SYNAPSE ML TASK 3.3

# IDENTIFYING SPAM FROM UNLABELLED DATA

## Introduction

I delve into the realm of spam detection using unlabelled data. I meticulously explore the nuances of unsupervised learning techniques, particularly focusing on the power of clustering algorithms. Through this approach, we unravel hidden patterns within messages, shedding light on effective strategies for identifying spam content.

## Problem Statement

The objective is to categorize unlabelled paragraphs, messages, and emails as spam or not spam, employing techniques that suit a beginner's understanding of machine learning.

## Approach

- **Splitting dataset into test and training sets**
  - Split data into training and testing sets.
- **Data Processing**
  - Convert all text to lowercase to ensure uniformity.
  - Tokenize text into words for analysis.
  - Making a vocabulary of unique words.
- **Feature Extraction**
  - Convert text data into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency).
  - Use unsupervised techniques to identify patterns in the unlabelled data. Techniques like clustering (e.g., K-Means)) can help you group similar messages together.

- ❖ **Manual Labeling (Seed Data)**
    - ➢ Manually label a small subset of the clustered data as "spam" or "non-spam." This forms my initial training set.

- ❖ **Making the Model**
    - ➢ Utilize Multinomial Naive Bayes Algorithm due to its simplicity and effectiveness for text classification.

- ❖ **Checking accuracy of Model**
    - ➢ Evaluate model performance using metrics like accuracy, precision, recall, and F1-score.

## K-Means Clustering

K-means clustering is an unsupervised machine learning algorithm that groups unlabeled data into k clusters based on similarity. It iteratively assigns data points to their closest cluster centroid and updates the centroid until convergence. The resulting clusters are used to labelise the data.

## TF-IDF

TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a numerical measure used to score the importance of a word in a document based on how often it appears in that document and a given collection of documents1.

## Conclusion

Building a Convolutional Neural Network (CNN) for multi-label classification involves understanding deep learning principles, the significance of convolutional layers, and data preprocessing techniques. By iteratively optimizing cluster centroids, K-Means Clustering forms clusters that help uncover underlying patterns and structures in data.