

**COS30049**

**Computing technology Innovation Project**

**Assignment 2 – AI-based Project**

***Spam Emails/Messages Detection Using Machine Learning***

Swinburne University of Technology – Semester 2, 2025

Tutorial: Monday 12:30 AM – Jianwen Liu

Session 26 – Group 7 Members:

* Hong Dat Cao – 104497198
* Vu Thuan Huynh – 105218471
* Duy Le – 104180126

Table of Contents

[I. Introduction 3](#_Toc210122113)

[1. Project background 3](#_Toc210122114)

[2. Intended Users 3](#_Toc210122115)

[II. Problem Framing 3](#_Toc210122116)

# Introduction

## **Project background**

Thanks to the rise of digital communication, various unprecedented opportunities for information sharing has been created. However, it has also introduced significant risks in the form of spam. Spam mesages, whether delivered through emails for SMS, can compromise users’ privacy, waste resources or often even serve as entry points for phishing campaigns. These unwanted situations not only reduce productivity but can also trick individuals to have their sensitive information exploited by others or download harmful content. Traditional rule-based systems and signature detection methods struggle to keep up with the ever-changing tactics of spammers, who continuously disguise malicious intent to bypass static filters.

This project, ShieldMail, focuses exclusively on spam detection through the use of machine learning techniques. By leveraging datasets of spam and legitimate text messages and emails, we aim to train models that can distinguish between the two with high accuracy. The solution will be delivered through an interactive web application, allowing users to input or paste text for classification. Alongside classification, the system will also provide supporting visualizations, for instance, spam-to-legit ratios and word frequency charts, helping users better understand the patterns that differentiate spam from legitimate communication.

## **Intended Users**

**General Internet Users:** individuals who rely heavily on email or SMS for daily communication, including students, professionals and families.

**Small/Medium Businesses:** organizations with limited IT resources that depend on efficient and secure communication channels to conduct operations.

**Email Service Providers:** companies which manage and deliver email platforms to large user bases, requiring reliable spam filtering to maintain trust and usability.

**Cybersecurity Researchers:** academics and practitioners studying spam detection trends and machine learning applications in cybersecurity.

**Educational Institutions:** universities, schools and training providers that utilize email systems extensively for communication between staff and students.

# Problem Framing

The core challenge lies in accurately distinguishing between harmless communication and malicious spam. Most existing filtering solution nowadays fail to detect newly emerging spam tactics, as they rely heavily on static blacklists and manually curated rulesets. These systems sometimes also generate high false positive rates that disrupt users’ experience by blocking legitimate messages. As communication rapidly grows in scale and diversity, traditional methods mentioned above is unable to provide sufficient protection.

Machine learning, on the other hand, provides a more dynamic approach by constantly learning patterns from large datasets of spam and non-spam messages. This enables the system to generalize and identify various forms of spam content that do not match the pre-defined rules. In this project, the classification task is framed as a binary classification problem, where inputs are categorized as either Spam or Legit.

In order to achieve that, the project will explore algorithms such as Support Vector Machines (SVM), which are known to be well-suited for high-dimensional text classification. Then we compare them against other methods like Naïve Bayes or Logistic Regression, ensuring robustness.

The key challenges include handling imbalanced datasets, ensuring fast and interpretable predictions, and mitigating adversarial manipulation where spammers deliberately modify content to evade detection. Successfulness will be measured by achieving at least 80% accuracy, with balanced precision and recall, ensuring harmful messages are efficiently captured without an overwhelming number of users with false alerts.

# Dataset Collection

## **Spam Email Dataset**

This dataset consists of email text messages labeled for spam detection, where each entry includes the email body along with any subject lines or headers. A binary label, spam, indicates whether an email is considered spam as “1”, or not spam as “0”. The dataset is designed for training and evaluating spam classification modes.

* [Spam Email Dataset on Kaggle](https://www.kaggle.com/datasets/jackksoncsie/spam-email-dataset) – Created by **\_W1998**