In this repo we build a classifier for the "fetch\_20newsgroups" data set contained in the sklearn library.

Summary of the code file :

1)**Exploring the dataset** : consists of texts classified in 20 different categories based on their content.

2)**Constructing Tokenizer1 and Tokenizer2 functions** : these functions are constructed based on the build-in word\_tokenizer function along with other word operations/transformations, such as removing stopwords, converting upper letters to lower letters, word stemming etc.

3)**Defining text classifiers** : We consider the TfidfVectorizer transformer to get a vectorized version of the text data. We then feed them to the following classifiers :SGD classifier,Multinomial Naive Bayes classifier and Gradient Boosting classifier

4)**Training the models ( over reduced version of the dataset ie consider 4 labels only, due to limited computational power )**

5)**Evaluating the models** : accuracy, precision, classification report and confusion matrix

6)**Further analysis of the best classifier (Tokenizer1 + MNB ) using n-grams**

7)**Tuning and Evaluating the final model** : we consider the classifier which used both unigrams and bigrams and, tune its hyperparameters and evaluate its performance

8)**Behaviour of the final model over the 20-label dataset**