**ML Assignment 1**

**Naive Bayes**

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where the assumption is that every pair of features is independent of each other.

**Steps taken to implement algorithm**

1. Remove special characters and stopwords from the sentences.
2. Generate vocabulary by making a sorted list of all words in all sentences.
3. Generate an equivalent vector for each sentence.
4. Generate bag of words representation by stacking all vectors.
5. Calculate conditional prior probability for each word (with Laplace Smoothing).

*P(word = 1/sentiment = k) = (nx+1)/(n+2)*

*where nx = number of times word is present when sentiment is k*

*n = number of times sentiment is k*

*k = 0 or 1*

1. Calculate log posterior probabilities for each sentence.

*P(sentiment/sentence) ∝ P(sentiment) \* P(sentence/sentiment)*

*= P(sentiment) \* ∏ P(word/sentiment)*

*ln(P(sentiment)) ∝ ln(P(sentiment)) + ∑ ln(P(word/sentiment))*

1. Sentence is classified as positive if log posterior probability of it being positive is greater, otherwise it is classified as negative.
2. Perform 5-fold cross validation.

**Steps taken for basic text pre-processing**

1. Removal of special characters and numbers

Without removing symbols accuracy = 0.79 ± 0.02

1. Removal of stop words - Removal of common words that can be ignored

Without removing stopwords accuracy = 0.79 ± 0.03

After text pre-processing accuracy = 0.80 ± 0.01

Removal of special characters increased accuracy by 1%

Removal of stop words increased accuracy by 1%

**5-fold cross validation results:**

Test fold 1 : Accuracy = 0.82 F-score = 0.83

Test fold 2 : Accuracy = 0.80 F-score = 0.82

Test fold 3 : Accuracy = 0.81 F-score = 0.83

Test fold 4 : Accuracy = 0.77 F-score = 0.79

Test fold 5 : Accuracy = 0.79 F-score = 0.79

Accuracy = 0.80 ± 0.02

F-score = 0.81 ± 0.02