**PSG College of Technology**

**Department of Applied Mathematics and Computational Sciences**

**MSc SS / TCS / DS**

**Natural Language Processing Lab**

1. Given a dataset for drug sentiment with Medicine, disease, tweet about the adverse effect of medicine for a disease and the sentiment, do the following tasks:
   1. Preprocessing steps for sentiment feature:
      1. Tokenization in which convert the tweet into words
      2. Remove stop words
      3. Do stemming
      4. Formulate bag of words
   2. Formulate training set using features and class label (Drug, disease, bag of words, sentiment)
   3. Do visualization using different plots which depict
      1. Frequency distribution of words disease wise / sentiment wise / medicine iwse
      2. Adverse effects Vs Drug for disease wise
      3. Drugs Vs disease Vs sentiment
      4. Etc many more (Innovative themes and plots are welcome)
   4. Develop a UI
   5. Apply the following ML algorithms on the data and choose the best algorithm
      1. SVM
      2. Logistic regression
      3. KNN
      4. Naïve Bayes