Details of Classification Problem

The objective of the classifier is to classify emails into 6 distinct labels of:

- Academics
- Miscellaneous
- News
- Placements
- Promotions
- Social

The dataset is the same as the one used in the previous set, procured from our personal mail accounts. They are stored in the form of a CSV file.

The assignment is programmed in Python. All the classifiers were imported from the scikit-learn library.

The dataset is initially partitioned into training and testing sets and the training set is further partitioned into 10% initial labelled data and 90% active training data.

For Uncertainty Sampling, all four methods of least confidence, smallest margin, largest margin and entropy were tried out. The different disagreement strategies were implemented with help of modAL framework.

For Query-by-Committee both Vote Entropy and KL-Divergence were used.

Query-by-Committee was implemented with bagging on 5 different classifiers.

The different classifiers used were:

- Decision Tree
- Random Forest
- Linear SVM
- RBF SVM
- Sigmoid SVM

All the active learning models were tried out using 6 different classifiers and graphs were plotted including the passive learning performance for the corresponding amount of labelled data. The graphs have x-axis as percentage of actively learnt data and y-axis as the corresponding accuracy.

The different classifiers used on the Active Learning models:

- Decision Tree
- Random Forest
- Linear SVM
- RBF SVM
- Sigmoid SVM
- Polynomial SVM