Project Topic:

**20 Newsgroup Document Classification**

The project will focus on classifying the 20 news data using machine learning algorithms such as Naïve bayes, Random forest, SVM etc. It is a supervised classification problem, there are news of 20 categories, each piece of news belongs to one category, the goal is to extract proper features and build an effective model to assign each piece of news to the correct.

I will explore the dataset in the beginning on the training part by extracting useful keywords and build vectors of features from the texts of news. Then I will use several classification methods to do classification based upon those vectors and then compare the efficiency of these classifiers on the test data and choose one. I will also use gridsearch to find the best parameters for classifiers.