Experiment 5
Text classification: Naive Bayes, Rocchio and K-Nearest Neighbor

I. EXPERIMENTATION DONE

Perform classification task on 20 Newspaper Dataset using Naive Bayes classifier, Rocchio classifier, K Nearest Neighbor classifier and analyze the results obtained.

II. APPROACH

A. Setting up the environment

We downloaded the 20 Newsgroup dataset and imported different python libraries for our task such as numpy, pandas, natural language toolkit (nltk), re(for regular expressions) and sklearn (scikit-learn learn) for classification models.

B. Data analysis

The dataset of the 20 Newsgroup contains two primary segments as information and target. The information section contains the genuine article of the paper and the objective segment shows the class to which this article has a place.

C. Data Cleaning

Prior to training our models on these articles we need to perform to cleaning task to retrieve better semantics from our information. For this reason as a matter of first importance we bring down the entire text followed by removing stop words and punctuations. Then, at that point we saw that our dataset contained some exceptional articulation like @, , sections and so on so we eliminated those images utilizing regex articulations.

D. Text Data Vectorization

Our ML models don't comprehend text based information along these lines,Firstly, we need to change our information in an format which our model can understand for example numerical information. For this, we initially change our information into its Bag of words representing,using the CountVectorizer technique for sklearn. This Bag of Words representation is then used to make TF-IDF portrayal of our informational index which is then at last took care of into our ML models for preparing. As the information size is huge and we have " words in our jargon this framework comes out to be an extremely scanty network.

E. Training the models

Utilizing sklearn we make three distinct models for each Naive Bayse, Rocchio and KNN classifier. TF-IDF lattice was our contribution for this load of models and the objective variable for which we needed to prepare our models is accessible under the objective section of the dataset.

F. Prediction and evaluation

After we train our model we play out every one of the information cleaning and preprocessing on our traning dataset also. Then, at that point utilizing this test dataset we foresee our marks and think about the expectations created by our model against the genuine names which are again accessible under the objective section of the test dataset. We find the f1 score of our forecast utilizing the inbuilt library of the sklearn.

III. RESULTS DISCUSSION

- The f1 score obtained from the Naive Bayes classifier is as follows:- **0.8191715347849177**
- The f1 score obtained from the Rocchio classifier is as follows:- **0.7573021773765268**
- The f1 score obtained from the KNN classifier is as follows:- **0.7428305894848647**

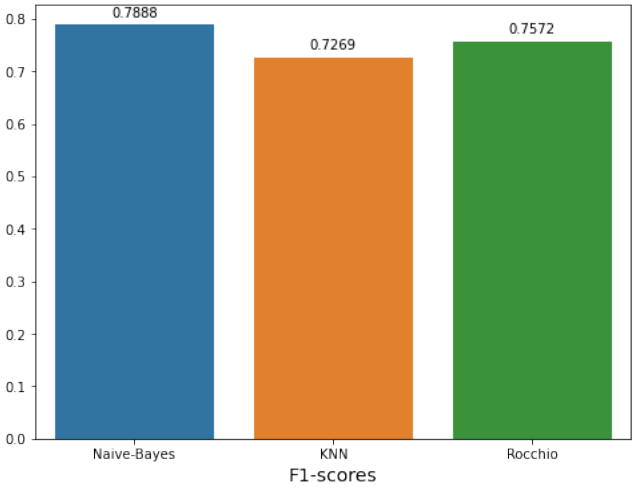

fig 5.1
f1 scores for different classifiers.

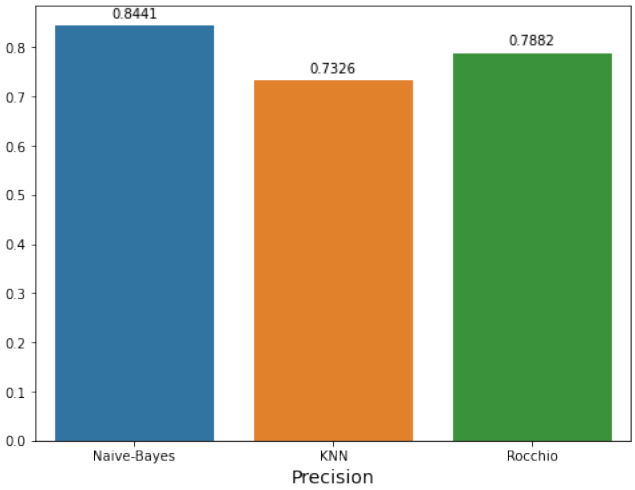

fig 5.2
precision scores for different classifiers.

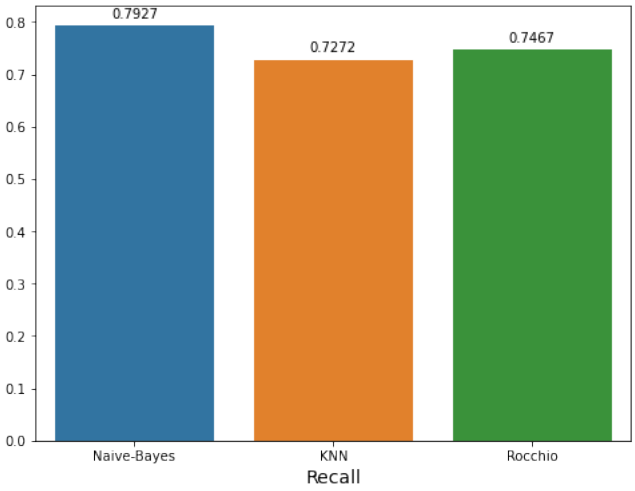

fig 5.3
recall scores for different classifiers.

IV. INTERPRETATION

According to the f1 score, our Naive Bayes classifier was the best fit on our dataset followed by Rocchio and KNN respectively. We can also see that for Naive Bayes, Rochio and KNN the precision is always better than recall both again Naive bayes has both higher precision and recall followed by ration and KNN which is reflected in the f1 score as well as f1 score is harmonic mean of precision and recall.

V. CONCLUSIONS

Naive Bayes classifier is the best classifier for the given dataset in terms of the F1 score. It outperforms other classifier in precision and recall as well.