ROOT2AI Technology Private Limited

Classification Assignment Report

Presented By-

Name-Janvi Gupta

College-GLA University ,Mathura

Course-Btech

Branch-Computer Science with specialization in AIML

Year-2nd

E-mail-janvi.gupta_cs.aiml19@gla.ac.in

About the Assignment

- The assignment is based on the classification problem.
- We have to classify the dataset on the bases of given target.
- We have to predict for the certain text which target domain is best suited.
- We have to use Text classification techniques to solve the problem.

My thoughts about the problem:-

It was a very interesting problem. I have a solved a lot of classification problem but this problem teaches me a lot of new things and a new ways to approach. I really enjoy solving that problem. Thank you for giving me this opportunity .

Approach to solve the problem

- First, I have imported the dataset as csv file and then dropped all the 3 null values from ‘Text’ column.
- Second, I have used NLP, Tfidf Vectorizer for processing and understanding the dataset.
- Third, I have split the dataset for training and testing purpose.
- Fourth, To make the vectorized classifier easier to work , I have use PipeLine class in Scilkit-Learn that behaves like a compound classifier.
- I have made models using Logistic Regression & Support Vector Classifier.
- Last, I have calculated the Testing ,Training, Accuracy Score.

Model Interpretation

- As we can see the testing and training scores from the model in the next slide. We can say that the model in fitted good enough.
- On comparison of the model score it is clear that the Logistic model in better than the other two as it has maximum accuracy score.
- But further we can’t say this because the accuracy score of any of the algorithm is not that good.
- Therefore we can say all the models are just good but not the best to predict the best target.

Train & Test accuracy score

- As I have made three models:
- By Logistic Regression:-
- Testing Score:- 0.6156217882836588
- Training Score:- 0.9206419131529263
- Accuracy Score:-0.6156217882836588

- By SVM:-
- Testing Score:- 0.6100425781823521
- Training Score. 0.6770925110132159
- Accuracy Score:-0.6100425781823521

- By Naïve Bayes’ Theorem:-
- Testing Score:- 0.5071208339450888
- Training Score:- 0.5494021397105098
- Accuracy Score:-0.5071208339450888

Limitation of the models

Logistic Regression:-Drawbacks of the model is few of the assumptions of logistic regression i.e. there is no high inter-correlation among the predictors, there is a linear relationship between the sigmoid of the outcome and the predictor variables.

SVM-As this model learn from the support vectors, unlike other machine learning models that learn from the correct and incorrect data. Therefore, they perform better most of the time. The drawbacks are it is not appropriate for non-linear problems, not the best choice for large number of features.

Naive Bayes Classifier:- It works on the basis of Bayes' Theorem The fundamental assumptions made are that all the features are independent of one another and contribute equally to the outcome; all are
of equal importance. But these assumptions are not always valid in real life and that is the drawback of Naive Bayes Classifier.

THANK YOU