**

COMP814 – Text Mining Lab

Classification and FPR Evaluation

# Objective

1. To be able to do classifications using python and use appropriate libraries to evaluate the results.

# Task

1. You will need the sample code snippets from lectures for the following tasks.
2. Download the classification data (txt\_sentoken.7z) as used in the lecture sample code.
3. Use a list of different classifiers (the list given below) from sklearn, train them on 80% data and test on the remaining 20% of the data as demonstrated in lectures.

**from** **sklearn.neural\_network** **import** [MLPClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html#sklearn.neural_network.MLPClassifier)

**from** **sklearn.neighbors** **import** [KNeighborsClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier)

**from** **sklearn.svm** **import** [SVC](https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC)

**from** **sklearn.gaussian\_process** **import** [GaussianProcessClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessClassifier.html#sklearn.gaussian_process.GaussianProcessClassifier)

**from** **sklearn.gaussian\_process.kernels** **import** [RBF](https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.kernels.RBF.html#sklearn.gaussian_process.kernels.RBF)

**from** **sklearn.tree** **import** [DecisionTreeClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier)

**from** **sklearn.ensemble** **import** [RandomForestClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier), [AdaBoostClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html#sklearn.ensemble.AdaBoostClassifier)

**from** **sklearn.naive\_bayes** **import** [GaussianNB](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html#sklearn.naive_bayes.GaussianNB)

**from** **sklearn.discriminant\_analysis** **import** [QuadraticDiscriminantAnalysis](https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis.html#sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis)

1. Tabulate the results and make conclusions as to which one is the best for the data.
2. Now make the data **imbalanced** (make one of the categories 50% less data points. ie 500 instead of 1000) and repeat the process. Make conclusions as to which one is the best performer.
3. Download the second set of classification data (Classification data computers.7z) and test your best model on it. Comment on the accuracy.
4. Upload our comparisons and code as evidence for attempt at this lab.