Sasha Ena Ofori

Mr. Dennis Owusu

Lab 4 Report

Natural Language Processing

**LAB REPORT: EVALUATION OF PERFORMANCE BETWEEN NAÏVE BAYES AND LOGISTIC REGRESSION SENTIMENT CLASSIFIERS**

LIBRARIES USED AND REASONS

**SKLEARN** This was the main library I used in the classification of both naïve Bayes and logistic regression classifiers. This was because this library provided me with tools for extracting features for classifiers, training and testing. Also, this classifier provided clear and concise documentation from which I could easily understand and therefore can be used by programmers of different skill level.

**The Natural Language Tool kit (NLTK)** in conjunction with the “re module” (Regex Library) was used in the preprocessing of data. Preprocessing meaning, I was able to clean my data of punctuations and other unnecessary characters that would not change results in any way. For Stemming and Lemmatization I used the NLTK Lancaster stemmer to help improve classifier results

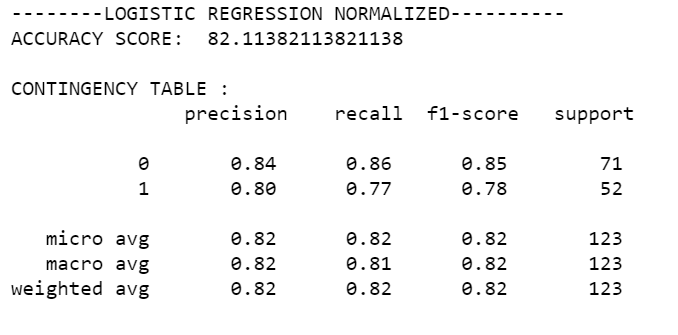
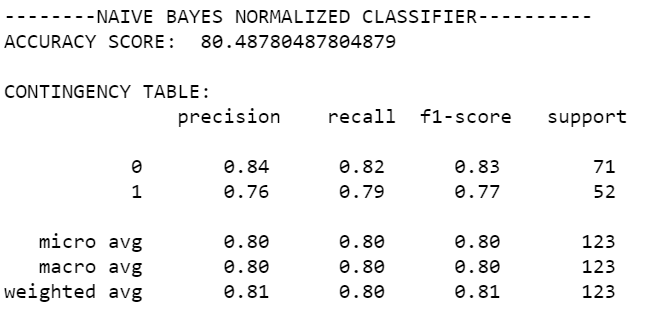
***Pickle:*** *This library helped me to serialize and de-serialize my object structure. This means, the object can be saved on the disk after being converted into a byte stream. Thereby to recreate the original object, I just needed to unpickle and this saved a lot of time.*

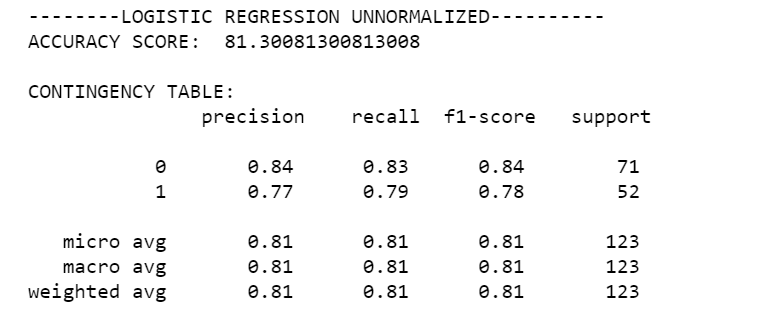
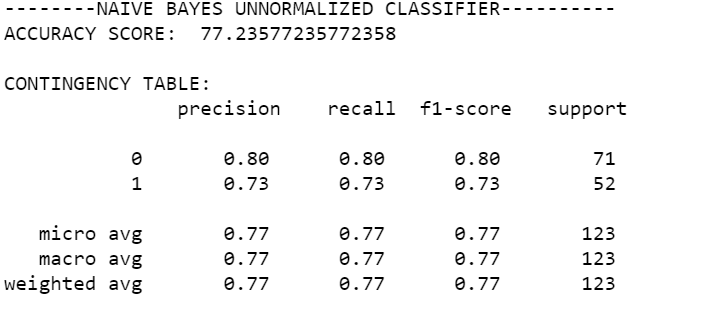
**SYS**: Sys module provided me with the functionality to access variables through the command line which could be used by the interpreter.

EVALUATION OF CLASSIFIERS AND RESULTS

To evaluate both normalized and unnormalized versions of Naïve Bayes and Logistic Regression Classifiers, I used the evaluation checks i.e. Accuracy, Precision, Recall, F- Measure. In order to give a full report on the efficiency of the classifiers, accuracy was also calculated for as it helps to bring out this feature.

*Insert images here*

**

** **

DISCUSSION OF RESULTS *(compare results and give opinions on why certain results were achieved)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sentiment Classifier | Accuracy Score | Precision Score | Recall Score | F1 Measure | Support Score |
| Logistic Regression Classifier |  |  |  |  |  |
| * Normalized Version | 82% | 0: 84%  1: 80% | 0:86%  1:77% | 0:85%  1:78% | 0:71  1:52 |
| * Unnormalized Version | 81% | 0:84%  1:77% | 0:83%  1:79% | 0:84%  1:78% | 0:71  1:52 |
|  |  |  |  |  |  |
| Naïve Bayes Classifier |  |  |  |  |  |
| * Normalized Version | 80% | 0:84%  1:76% | 0:82%  1:79% | 0:83%  1:77% | 0:71  1:52 |
| * Unnormalized Version | 77.2% | 0:80%  1:73% | 0:80%  1:73% | 0:80%  1:73% | 0:71  1:52 |

The results from both classifiers showed that normalized classifiers perform better than unnormalized classifiers no matter the type of classifier. This performance was measured using recall, precision and the F1 score predominantly with accuracy supporting.

Logistic Regression Classifier performed relatively better than Naive Bayes Classifiers