Chương trình phân loại thư rác

# 1. Mô tả vấn đề nghiên cứu

Trong thời đại số hóa hiện nay, số lượng email được gửi đi mỗi ngày lên tới hàng tỷ, trong đó có một tỷ lệ lớn là thư rác (spam). Những email này không chỉ gây phiền toái cho người dùng mà còn có thể chứa mã độc, lừa đảo, hoặc quảng cáo không mong muốn. Do đó, việc xây dựng một hệ thống phân loại thư rác tự động, chính xác và hiệu quả là vô cùng cần thiết trong lĩnh vực bảo mật và xử lý ngôn ngữ tự nhiên.

Vấn đề đặt ra: Làm sao để xây dựng một mô hình học máy có thể phân biệt chính xác giữa email hợp lệ (ham) và email rác (spam)?

# 2. Phương pháp nghiên cứu

## 2.1. Mục tiêu

Sử dụng thuật toán Naive Bayes, một mô hình học máy đơn giản nhưng hiệu quả cao, đặc biệt phù hợp cho bài toán phân loại văn bản như thư rác.

## 2.2. Quy trình thực hiện nghiên cứu

1. Thu thập và tiền xử lý dữ liệu

* Sử dụng bộ dữ liệu công khai:…
* Làm sạch văn bản: Tách từ, chuyển chữ thường, loại bỏ từ vô nghĩa, chuẩn hóa từ

1. Biểu diễn văn bản thành vector

* Áp dụng mô hình **…** để chuyển đổi văn bản thành các đặc trưng định lượng.

1. Xây dựng mô hình Naive Bayes
2. Huấn luyện và đánh giá mô hình

* Chia dữ liệu huấn luyện
* Đánh giá qua các chỉ số: Accuracy, Precision, Recall, F1-score, Confusion Matrix.

# 3. Cách tiếp cận giải quyết vấn đề

## Mục tiêu

* Tự động phát hiện thư rác với độ chính xác cao.
* Đảm bảo thời gian xử lý nhanh, chi phí tính toán thấp (ưu điểm của Naive Bayes).

## Tiếp cận kỹ thuật

1. Preprocessing (Tiền xử lý):

* Tách từ, chuyển chữ thường, loại bỏ từ vô nghĩa, chuẩn hóa từ.

1. Vector hóa văn bản:
2. Huấn luyện mô hình Naive Bayes:
3. Đánh giá mô hình:

* In ra confusion matrix và các chỉ số hiệu suất để đo độ chính xác trong phân loại.