Dalhousie University

CSCI5408 Data Management and Warehousing Analytics

Project Report

Assignment 6

Submitted By :
Deeksha Behara :B00784704
Tushar Gupta :B00782699

Supervisor:
Suhaib Qaiser

July 31, 2018

Contents

1 Introduction

The motivation of the assignment to learn to use Scikit machine learning tool for data analysis and perform some visualizations to explore the data and use the insights for further analysis. The goal of the project is also to identify patterns through the simulation of the data. (https://github.com/bio33/Apache-Spark-Sentiment-Analysis/tree/master/ass6)

2 Experimental Setup - Task Description

The initial task implementation started off by setting up the environment for the assignment and performing predictive analysis on the data. Certain python libraries were used like Sci-kit machine learning tool, Matplotlib for visualization to perform predictive analysis and to identify the patterns after obtaining the results.

2.1 Tools, packages , technologies and libraries used

List of tools,technologies and libraries used for the study are as follows:

1. Python SciKit:

   Python Scikit provides efficient tools for data mining and data analysis.The library is open source and is reusable on various contexts.

2. Python Matplot:

   This library offered by python allows to plot graphs like histograms, bar charts, scatter plots and various other charts.

3. Principal Component Analysis:

   Principal component analysis performs linear transformation on the data which can be used in numerous applications such as stock market predictions.

4. Pycharm:

   PyCharm is the integrated development environment specifically suitable for python, was used to program for this project.

5. Pandas Dataframe:

   Pandas Dataframe was used to make data transformations.

3 DataSet

3.1 Dataset Description

The training data set for the assignment consists of 13 different languages. Certain languages present are closely related to each other. Languages were classified based on similarities. Similar languages were clubbed and groups of A to F were made. Group A consists of Bosnian, Croatian, Serbian, Group B consists of Indonesian, Malaysian, Group C consists of Czech, Slovakian , Group D consists of Brazilian Portuguese, European Portuguese, Group D contains Peninsular Spain, Argentine Spanish and finally Group D contains American English, British English.

3.2 Feature Extraction and Selection :

The feature extraction is the process of converting the raw text to features which gives the classifier a simpler and more focused view of text.

Suitable features were extracted using LDA and select KBEST chi square offered by Sklearn. The best features from both the selectors were combined using feature union.

For the experiment purpose ,the number of features that was extracted from the data were 10, 50 and 100. There was a limitation on memory to increase the features beyond a hundred. The experiments were conducted on the combination of the features extracted from the feature union. We had to restrict ourselves to go for the best 50 features because, going beyond 50 features was

causing memory issues.Additionally , the "countVectorizer" was also used to convert the documents to a matrix of tokens before performing feature extraction.

1. LDA

   Linear discrimination Analysis is commonly used for reducing the dimensionality for extracting the features.

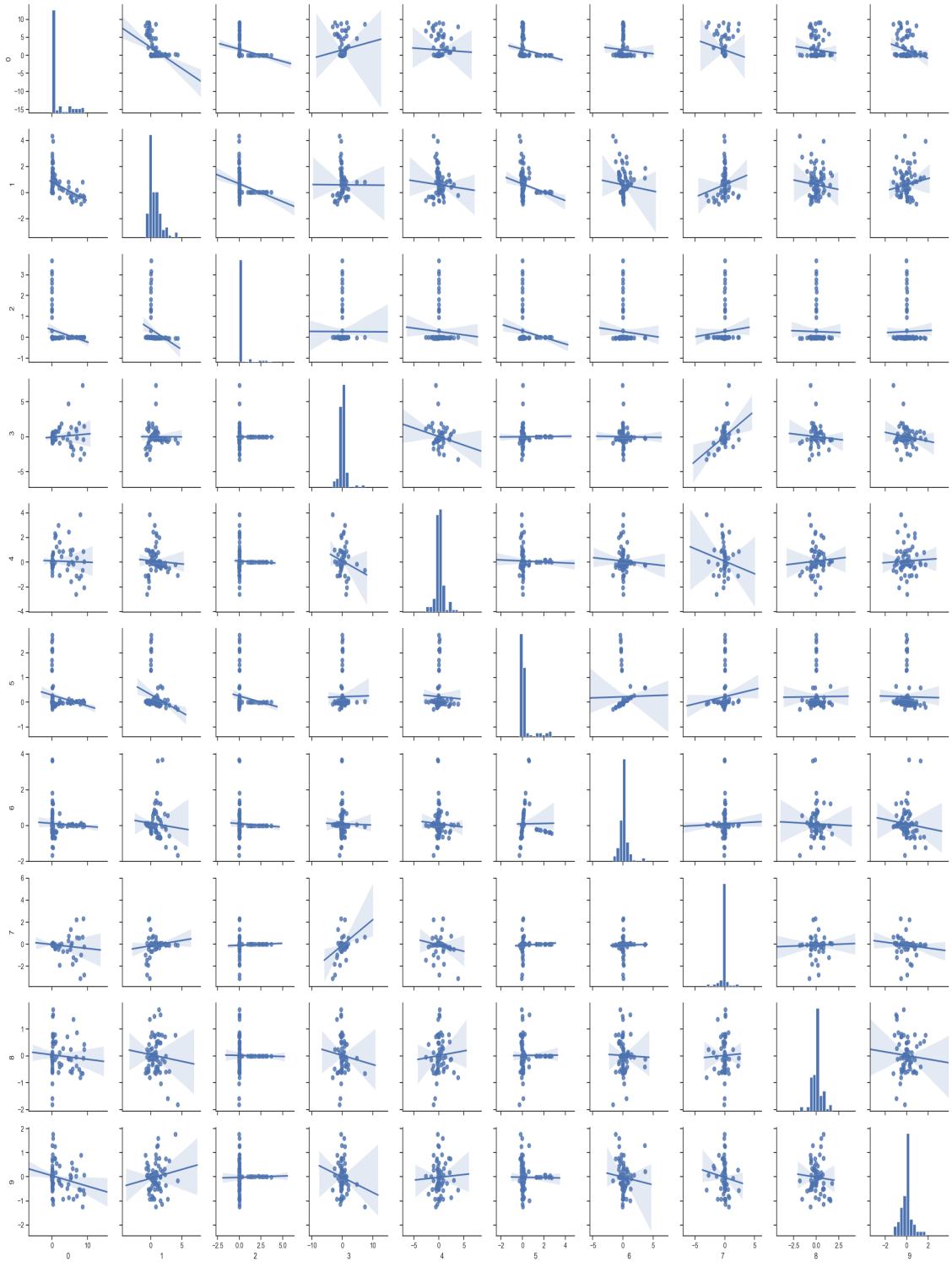

Figure 1: 10 by 10 pair wise plot was constructed for the selected top ten features.

2. Select KBEST

   Select Kbest selects the top K features that have maximum relevance to the target variable. It takes two parameters as input which are 'K' that stands for the features and the the score function to rate the relevance of each feature.

3. Feature Union

   Feature Union was used for combining the features from the feature sets obtained from the previous sets. Feature union is helpful to concatenate the results from different transformer objects.In this case, 25 features from each of 'selectKbest' and 'LDA' were taken into consideration.

4 Classification Algorithms:

1. Logistic Regression

   Logistic Regression is a linear model for classification. It is a statistical method for analyzing the dataset in which there are one or more independent variables that determine the outcome.

2. Linear SVM

   Support Vector Machine is a supervised machine learning algorithm which can be used for both classification and regression challenges. SVM is also available in scikit-learn library.

3. Decision Tree

   Decision Tree algorithm belongs to the supervised learning algorithms. Decision tree makes feature selections based on specific criterion that further computes the importance of each of the attribute and arrive at possible solutions to these possibilities respectively.

4. Naïve Bayes

   Multinomial Naive Bayes algorithm is suitable for a classification problem with discrete features. The Naive bayes algorithm usually takes integer feature inputs.

5 Approach Followed:

1. The training dataset was downloaded from the source and transformed to a feature set using linear discrimination analysis and select kbest.
2. Count Vectorizer was used to convert the collection of documents into a matrix of token counts. The transformation of the documents using count vectorizer results in a sparse representation of the counts using the sparse matrix.
3. The features extracted using the linear discrimination analysis and the select kbest was combined using feature union.
4. The training data was used for training the models for the classifier Naive Bayes, Linear SVM , Logistic Regression and Decision Tree.
5. A Pipeline was constructed to combine all the steps into one multi-stage model for each of the classifier type.
6. The accuracies were calculated for each of the models using the test data.
7. Comparison was drawn for the models used individually.
8. Finally , comparison and the confusion matrix was done for each of the models.

6 Output - Confusion Matrix and Accuracies

The classifiers were tested using 10, 50 and hundred extracted features components from the dataset to draw a comparison. The confusion matrix for the combined features from LDA and KBEST was taken into consideration for drawing results (confusion matrix) and constructing graphs.Below are the results obtained for each of the classifiers.

1. Linear SVM

   Linear SVM constantly failed during the experiment since the number of rows in the dataset is 252,000 which is way above the capacity of the classifier. The classifier has a limitation of processing 10,000 rows at a time according to the sklearn documentation(http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC).

   However , due to the performance issue we restricted the number of features for the classifier to 1000 and the accuracy obtained is 35

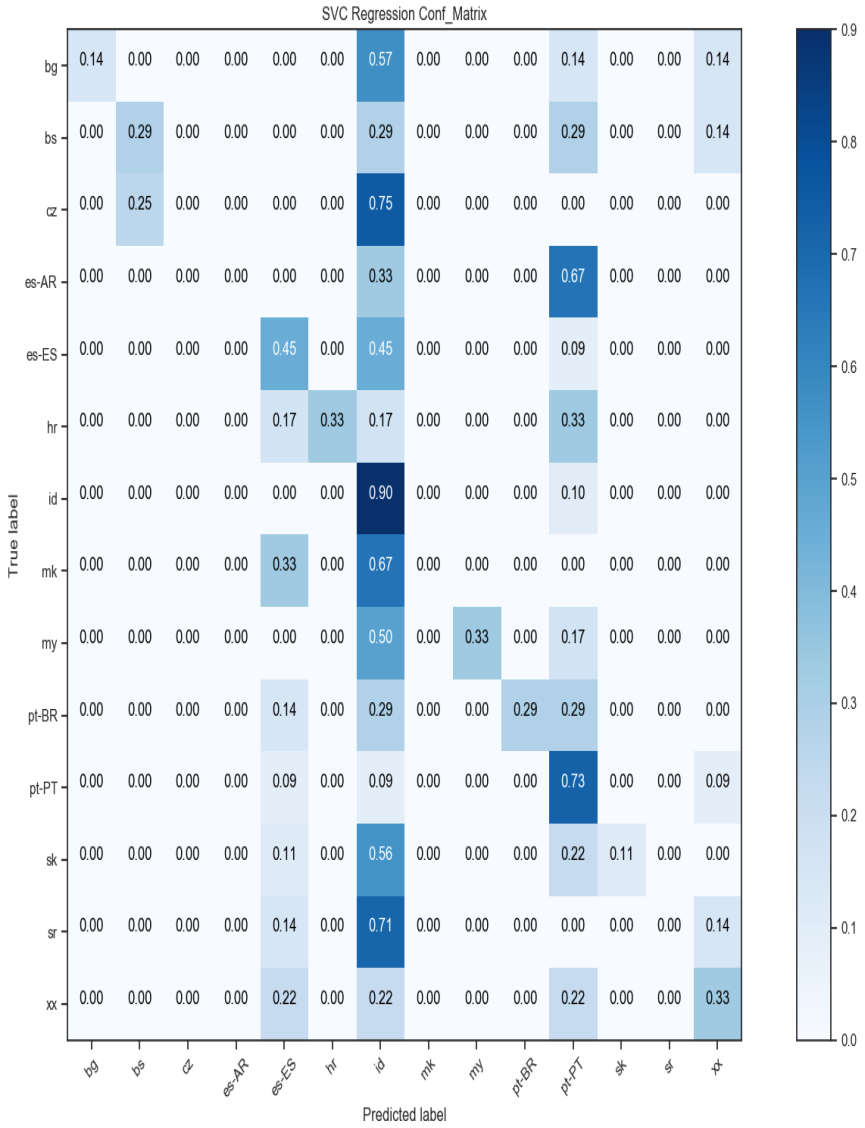

Figure 2: Confusion Matrix for Logistic Regression is shown in the diagram above.

2. Logistic Regression

The Confusion matrix for the Logistic regression is shown in the diagram below.

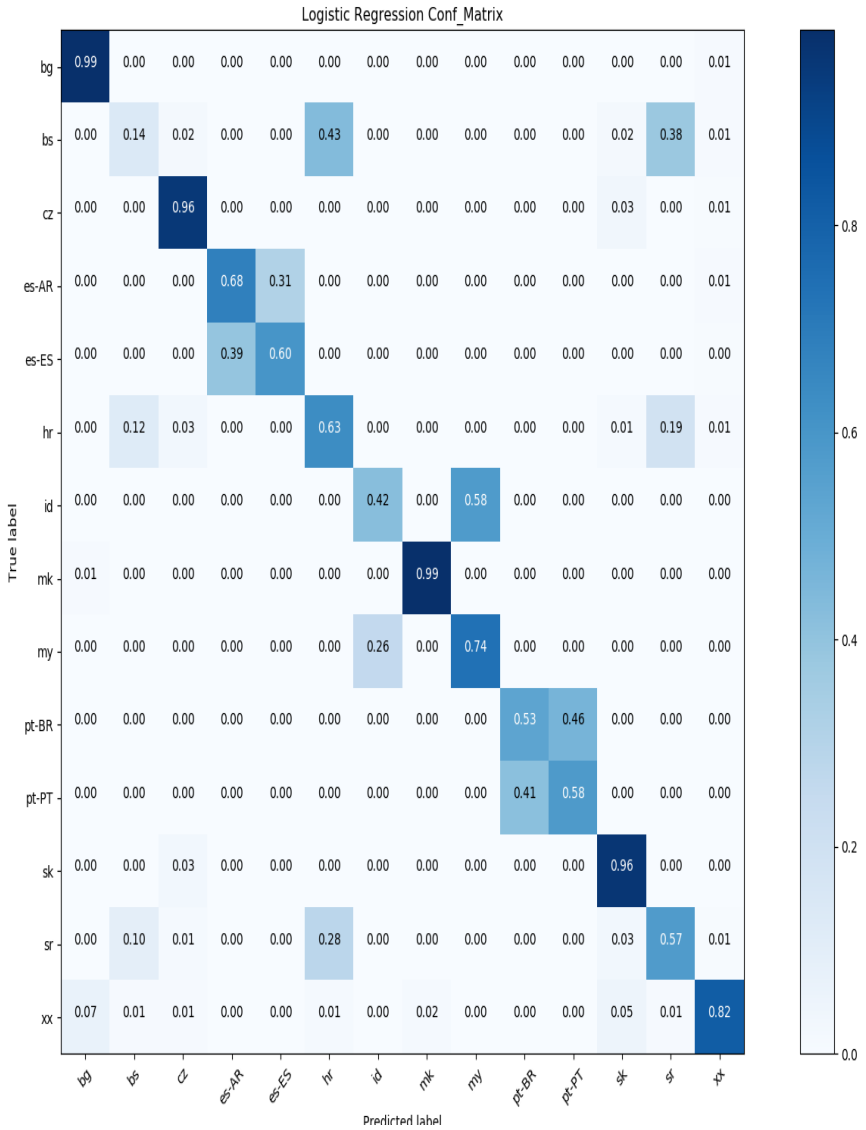

Figure 3: Confusion Matrix for Logistic Regression is shown in the diagram above.

3. Decision Tree

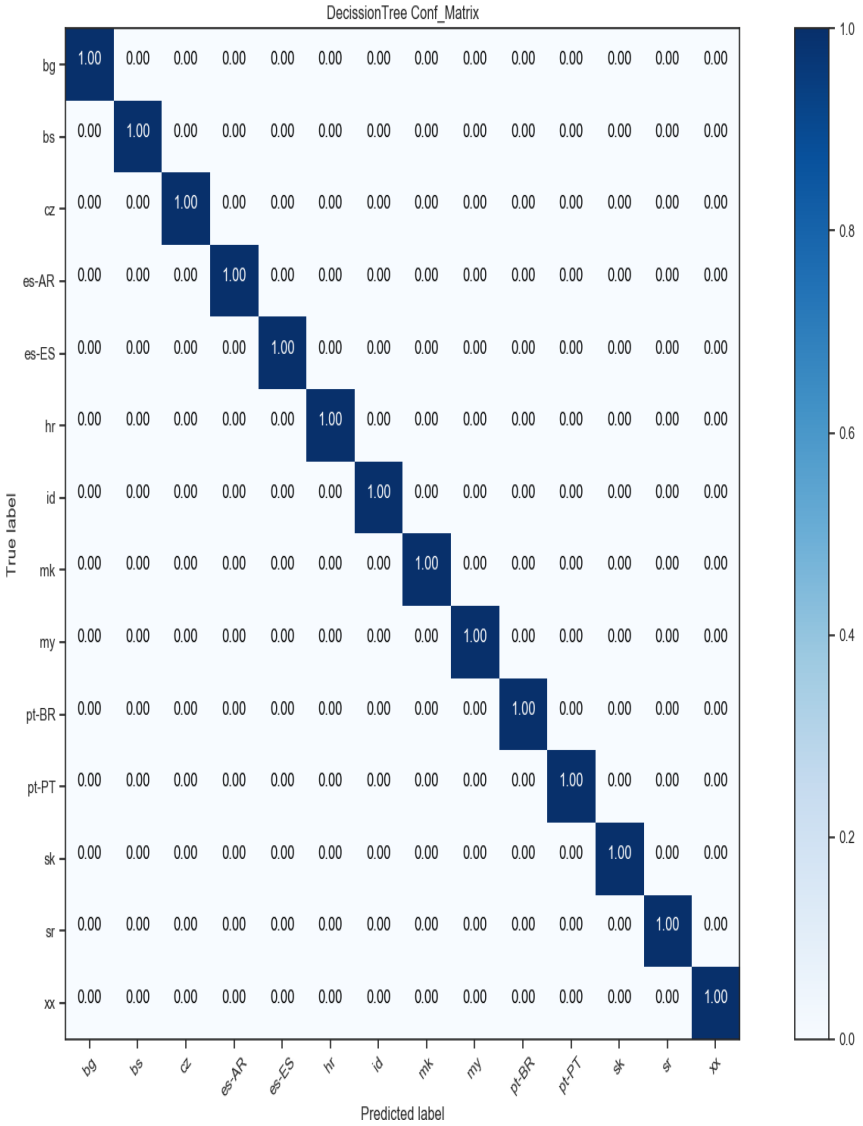

Figure 4: Confusion Matrix for Decision Tree is shown in the diagram above.

4. Naïve Bayes

   The Confusion matrix for the Naive baye's classifier was constructed as given in the diagram below.

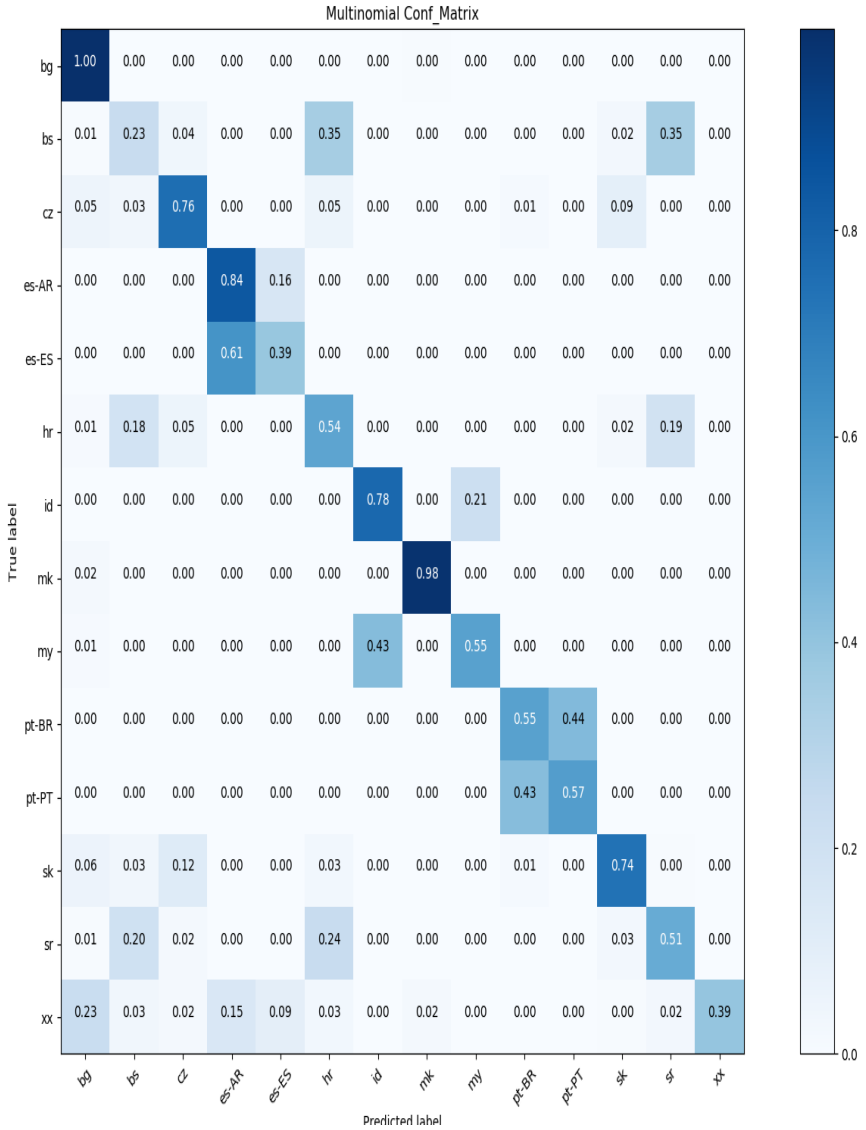

Figure 5: Confusion Matrix for Bournalli's Naive Bayes is shown in the diagram above .

7 Comparison

Classifier	Ten features	Fifty features	Hundred features
SVM	32%	45%	NA
Logistic Regression	52%	71%	77%
Decision Tree	98%	99%	98%
Bernoulli Naive Bayes	43%	55%	61%
Multinomial Naive Bayes	27%	57%	63%

* The classifier Decision Tree performs the best with an average of 98 percent accuracy.

* The classifier Logistic Regression gives an average accuracy of 70 percent.

* The classifiers MNB, SVM gives the next best accuracies of 55 and 37 percent respectively.

Classifier	Combined Features
SVM	37%
Logistic Regression	68%
Decision Tree	98%
Bernoulli Naive Bayes	57%
Multinomial Naive Bayes	negative values

* The comparison between the classifiers is shown in the diagram below. SVM is not taken into consideration because it was performing poorly for the given problem statement.

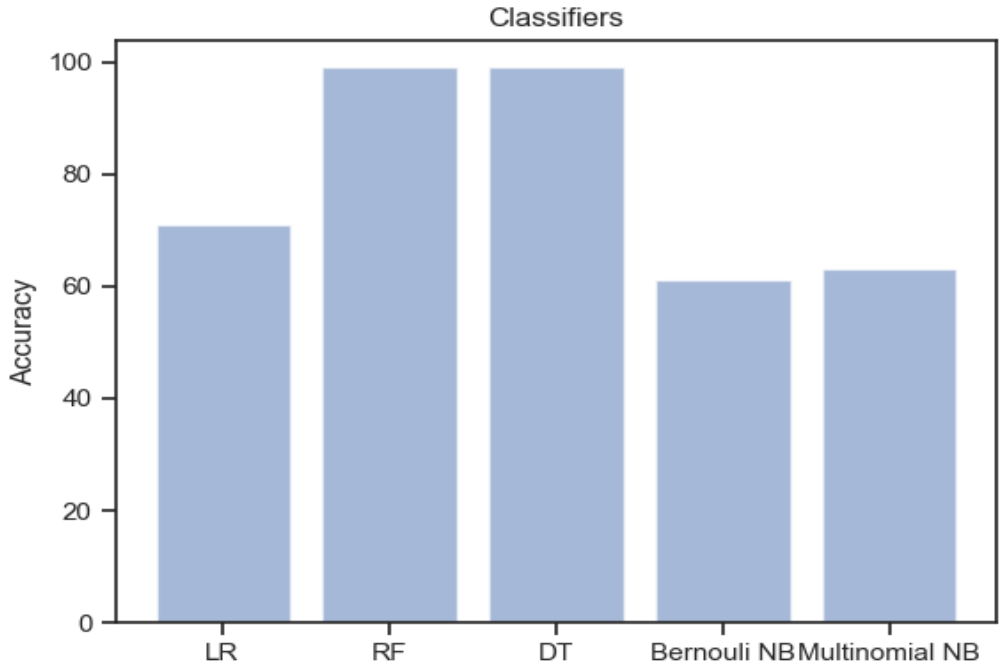

Figure 6: Bar chart representing the comparison between different classifiers which shows that the random forest and the decision tree classifiers performs the best.

8 Conclusion

In this project, we practiced the process of feature extraction, data preprocessing. We also constructed pipelines to combine the steps individually for each of the classifiers. The accuracies obtained from each of the classifiers were compared to conclude the best classifier for the problem defined. The confusion matrix was constructed to discover the actual performance of the classifier.

References

[1] https://github.com/Simdiva/DSL-Task
[2] http://scikit-learn.org/stable/
[3] https://matplotlib.org/
[4] http://corporavm.uni-koeln.de/vardial/sharedtask.html