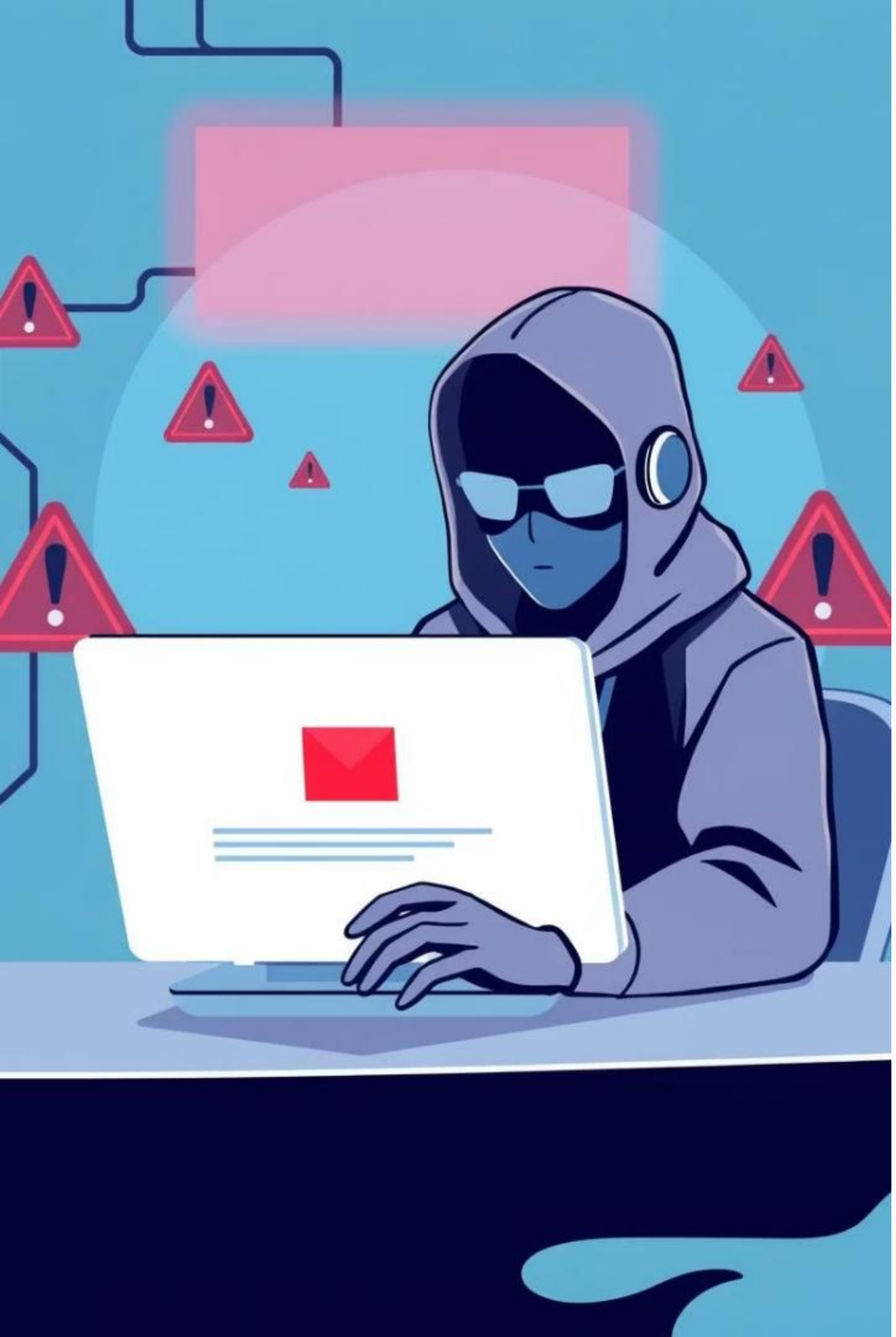
Bài Tập Lớn:
Phân tích và Dự đoán Email chứa Mã độc

Nhóm 4

Phạm Tuấn Dũng – 21010613

Nguyễn Mạnh Cường – 21011583

GVHD: Th.S Nguyễn Văn Thiệu



Tổng quan về Đề tài và Lý do chọn

Tóm tắt đề tài

Nghiên cứu xây dựng hệ thống dự đoán email chứa mã độc hoặc lừa đảo phát hiện sớm với mô hình học máy kết hợp xử lý ngôn ngữ tự nhiên.

Lý do chọn đề tài

Email độc hại ngày càng tinh vi, gây thiệt hại tài chính và uy tín, đòi hỏi giải pháp bảo mật hiệu quả.

Ý nghĩa và ứng dụng

Giúp cá nhân và tổ chức bảo vệ dữ liệu, phát hiện nhanh các mối nguy hiểm qua email, tích hợp công nghệ AI để nâng cao an toàn thông tin.

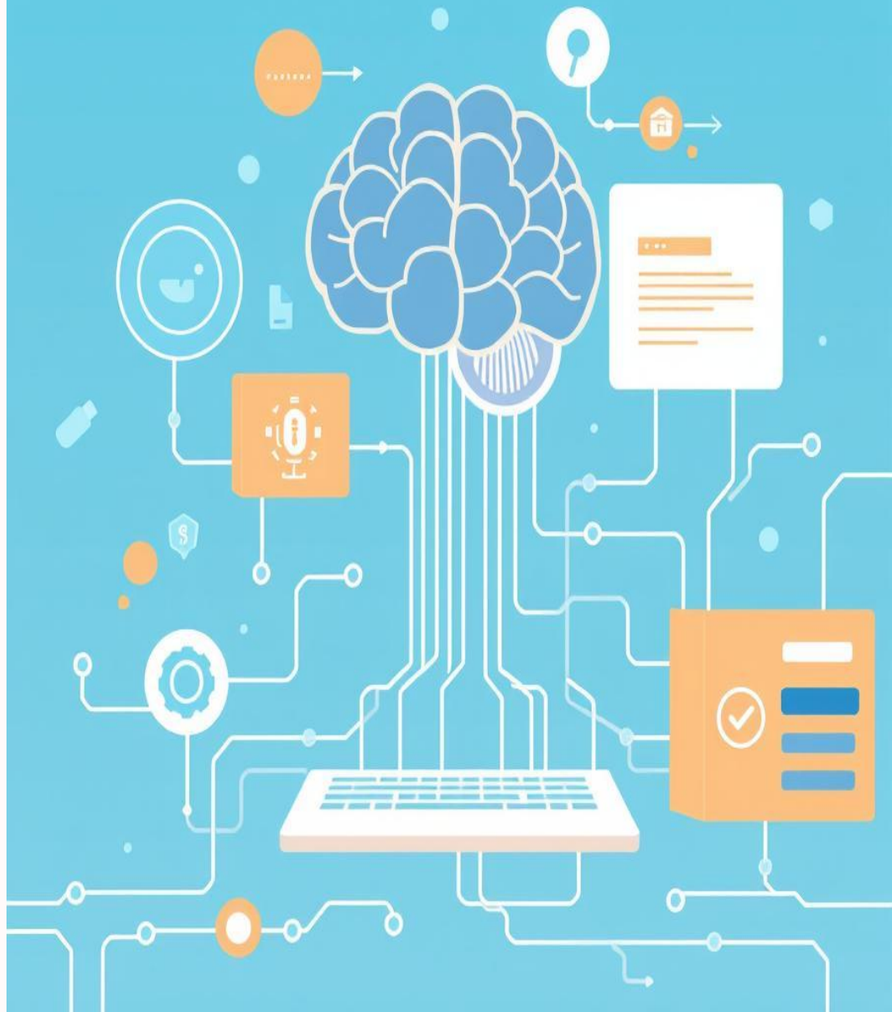
Kiến thức nền tảng và Công nghệ sử dụng

Kỹ thuật học máy

- Logistic Regression: Phân loại nhị phân, tính xác suất sự kiện
- TF-IDF Vectorizer: Biến đổi văn bản thành đặc trưng số hóa
- Regex: Trích xuất từ khóa, URL, tệp đính kèm

Công cụ lập trình và thư viện

- Python: Ngôn ngữ chính
- Pandas: Xử lý dữ liệu bảng
- Scikit-learn: Xây dựng và đánh giá mô hình
- Matplotlib & Seaborn: Trực quan hóa dữ liệu



Thu thập và Xử lý Dữ liệu

1

Quá trình thu thập

Dữ liệu email từ file CSV phishing_email.csv, gồm nội dung và nhãn an toàn/mã độc.

2

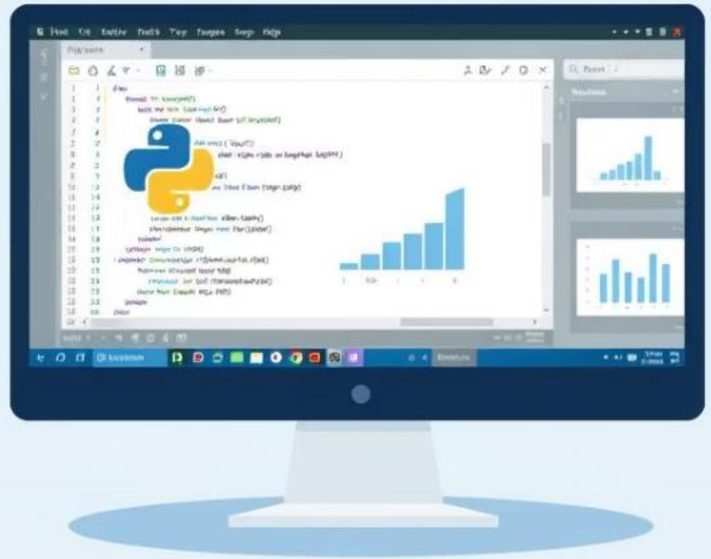
Xử lý dữ liệu

Loại bỏ dữ liệu lỗi, mã hóa nhãn, tách dữ liệu huấn luyện và kiểm tra 80-20.

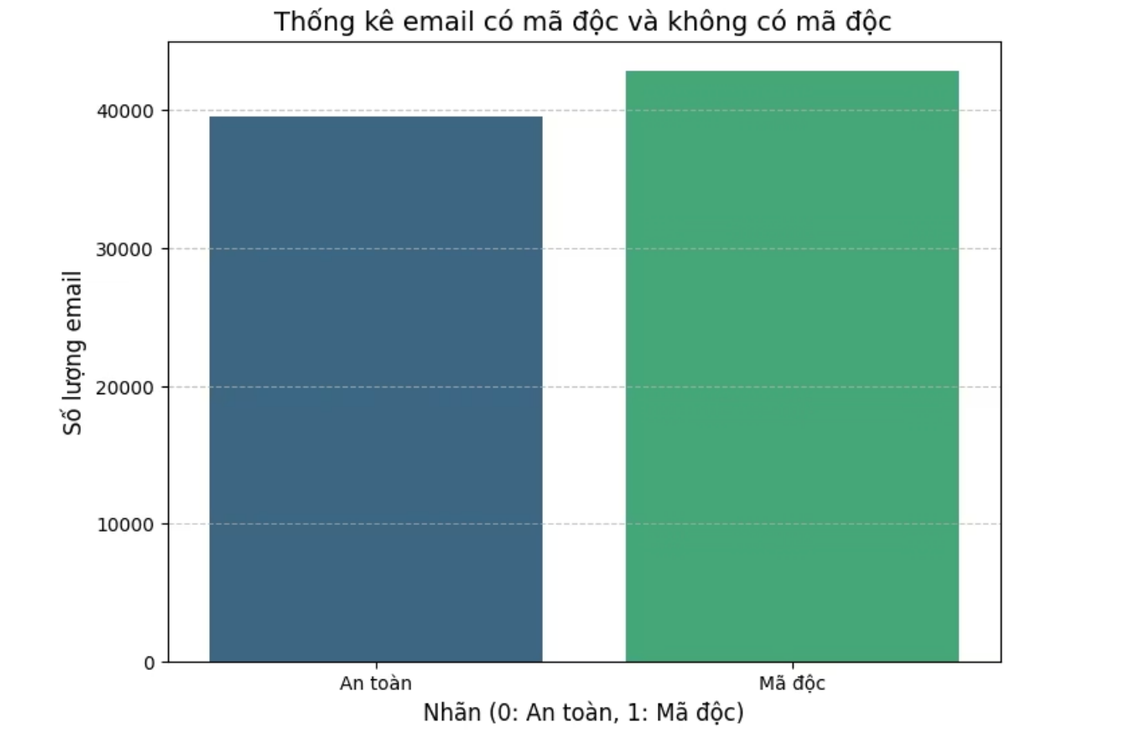
3

Trích xuất đặc trưng

Kết hợp TF-IDF với đặc trưng thủ công như số URL, từ khóa nguy hiểm, tệp đính kèm.



Phân tích Dữ liệu và Đánh giá Mô hình



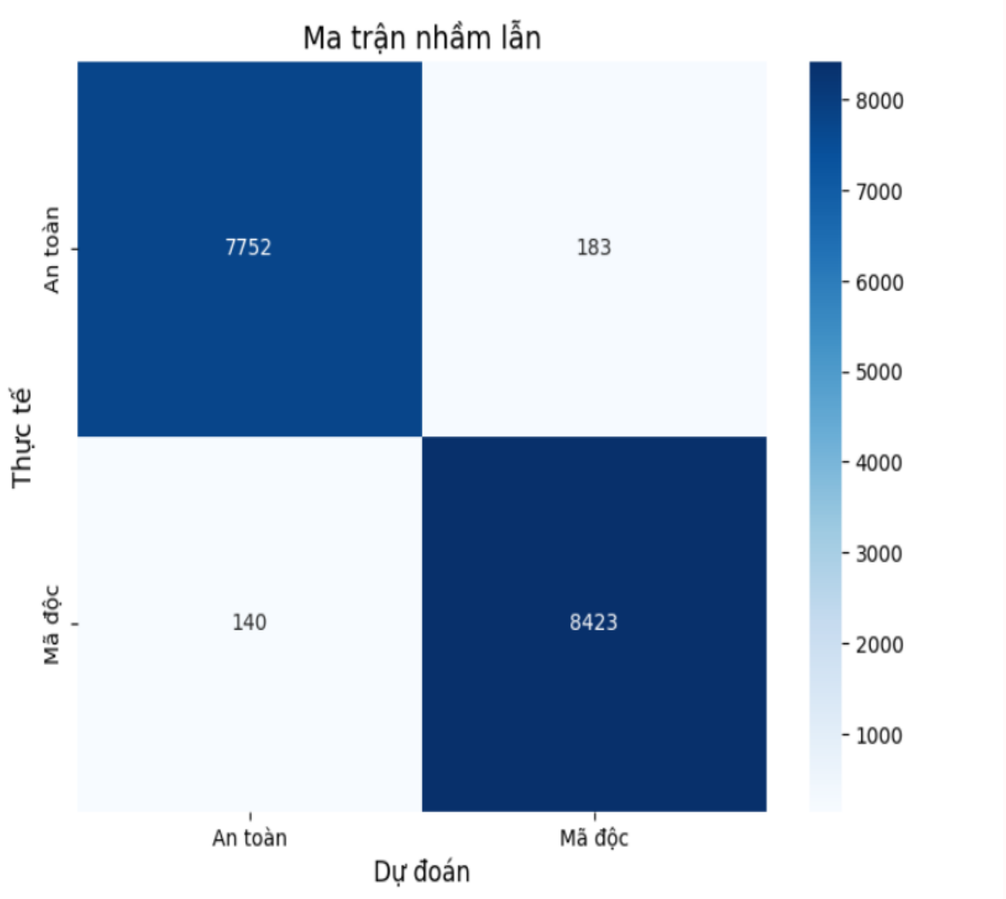
Phân bố dữ liệu

Hơn 80,000 email được gán nhãn, tỷ lệ cân bằng đủ để huấn luyện mô hình hiệu quả.

Đánh giá mô hình

Độ chính xác 98.04%, Precision và Recall đạt 0.98, F1-score cũng tương tự, cho thấy phân loại chính xác cao.

Biểu đồ Ma trận nhầm lẫn và ROC



1

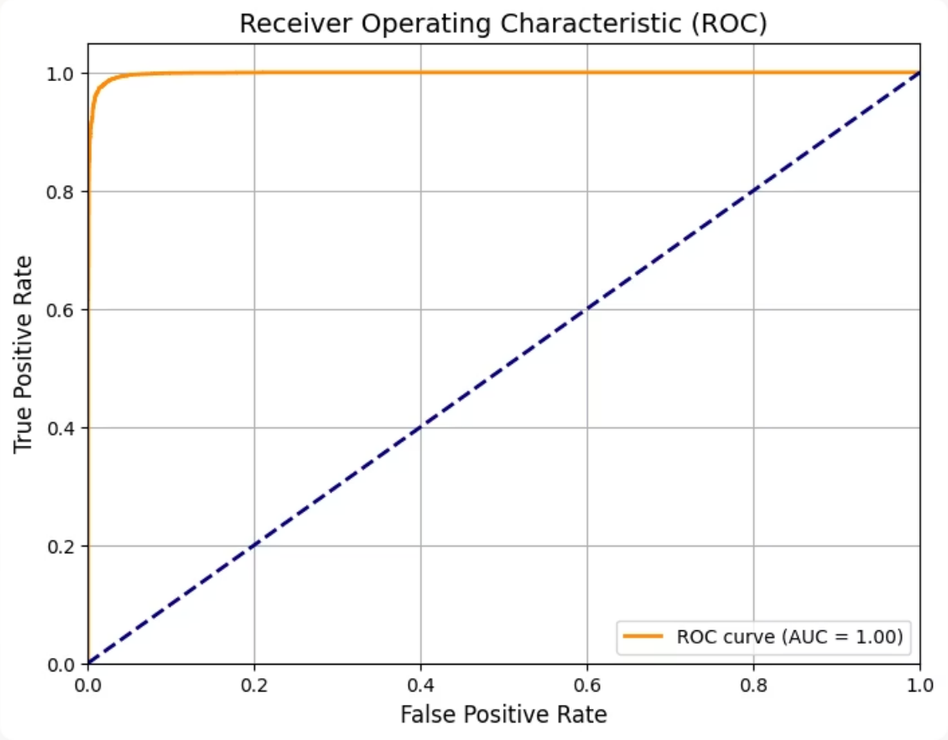
Ma trận nhầm lẫn

Mô hình phân loại chính xác hầu hết các email, với tỷ lệ nhầm lẫn rất thấp giữa hai lớp an toàn và mã độc.

Ma trận nhầm lẫn giúp xác định số lượng các trường hợp dự đoán đúng và sai, bao gồm các trường hợp dương tính thật, dương tính giả, âm tính thật và âm tính giả.

Điều này cho phép đánh giá chi tiết hiệu quả mô hình trong việc phát hiện email chứa mã độc, giảm thiểu cảnh báo sai làm phiền người dùng.

Biểu đồ Ma trận nhầm lẫn và ROC



2

Biểu đồ ROC

AUC đạt 1.00, thể hiện hiệu suất phân biệt hoàn hảo giữa email an toàn và email chứa mã độc.

Đường cong ROC minh họa khả năng phân loại chính xác của mô hình ở các ngưỡng khác nhau. Giá trị AUC gần 1 chứng tỏ mô hình rất tin cậy trong việc phân biệt dương tính thật và âm tính giả, từ đó giảm thiểu cảnh báo sai.

Quy trình Xây dựng Mô hình và Kết quả

Quy trình xây dựng

- Tiền xử lý dữ liệu và trích xuất đặc trưng
- Chia tập huấn luyện và kiểm tra
- Huấn luyện Logistic Regression với tham số cân bằng nhãn

Kết quả

Mô hình đơn giản đạt độ chính xác 98%, phù hợp phân loại email độc hại hiệu quả với độ tin cậy cao.

Tổng kết và Kế hoạch Phát triển

Kết quả đạt được

Hoàn thành xây dựng và đánh giá mô hình dự đoán email chứa mã độc, trực quan hóa kết quả và báo cáo chi tiết.

Ưu nhược điểm

Mô hình hiệu quả cao, đơn giản dễ triển khai nhưng chưa so sánh với các mô hình phức tạp hơn hoặc tối ưu tham số.

Kế hoạch tương lai

Thử nghiệm mô hình khác, áp dụng học sâu và đánh giá trên dữ liệu thực tế để nâng cao khả năng phát hiện email độc hại.