

# IDENTIFYING SPAM TEXTS

---

## Introduction

In the digital age, spam texts have become a prevalent issue, cluttering communication channels and posing security risks. This report outlines an approach to identify spam texts using basic preprocessing steps and a simple classification method.

## 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

- Transform text data into a matrix of word frequencies.
-

---


### ❖ **Checking for Spam**

- 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.

## **Code Link**

 `Synapse_Task_3.ipynb`

## **Conclusion**

In conclusion, the approach outlined presents a systematic methodology for tackling the challenge of identifying spam texts within unlabelled messages. By employing text preprocessing techniques and utilizing a Multinomial Naive Bayes Classifier, I demonstrate a foundational framework for distinguishing spam from legitimate content. While effective, this approach may face limitations with evolving spam tactics and context-rich content. Through continuous refinement and vigilant monitoring, I can enhance the accuracy and relevance of our classification models in real-world scenarios.