Áp Dụng Máy Học Vào Phân Loại Thư Rác Điện Tử

Nguyễn Phạm Tuấn Khôi, Ngô Thượng Bảo, Nguyễn Trọng Nhân

Tổng quan vấn đề

Trong bối cảnh chuyển đổi số mạnh mẽ, email trở thành phương tiện giao tiếp phổ biến, song cũng là mục tiêu tấn công của các hình thức gian lận tinh vi như thư rác (spam) và email lừa đảo. Với hơn 50% lưu lượng email toàn cầu mỗi ngày là thư rác, việc phát hiện và ngăn chặn hiệu quả các email độc hại là nhu cầu cấp thiết nhằm bảo vệ người dùng và tổ chức khỏi các rủi ro bảo mật thông tin.

Mục tiêu

- Xây dựng mô hình phát hiện gian lận: Phát triển mô hình máy học có khả năng phân loại chính xác giữa email hợp lệ và email gian lận và tối ưu để đạt được hiệu suất cao về độ chính xác, độ nhạy và độ đặc hiệu, giảm thiểu lỗi phân loại.
- - Lựa chọn và trích xuất đặc trưng: Áp dụng kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) để trích xuất thông tin từ nội dung email cùng với đó là sụ kết hợp với siêu dữ liệu để xây dựng tập đặc trưng có giá trị cho mô hình.
- - So sánh và lựa chọn thuật toán tối ưu: Thử nghiệm và đánh giá nhiều thuật toán máy học (Naïve Bayes, SVM, Decision tree, K-nearest neighbor). Cùng với đó là đề xuất ra phương pháp cải tiến nhằm cải thiện hiệu năng, độ chính xác của mô hình

Phương pháp nghiên cứu

- Ứng dụng phương pháp nghiên cứu định lượng kết hợp với nghiên cứu thực nghiệm để đưa ra được kết quả chính xác và trực quan nhất có thể
- Tham khảo tài liệu liên quan tới vấn đề để nắm bắt được các giải pháp hiện tại của bài toán, điểm còn thiếu cũng như điểm mạnh của các giải pháp trước đây
- Sử dụng các thuật toán máy học như Naive Bayes, K-nearest neighbor, Support Vector Machine, mô hình mạng học sâu nơ ron network để phân loại thư điện tử, từ đó đánh giá hiệu quả thông qua các thông số như Accuracy, Precision, Recall

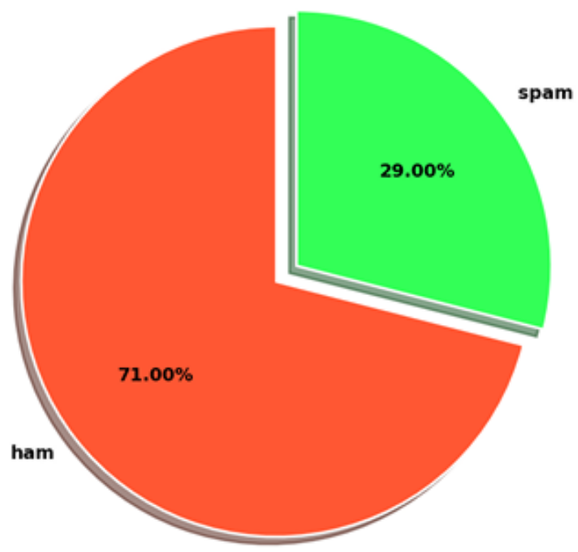
Đối tượng nghiên cứu

Đầu vào: Dữ liệu văn bản từ thư điện tử, tin nhắn văn bản
Đầu ra: Kết quả phân loại văn bản/thư điện tử là hợp lệ hoặc không

Giả thuyết khoa học: Tìm ra được thuật toán phân loại hiệu quả nhất, thử áp dụng mô hình học sâu, mạng nơ ron nhằm giải quyết bài toán phân loại thư điện tử.

Tập dữ liệu

Bộ dữ liệu Email Spam Classification Dataset CSV, bao gồm 5172 email được lựa chọn ngẫu nhiên và phân loại thành thư hợp lệ và thư không hợp lệ, cùng với 3000 từ khóa xuất hiện đa số trong các email.
Đường dẫn: https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv



Tỉ lệ giữa thư hợp lệ (ham) và thư rác (spam)

Kết quả

Thuật toán	Recall (%)	Precision (%)	Accuracy (%)
Naïve Bayes	98.46	99.66	99.46
Support Vector Machine	95.00	93.12	96.90
K-Nearest Neighbor	97.14	87.00	96.20
Neural Network	96.92	96.02	96.83

Mặc dù là thuật toán kém phước tạp và giả định đơn giản nhưng thuật toán Naïve Bayes lại là thuật toán có chỉ số ấn tượng nhất, với các độ đo như Recall, Precision và Accuracy cao vượt trội, cao nhất so với 4 thuật toán được sử dụng trong bài toán phân loại thư điện tử. Đối với mạng Nơ ron network, bù lại cho việc không đứng đầu về các chỉ số, mô hình học sâu này có tốc độ xử lý nhanh nhất trong 4 mô hình

Kết luận

Nghiên cứu ứng dụng thành công mô hình máy học vào việc giải quyết bài toán phân loại thư điện tử, tuy nhiên thì kết quả của mô hình học sâu có phần không như mong đợi, tuy nhiên thì cũng từ đó mà rút ra được kết luận rằng không phải cứ áp dụng mô hình máy học cao vào bài toán thì ta sẽ cải tiến hiệu năng của nó, mà còn phải xét về độ đa dạng và phức tạp của tập dữ liệu liên quan tới bài toán, tùy vào tập dữ liệu và độ phức tạp của bài toán mà ta sẽ ứng dụng các mô hình khác nhau vào việc giải quyết những bài toán ấy