**Assignment 5**

**Naïve Bayes Learning**

**Submitted By: Neel Vijay Patil (11436004)**

**Submitted To: Dr. Junhua Ding**

**INFO 5505 Applied Machine Learning for Data Science**

**UNT**

The main objective of assignment is to execute the classification technique (Naïve Bayes Algorithm) and utilizing the technique on the reviews given by customer for the classification of the sentiments as well.

The NB Classifier expects that there is not any relation between the attribute and any other present attribute.

**Dataset:** I have collected the dataset from

URL: <http://www.cse.chalmers.se/~richajo/dit862/data/all_sentiment_shuffled.txt>

The Data is gathered from 6 different topics (Reviews) which are given by various customers.

The Dataset is arranged in such a way that each line has 1 review and tokens are used to split the text and converted to lower-case.

Here the topic labels are denoted by 0, The polarity of sentiments (positive or negative) are denoted by 1, identifiers by 2 and 3 for words in document.

**Step 1: Importing the required libraries**

Text

Description automatically generated

**Step 2: Importing Dataset using the URL**

Graphical user interface, application

Description automatically generated

**Step 3: Defining different functions for preprocessing of data.**

Imported the Regular expression through libraries and defined the function for removing the single, 2 words from the customer reviews

Text

Description automatically generated

Install the stop words package and eliminated stop words, spaces from the customer reviews.

Graphical user interface, text, application

Description automatically generated

**Step 4:** Defined the review extract function

The lines were split using split line’s function and then append from 3 index in this way the reviews are extracted.

Graphical user interface, text, application

Description automatically generated

**Step 5:** Defining labels extract function

The data is split using the split line’s function and 1st index is used to extract the labels. The labels would be either 0 or 1.

Text

Description automatically generated

**Step 6:** Defining topics extract function

The data is split using split line’s function and 0Th index is used to extract the topics.

Text

Description automatically generated

**Step 7:** DefiningDoc\_ID function

The data is split using split line’s function and 2nd index is used to extract the Doc\_ID

Text

Description automatically generated

**Step 8:** Creating Data dictionary and assigning respective keys and values and then creating the data frames from the data dictionary.

Text

Description automatically generated

Table

Description automatically generated with low confidence

**Step 9:** Carrying out train test split, Defining X as reviews and Y as labels

Split the train and test data in ratio of 80:20

Icon

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

Checking the train and test data shape for both X, y

Text, letter

Description automatically generated

**Step 10: Using the Counter vectorization to transform data which is text format into vector**

By utilizing the term freq and the TF-IDF we transform the train test data into vector format.

Graphical user interface, text, application

Description automatically generated

**Imported Multinomial NB from the Sklearn.naive\_bayes and fitting the model train vector and y train.**

Graphical user interface, text, application

Description automatically generated

**Step 11:** Imported cross validate from sklearn package and evaluated the accuracy of the model by utilizing the train data. I checked with the 2 different cv values 4 and 5.

Where I found that accuracy score when the cv value is 4 is 79.61%.

While the accuracy score when the cv value is 5 is 80.22%.

Text

Description automatically generated

**Step 12:** I applied the NB model on test data and evaluated the cross\_validate score and accuracy score for the test data.

Table

Description automatically generated with low confidence

Table

Description automatically generated

Table

Description automatically generated with medium confidence

Text

Description automatically generated

The Model Accuracy on the test data was estimated to be 82.24% while the error rate was the rest of the percentage.

**Step 13: Confusion matrix**

Chart, treemap chart

Description automatically generated

From the above confusion matrix, it can be illustrated that the count of the reviews which were negative and predicted to be as positive was found to 194. While count of review was found to be 229 which were projected negative but are actually positive reviews. The models was correctly able to classify as the positive-1015 and negative-945 reviews.

**Step 14: Evaluated the AUC score and plotting ROC Graph**

Plotted the graph for TP\_Rate VS FP\_Rate and evaluated the AUC was found to 88.3%.The performance of the model can be evaluated from AUC score greater the area under curve the model founds to be more accurate.



Graphical user interface, text, application

Description automatically generated

Chart, line chart

Description automatically generated

**Conclusion: The Area under curve was evaluated to be 0.883 and Accuracy score was approximately 80%. Hence, we can infer that prediction where found to be accurate when it comes to classification of the 2 different label categories(POS, NEG).**