# Machine Learning

Assignment #05

**Due Date:** 6-Jan 2020

Assignments are to be done individually. No late assignments will be accepted.

**Submissions that do not comply with the specifications given in this document will not be marked and a zero grade will be assigned.**

You are needed to submit a single .zip file (file name is your ID and Name) and printed file in class.

## Introduction

Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

## Task

- Upload the attached Urdu Classification Dataset into Pandas
- Clean the data.
- Build a tokenized list of words from the data.
- Remove stop words, **Lemmatizing** and **Stemming**, remove words with less frequency(Outliers)
- Feature Extraction
- Apply a **Logistic Regression** model using a feature extraction algorithm bag of words (CountVectorizer)
- Apply a **Logistic Regression** model using feature extraction algorithm **TfidfVectorizer**.
- Compare the output of each model above.
- Apply **SVM (support vector machine predictor)** to bag of words (**CountVectorizer**) and compare the output with above.

**Honor Policy**

This task is a learning opportunity that will be evaluated based on your ability to work through a problem in a logical manner and write a report on your own. You may however discuss verbally or via email the assignment with your classmates or the course instructor, but you are to write the actual report for this task without copying or plagiarizing the work of others. You may use the Internet to do your research, but the written work should be your own. **Plagiarized code will get a zero.**