removing the stopwords is to better allow the model to process the words and provide more accurate sentiment. To provide the best the most accurate results the five different iterations of the model was run. The first iteration of the model measured sentiment on tokens that appeared five or more times in the data and returned a ngram range of 1,1, and this returned an accuracy result of 92.1%. The next iteration measured sentiment with tokens that appeared a minimum of three times and a ngram range of 2,2, and this returned an accuracy result of 91.2%. The third iteration analyzed tokens appearing two or more time with a ngram range of 3,3 and returned an accuracy of 90.9%. The fourth iteration analyzed tokens appearing five or more times and a ngram range of 3,3 which is 92% accurate. The fifth iteration analyzed tokens appearing five or more times with an ngram range of 1 to five, which returned an accuracy of 91.8%.

Based on all the versions of the MNB tests conducted the most accurate was the first iteration where tokens appearing five times or greater were analyzed with stopwords applied and a ngram range of 1,1.

**Bernoulli Naïve Bayes**

The next test conducted is the Bernoulli Naïve Bayes. The same five tests that were conducted with the MNB were used for Bernoulli NB. Furthermore, the data was pre-processed in the same manner, that is tokenization of the data, the tokens were lowercased, and NLTK stopwords were applied. The first test returned an accuracy of 87.9% which is slightly lower than the 92.1% from the MNB test. The second test returned an accuracy of 90.6% which was lower than the MNB which had a score of 91.2%. The third test returned an accuracy of 89.5% which was lower than the MNB score of 90.9%. The fourth and fifth tests both returned an accuracy of 90%. The fourth test scored lower than the MNB of 92%, and the fifth test scored lower than the MNB score of 90.2%. The sixth and final test scored an 89.2% which like the other results scored lower than the MNB of 91.8%.

When analyzing the results of the Bernoulli Naïve Bayes the results were not terrible, however it is very clear when compared to the MNB it is less accurate. An interesting finding is the MNB is more accurate with a ngram range of 1,1. Whereas the Bernoulli NB seemed to score highest when looking at the bigrams (a ngram range of 2,2).

**Logistic Regression**

The last analysis conducted is the Logistic Regression. Prior to conducting this analysis in a similar fashion to the other tests the data is tokenized, and all tokens were lower cased. Next the NLTK package is used to implement stopwords. After conducting the first test it yields an accuracy of 93.8%. The second test returned an accuracy of 95.3%. The third test returned an accuracy of 95.8%. The fourth test returned an accuracy of 96.5%. The fifth test returned an accuracy of 90.2%. The sixth test returned an accuracy of 95.9%.

When interpreting the results of the Logistic Regression out of all three models this is the most accurate, and the most accurate of all the models is test four. This is interesting because test four like the other tests uses tokens that appear five or more times, returns trigrams (ngram range of 3,3). This is interesting because when compared to the other tests it did not score nearly this high.

**Interpreting Results**

Chart, line chart

Description automatically generatedAfter conducting the analysis on the data, a classifier comparison was conducted using the Receiver Operating Curve. The ROC measures the three different analysis and scores them based on the True Positive Rate and the False Positive Rate. This measure returns a score called Area Under Curve or AUC, the best way to interpret this result is the higher AUC score the more accurate the test is. The MNB has an AUC score of 0.91, Bernoulli NBs AUC is 0.86, and finally the Logistic Regressions AUC is 0.97. Based on these results and what is alluded to from the individual testing the Logistic Regression is the most accurate between the models, followed by the MNB, and lastly the Bernoulli NB.

Table

Description automatically generatedNext a closer look at the results was analyzed to identify how well the models did in predicting positive and negative words and analyzing their f1-score and weighted averages. First the MNB is analyzed, and it returned an overall accuracy of 91.8%. This model accurately identified 96% of the positive reviews and 90% of the negative reviews. It returned an overall f1 score of 90% and a weighted average of 91%. Next the Bernoulli NB modes is examined, and it identified 51% of the positive reviews and 92% of the negative reviews. It had an overall f1 score of 90% and a weighted average of 88%. Lastly the Logistic Regression is interpreted, and it returned an overall accuracy of 96%. This model accurately interpreted 88% of the positive reviews and 97% of the negative reviews. It had an overall f1 score of 96% and a weighted average of 96%. Between all the models the MNB did the best at interpreting the positive reviews with 96% and Bernoulli did the worst with 51%. Whereas the Logistic Regression did the best at interpreting the negative reviews with 97% and the MNB performed the worst with 90%.

After identifying what model had performed the best, the analyst wanted to identify what products had the most reviews. To identify what products had the most reviews the data was sorted by the ‘asins’ variable. This revealed there were 24 unique items in the data set with the number of reviews ranging from 797 to 4. For this analysis only the top 5 products were reviewed which ranged with reviews from 797 to 467. To create word clouds the products were identified as the product names and the positive and negative words were obtained from the reviews. The top 5 reviewed products in order from most to least consist of: ‘All-New Fire HD 8 Tablet’, ‘Amazon Echo Show’, ‘Amazon-Echo Plus w/Built-in Hub’, ‘Fire Kids Edition Tablet’, and ‘Brand New Amazon Kindle Fire’. Something interesting about the most reviewed items are all forms of technology and are all Amazon products.

**Word Clouds**

**Negative Reviews**

A picture containing text, newspaper, screenshot

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

A picture containing text, newspaper, screenshot

Description automatically generated

Text

Description automatically generated

**Positive Reviews**

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

**Conclusion**

In conclusion after conducting analysis on the Amazon Reviews data the most accurate modeling technique used was the Logistic Regression, followed by the Multinomial Naïve Bayes, and then the Bernoulli Naïve Bayes. In each of the different analysis there is a common theme, that tests that had tokens appearing five times or greater performed better than tests that had a lower threshold of tokens. Furthermore, models that had a lower ngram range in most cases performed better than when the ngram range is higher. When conducting a sentiment analysis, it is the analyst’s recommendation to train multiple models to interpret results this is because each model interprets the data slightly differently and this produces different accuracy. Moreso, when looking at the results from above it is important to understand how the model is performing and where its shortfalls are, this can potentially be used to retrain models in the future to try and increase accuracy.