The George Washington University **Columbian College of Arts and Science**

Final Project Proposal

EMAIL SPAM FILTER

DATS 6202 - Machine Learning I : Algorithm Analysis

Professor: Huan Zhang

Team Members
Aswin Balaji Thippa Ramesh - Data Specialist, Model Engineer & Presenter
Gowri Sriram Lakshmanan - Data Specialist, Model Engineer & Presenter

---

Problem Statement

The chosen problem is binary text classification with a focus on email spam detection. The objective is to classify emails into two categories: **"spam"** and **"ham"** based on their textual content.

This problem was selected due to its practical importance in filtering unwanted communication and enhancing cybersecurity. It also presents a compelling challenge for applying natural language processing (NLP) techniques combined with supervised learning methods, making it ideal for exploring the effectiveness of machine learning in real-world text classification tasks.

Dataset Description

This project uses the **Email Spam Collection Dataset** sourced from **Kaggle**, containing over **5,500 labeled messages** classified as either **"spam"** or **"ham"**.

It includes two main columns: the **label** (v1) and the **message content** (v2). The dataset provides a realistic sample of text communication and is well-suited for applying NLP and supervised learning techniques in binary classification tasks.

Approach

This project addresses a binary classification task using supervised learning, aiming to predict whether an email is spam or ham based on its text content. The process begins with standard text **preprocessing** steps, including **lowercasing**, removing **stop words**, **punctuation**, and **special characters**. The cleaned text is then tokenized & transformed into numerical representations using word embeddings.

For modeling, the initial focus will be on classical **machine learning algorithms** such as **Logistic Regression, Naive Bayes, and Random Forest,** which are known for their interpretability and strong baseline performance on text data. Depending on available computational resources and time constraints, the pipeline will be extended to include a **deep learning-based Multilayer Perceptron (MLP)** to explore potential performance improvements.

Expected Challenges

A key challenge in this project is the **class imbalance** in the dataset, where spam messages are significantly fewer than non-spam (ham) messages. This imbalance can lead to biased models that perform well on the majority class but poorly on the minority class.

To address this, techniques such as **oversampling** the spam class or assigning **higher class weights** during model training will be applied to ensure balanced learning and improve the model's ability to detect spam accurately.

Evaluation Metrics

Metrics For evaluating the model's performance, a combination of standard classification metrics will be used. The primary evaluation metrics include **Accuracy**, which measures overall correctness; **F1-Score** which balances precision and recall, and the **Confusion Matrix**, which provides a detailed breakdown of predictions across sentiment classes.

If applicable, **AUC-ROC** will also be used to assess the model’s discriminative ability, especially in binary comparisons within the multiclass setting. To contextualize performance, results will be compared against simple benchmarks, which serves as a standard baseline for text classification tasks.

Expected Outcome

The goal is to build a high-performing spam detection system that can accurately classify emails and protect users from unwanted or malicious content. This solution addresses a real-world need across industries—from **email providers** to **cybersecurity firms** where spam and phishing attacks threaten communication integrity and user safety.

By combining traditional machine learning with deep learning models, the project aims to deliver a scalable, efficient, and adaptable classification tool. With proper handling of class imbalance and rich text preprocessing, the model will be well-suited for deployment in **email platforms, enterprise tools, or SaaS products** looking to enhance their filtering capabilities.