PHÂN LOẠI THƯ RÁC SỬ DỤNG RANDOM FOREST

Lê Anh Cường - 230202002

Trường Đại học Công nghệ thông tin - Đại học Quốc Gia Thành Phố Hồ Chí Minh

Mục tiêu

- Sử dụng thuật toán Random Forest dựa trên đặc điểm phù hợp với việc phân loại thư rác, bao gồm: có thể sử dụng được cho các bài toán phân loại, kết hợp nhiều cây quyết định (Decision trees) để đưa ra dự đoán, giải quyết được tập dữ liệu lớn, phức tạp; có khả năng giảm thiểu overfitting,...
- Xây dựng mô hình phân loại thư rác bằng cách sử dụng thuật toán Random Forest.
- Huấn luyện mô hình để có thể cho ra kết quả tốt nhất từ tập Dataset Spam Email của Kaggle.

Lý do

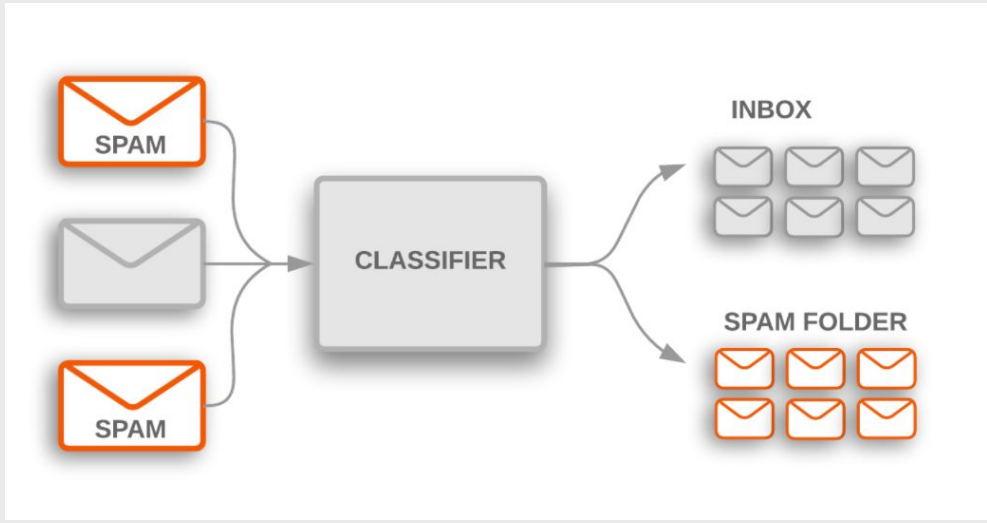
Email Spam không chỉ gây phiền toái mà còn tiềm ẩn nhiều rủi ro bảo mật:
- Nhiều Email chứa đường dẫn đến các trang web lừa đảo (phishing), nơi người dùng bị lừa cung cấp thông tin.
- Email Spam thường chứa các tệp đính kèm độc hại như virus, trojan, ...
- Nhiều Email Spam với nội dung chào mời đầu tư, quảng cáo kinh doanh giảm mạo, nhằm moi tiền người dùng.

Tổng quan

Sử dụng tập dữ liệu đã được chuẩn bị

Mô hình bắt đầu quá trình phân loại

Mô hình cho ra kết quả cuối cùng



Hình 1: Tổng quan về phân loại thư rác

Mô tả

1. Thu thập và chuẩn bị dữ liệu

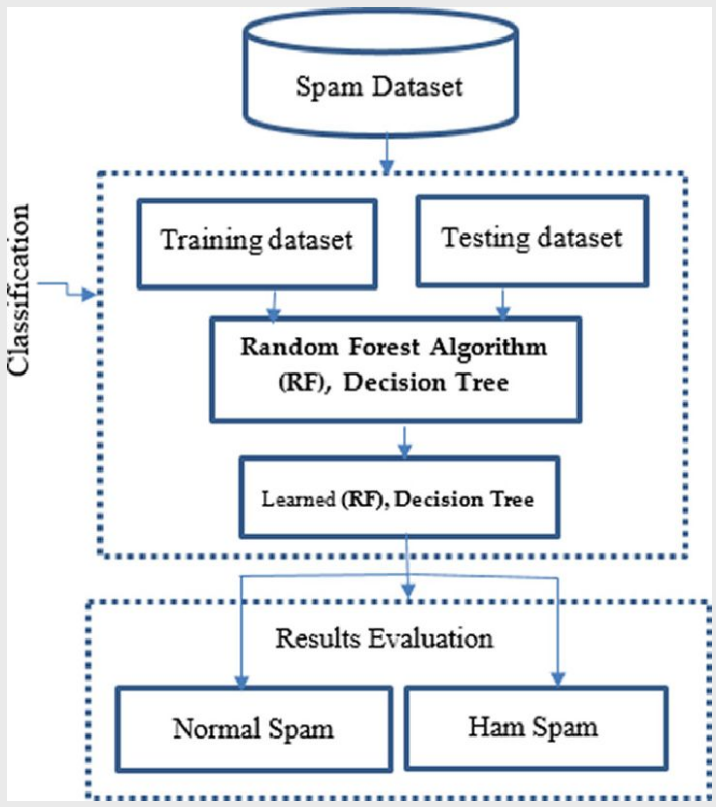
- Thu thập dữ liệu: tập hợp nguồn dữ liệu Email bao gồm hợp lệ và Spam. Có thể lấy từ nguồn cá nhân, cơ sở dữ liệu công khai và các tập dữ liệu trong các cuộc thi học máy.
- Chuẩn bị dữ liệu: làm sạch dữ liệu bằng cách loại bỏ các ký tự đặc biệt, dấu câu và các phần không cần thiết khác. Biến đổi các Email thành dạng có thể xử lý được.

2. Trích xuất đặc trưng

- Vector hóa: chuyển đổi văn bản Email thành các vector đặc trưng. Có thể sử dụng các kỹ thuật sau: Bag of Words (BoW), Embedding.
- Lựa chọn đặc trưng: lựa chọn các đặc trưng quan trọng để giảm kích thước của vector đặc trưng, giúp mô hình hoạt động hiệu quả hơn.

3. Xây dựng và đánh giá

- Xây dựng: mô hình sẽ được sử dụng thuật toán Random Forest. Huấn luyện mô hình trên tập huấn luyện.
- Kiểm tra mô hình: sử dụng tập kiểm tra để đánh giá hiệu suất của mô hình dựa trên các chỉ số như: độ chính xác, độ nhạy, độ đặc hiệu và điểm F1.



Hình 2: Mô tả quá trình của mô hình