**Assignment 02: Sentiment Analysis using NLP**

**Description:**

**Problem:**

Analyze the Sentiment dataset using NLP to:

1. View the observations,

2. Verify the length of the messages and add it as a new column,

3. Apply a transformer and fit the data in the bag of words,

4. Print the shape for the transformer, and

5. Check the model for predicted and expected values.

**Assessment:**

(Note: For code refer the attached notebook in pdf format)

**1. View the Observation**

* Import the required libraries.
* Using read\_csv (), get the sentiment dataset.
* Using df\_sentiment.head (10), View first ten records in the dataset.
* From the output of head (), in label column ‘1’ indicates positive sentiment and ‘0’ indicate negative sentiment.
* View more information about the sentiment data using describe and info method.
* View data using group by and describe method.

**2. Verify the length of the messages and add it as a new column**

* Verify length of the messages and also add it also as a new column (feature).
* Verify the dataset using head () method.

**3. Apply a transformer and fit the data in the bag of words**

* Import CountVectorizer from sklearn.feature\_extraction.text library.
* Define a function to get rid of stopwords present in the messages.
* Check and remove punctuations and stopwords in the dataset.
* Create a data variable bag\_of\_words by applying the function and fit the data (comment) into it.
* Apply transform method for the bag of words.
* Apply tfidf transformer and fit the bag of words into it (transformed version).

**4. Print the shape for the transformer**

* The shape of the tfidf is (748, 3259).

**5. Check the model for predicted and expected values**

* Choose Naïve-Bayes model to detect the spam and fit the tfidf data into it.
* Check model for the predicted and expected value say for comment#1 and comment #4.
* From the output, ‘1’ indicates positive sentiment and ‘0’ indicate negative sentiment.