THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):
  https://youtu.be/t7a6kTlFAQc
- Link slides (dạng .pdf đặt trên Github):
  mrbordean/CS2205.MAR2024: 230202002-LeAnhCuong (github.com)
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

• Họ và Tên: Lê Anh Cường • MSSV: 230202002 	• Lớp: CS2205.CH181 • Tự đánh giá (điểm tổng kết môn): 8/10 • Số buổi vắng: 1 • Số câu hỏi QT cá nhân: 3 • Link Github: mrbordean/CS2205.CH181: 230202002-LeAnhCuong (github.com) • Link youtube:

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÂN LOẠI THƯ RÁC SỬ DỤNG RANDOM FOREST

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

SPAM CLASSIFICATION USING RANDOM FOREST

TÓM TẮT *(Tối đa 400 từ)*

- Phân loại thư rác là một vấn đề quan trọng trong lĩnh vực học máy, nhằm bảo vệ người dùng Email khỏi những thông điệp không mong muốn và tiềm ẩn nguy cơ mất an toàn thông tin. Một phương pháp hiệu quả để giải quyết vấn đề này là sử dụng thuật toán Random Forest. Đây là một mô hình học máy mạnh mẽ, hoạt động bằng cách xây dựng một tập hợp các cây quyết định (decision trees) và tổng hợp dự đoán từ từng cây để đưa ra kết quả cuối cùng. Khi áp dụng vào phân loại thư rác, thuật toán này có khả năng xử lý các đặc trưng phức tạp của nội dung Email, từ đó phân biệt chính xác giữa thư rác và thư hợp lệ. Nhờ vào độ chính xác cao và khả năng xử lý tốt các dữ liệu lớn và đa dạng, Random Forest trở thành một lựa chọn lý tưởng cho các hệ thống lọc thư rác hiện đại.

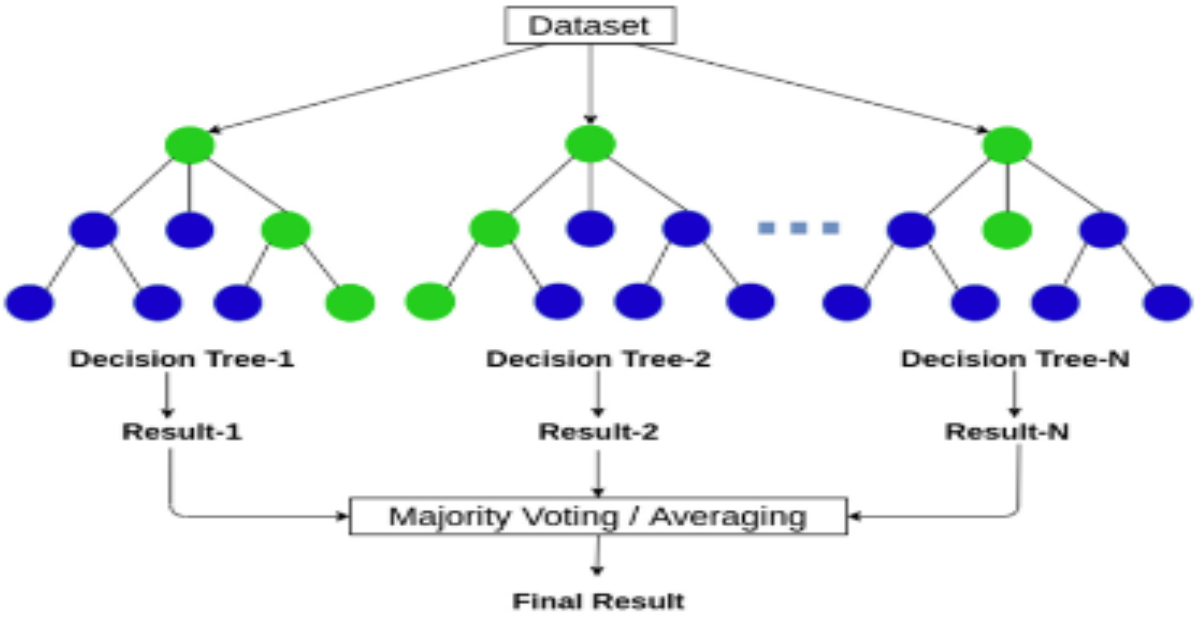
GIỚI THIỆU *(Tối đa 1 trang A4)*

- Thư rác (Email Spam) là các Email không mong muốn được gửi đến một số lượng lớn người nhận với nhiều mục đích khác nhau như quảng cáo, lừa đảo hoặc phát tán mã độc. Những Email này thường không được người nhận đăng ký và có thể gây phiền toái, mất thời gian và nguy cơ bảo mật cho người nhận.
- Một số đặc điểm của Email Spam là:
    - Thường được gửi đến người nhận mà không có sự đồng ý trước.
    - Thường được gửi với số lượng lớn nhằm mục tiêu tiếp cận nhiều người nhất có thể.
    - Nội dung của Email Spam có thể bao gồm quảng cáo sản phẩm, dịch vụ,

các chiến dịch marketing, hoặc các nỗ lực lừa đảo (phishing).

- Email Spam thường xuất phát từ các địa chỉ Email không xác định hoặc giả mạo.

- Thống kê từ Statista vào năm 2023 cho thấy, có tới 45.6% lưu lượng Email toàn cầu được đánh giá là Email Spam [1]. Và chúng ta đều biết rằng hầu hết các cuộc tấn công Phishing đều xuất phát từ Email. Trong đó có thể là một file đính kèm mã độc, hoặc là một tệp lệnh thực thi,... Có rất nhiều nguy hiểm tiềm tàng trong Email Spam mà bất kì người dùng nào cũng có thể là nạn nhân của chúng.
- Các kỹ thuật phân loại đang làm xuất sắc công việc của chúng trên nhiều bộ dữ liệu khác nhau có thể kể đến như Neural Networks (NN), Support Vector Machines (SVM) and Naive Bayesian (NB). Tuy nhiên trong bài viết này, chúng ta sẽ thảo luận về việc phân loại bằng kỹ thuật Random Forest (RF).[2]



Hình 1: Random Forest

- Khi so sánh với các phương pháp học máy khác, Random Forest có những ưu điểm tốt hơn các thuật toán học máy khác để sử dụng vào chủ đề Spam Classification, trong đó có thể kể đến như độ chính xác cao hơn cả đối với thuật toán SVM hay NB; giảm thiểu lỗi phân loại, giúp tăng hiệu suất tổng thể của hệ

thống lọc; xử lý tốt các bộ dữ liệu mất cân bằng cũng như dữ liệu bị thiếu; tốc độ huấn luyện thường nhanh hơn so với SVM và NN; đồng thời thuật toán cũng hoạt động tốt trong các tập dữ liệu lớn và phức tạp.[3]

- Input: tập dataset chứa thư rác và thư bình thường.
- Output: phân loại được đâu là thư rác, đâu là thư bình thường.

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Xây dựng một mô hình phân loại Email bằng thuật toán Random Forest.
- Huấn luyện mô hình để có thể cho ra kết quả tốt nhất từ tập dataset Spam Email của Kaggle.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

Nội dung thực hiện:

1. Thu thập và chuẩn bị dữ liệu:
    a. 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.
    b. 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:
    a. 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.
    b. 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. Chia dữ liệu thành tập huấn luyện và tập kiểm tra:
    a. Chia tập dữ liệu đã được chuẩn bị thành hai phần, một phần để huấn luyện mô hình và một phần để kiểm tra hiệu quả của mô hình.

4. Xây dựng và đánh giá mô hình Random Forest:
    a. Xây dựng: mô hình sẽ được sử dụng thuật toán Random Forest. Huấn luyện mô hình trên tập huấn luyện.
    b. 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.

Phương pháp thực hiện:

1. Sử dụng tập dataset đến từ Kaggle. [4]
2. Sử dụng thuật toán Random Forest để xây dựng mô hình.
3. Huấn luyện mô hình lọc Email Spam và đánh giá kết quả của mô hình.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Mô hình cho ra kết quả khả quan với những chỉ báo như độ chính xác cao, khả năng xử lý dữ liệu mất cân bằng tốt và thời gian huấn luyện hợp lý. Có khả năng phân loại chính xác các Email thành thư rác hay không.
- Có thể áp dụng cho người dùng thực tế.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Spam e-mail traffic share 2023 | Statista

[2]. (PDF) Classification of Spam Messages using Random Forest Algorithm (researchgate.net)

[3]. Basic Comparison Between RandomForest, SVM, and XGBoost | by Nattapoj Apichardsilkij | Medium

[4]. Spam Email Dataset (kaggle.com)