**DEPARTMENT OF INFORMATION TECHNOLOGY**

**COURSE CODE:** DJ19TEL7014  **DATE:** 10/1023

**COURSE NAME:** Machine Learning  **CLASS:** Final Year B.Tech

**EXPERIMENT NO. 4**

**CO Measured:**

**CO1** Solve real-world problems using suitable machine learning techniques.

**TITLE:** Study of Support Vector Machine

**AIM / OBJECTIVE:**

To build a spam classifier using Support Vector Machine (SVM).

**DESCRIPTION OF EXPERIMENT:**

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

## **Steps to follow for classification**

## **Text Analytics**

Find the frequencies of words in the spam and non-spam messages. The words of the messages will be model features. We can see that the majority of frequent words in both classes are stop words such as 'to', 'a', 'or' and so on. With stop words we refer to the most common words in a language, there is no single, universal list of stop words.

## **Feature engineering**

Text preprocessing, tokenizing and filtering of stopwords are included in a high level component that is able to build a dictionary of features and transform documents to feature vectors.

Remove the stop words in order to improve the analytics

## **Predictive Analysis**

Here main goal is to predict if a new SMS is spam or non-spam. Because, that is much worse misclassify non-spam than misclassify as spam. (User don't want to have false positives)

The reason is because one normally don't check the spam messages.

The two possible situations are:

1. New spam SMS in inbox. (False negative).

OUTCOME: User delete it.

1. New non-spam SMS in spam folder (False positive).

OUTCOME: User probably don't read it.

Normally the first option is preferred!!!

**Train test Split**

First transform the variable spam/non-spam into binary variable, then split data set in training set and test set.

**PROCEDURE:**

Develop a spam classifier using Support Vector Machine for a dataset of your choice. Find the confusion matrix, precision, recall and discuss on results.

**OBSERVATIONS / DISCUSSION OF RESULT:**

1. What do you mean by classification?
2. List and compare various classification techniques.

**CONCLUSION:**

**​Website References:​**

[1] <https://www.kaggle.com/code/pablovargas/naive-bayes-svm-spam-filtering>

[2] <https://techvidvan.com/tutorials/spam-detection-using-svm/>

[3] <https://www.geeksforgeeks.org/support-vector-machine-algorithm/>