TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT VĨNH LONG

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC KHAI PHÁ DỮ LIỆU

TÊN ĐỀ TÀI: PHÁT HIỆN EMAIL SPAM

HỌC PHẦN MÔ HÌNH KHAI PHÁ DỮ LIỆU

Sinh viên thực hiện:	20004223 – Văn Thị Mỹ Trang
Lớp:	ĐH CNTT 2020
Khóa:	45
Người hướng dẫn:	TS. Phan Anh Cang

Vĩnh Long, năm 2023

MỤC LỤC

LỜI CAM ĐOAN

Tôi xin cam đoan bài báo cáo : “Phát hiện Email Spam” là kết quả của quá trình học tập, nghiên cứu học tập nghiêm túc. Các số liệu trong luận văn là trung thực, có nguồn gốc rõ ràng, được trích dẫn và có tính kế thừa, phát triển từ các sách, tài liệu, tạp chí, các công trình nghiên cứu đã được công bố, và các website, … Các phương pháp nêu trong báo cáo được rút ra từ những cơ sở lý luận và quá trình nghiên cứu tìm hiểu từ các phương pháp đã học.

Vĩnh Long, tháng 06 năm 2023

LỜI CẢM ƠN

Trước hết, chúng em xin chân thành cảm ơn Trường Đại học Sư phạm kỹ thuật Vĩnh Long. Các Thầy, Cô trong Khoa Công nghệ Thông tin đã tạo điều kiện thuận lợi cho chúng trong suốt quá trình học tập và làm tiểu luận.

Em xin bày tỏ lòng biết ơn sâu sắc của mình đối với thầy Phan Anh Cang, người đã giảng dạy và hướng dẫn chúng em làm đề tài.

Em xin cảm ơn các Thầy, Cô đã quan tâm góp ý và nhận xét quý báu cho đề tài.

MỞ ĐẦU

1.Lý do chọn đề tài

Email spam là các email không mong muốn được gửi đến người nhận với mục đích quảng cáo sản phẩm hoặc dịch vụ, lừa đảo hoặc phân phối virus và phần mềm độc hại. Spam email có thể gây ra nhiều phiền toái và nguy hiểm cho người nhận, từ việc làm chậm hộp thư đến mất thông tin cá nhân và tài khoản. Đây là một vấn đề nan giải và cần xây dựng các phương pháp phù hợp để ngăn chặn nhằm bảo vệ người dùng.

Vì thế đề tài "Phát hiện Email Spam" là hết sức thiết thực và phù hợp với bối cảnh ngày nay

2.Mục đích nghiên cứu

Mục đích nghiên cứu của đề tài "Phát hiện email spam" là xây dựng các hệ thống và phương pháp giúp phát hiện và chặn các email spam trước khi chúng đến được hộp thư đến của người dùng. Điều này đóng vai trò quan trọng trong việc bảo vệ người dùng khỏi các email spam có thể gây ra nhiều phiền toái và nguy hiểm như lừa đảo, phân phối virus và phần mềm độc hại, mất thông tin cá nhân và tài khoản.

Mục tiêu của nghiên cứu phát hiện email spam là cải thiện hiệu quả phát hiện và chặn các email spam, giảm thiểu các ảnh hưởng tiêu cực của chúng đến người dùng và cộng đồng mạng. Nghiên cứu này có thể đóng góp quan trọng trong việc phát triển các công cụ và phương pháp chống spam hiệu quả hơn, giúp bảo vệ người dùng khỏi các email spam và tăng cường an ninh mạng.

3.Đối tượng nghiên cứu

–Đối tượng nghiên cứu là các email Spam và email bình thường được thu thập từ nhiều nguồn khác nhau

–Nghiên cứu các thuật toán phù hợp với độ chính xác cao để áp dụng vào mô hình đồng thời phân tích đánh giá kết quả

4. Phạm vi nghiên cứu

–Phạm vi nghiên cứu của đề tài phát hiện email spam có thể bao gồm những khía cạnh sau:

–Phân tích đặc trưng của email spam: Nghiên cứu có thể tập trung vào việc phân tích các đặc trưng của email spam, bao gồm nội dung email, chủ đề, từ khóa email để xác định các yếu tố quan trọng trong việc phát hiện email spam.

–Xây dựng thuật toán phát hiện email spam: Đề tài có thể tập trung vào việc phát triển các thuật toán để phát hiện email spam, bao gồm các phương pháp như học máy, phân tích cú pháp và phân tích nội dung.

–Đánh giá hiệu quả của các phương pháp phát hiện email spam: Nghiên cứu có thể tập trung vào việc đánh giá hiệu quả của các phương pháp phát hiện email spam bằng cách so sánh với các phương pháp với nhau và đo lường độ chính xác, độ nhạy và độ đặc hiệu của các phương pháp.

–Ứng dụng thực tế: Nghiên cứu có thể tập trung vào việc áp dụng các phương pháp phát hiện email spam vào các hệ thống bảo mật mạng thực tế để bảo vệ người dùng khỏi các email spam.

5. Phương pháp nghiên cứu

5.1. Phương pháp nghiên cứu lý thuyết

–Tìm hiểu các nghiên cứu liên quan: Ở bước này, nghiên cứu sẽ tìm hiểu các nghiên cứu trước đây về phát hiện email spam, bao gồm cả các phương pháp và thuật toán phát hiện email spam đã được sử dụng trong các nghiên cứu trước đó.

–Xác định các khái niệm và định nghĩa: Trong bước này, nghiên cứu sẽ xác định các khái niệm liên quan đến đề tài, bao gồm các khái niệm về email spam, các đặc trưng của email spam và các phương pháp phát hiện email spam.

–Phân tích và đánh giá các phương pháp phát hiện email spam: Ở bước này, nghiên cứu sẽ phân tích và đánh giá các phương pháp phát hiện email spam đã được sử dụng trong các nghiên cứu trước đó. Nghiên cứu sẽ xác định những ưu điểm và hạn chế của các phương pháp này và đề xuất các cải tiến hoặc phương pháp mới để phát hiện email spam.

–Xây dựng mô hình và thuật toán phát hiện email spam: Ở bước này, nghiên cứu sẽ xây dựng mô hình và thuật toán phát hiện email spam dựa trên các phương pháp và kỹ thuật đã được phân tích và đánh giá ở bước trước.

–Đánh giá và so sánh hiệu quả của các phương pháp phát hiện email spam: Ở bước này, nghiên cứu sẽ đánh giá và so sánh hiệu quả của các phương pháp phát hiện email spam đã được xây dựng.

–Đề xuất các hướng nghiên cứu tiếp theo: Ở bước này, nghiên cứu sẽ đề xuất các hướng nghiên cứu tiếp theo để cải thiện hiệu quả phát hiện và chặn email spam. Các hướng nghiên cứu này có thể bao gồm sử dụng các kỹ thuật mới như học sâu và mạng

lưới thần kinh để phát hiện email spam hoặc phát triển các giải pháp kết hợp nhiều phương pháp để cải thiện hiệu quả phát hiện và chặn email spam.

5.2. Phương pháp nghiên cứu thực nghiệm.

−Chuẩn bị dữ liệu: Ở bước này, nghiên cứu sẽ thu thập và chuẩn bị dữ liệu cho quá trình phát hiện email spam. Dữ liệu này bao gồm các email spam và các email không phải spam để dùng để huấn luyện và kiểm tra các thuật toán phát hiện email spam.

−Xây dựng mô hình và thuật toán phát hiện email spam: Ở bước này, nghiên cứu sẽ xây dựng mô hình và thuật toán phát hiện email spam dựa trên các phương pháp và kỹ thuật đã được phân tích và đánh giá trong phương pháp nghiên cứu lý thuyết. Sau đó, nghiên cứu sẽ tiến hành huấn luyện mô hình và thuật toán trên dữ liệu được chuẩn bị ở bước trước.

−Đánh giá hiệu quả của mô hình: Ở bước này, nghiên cứu sẽ đánh giá hiệu quả của mô hình và thuật toán phát hiện email spam bằng cách sử dụng dữ liệu kiểm tra. Nghiên cứu sẽ đo lường độ chính xác của mô hình và thuật toán và so sánh với các phương pháp với nhau.

−Đề xuất các hướng nghiên cứu tiếp theo: Ở bước này, nghiên cứu sẽ đề xuất các hướng nghiên cứu tiếp theo để cải thiện hiệu quả phát hiện và chặn email spam trong tương lai.

6. Ý nghĩa khoa học và thực tiễn của đề tài

6.1. Ý nghĩa khoa học

−Đóng góp vào việc mở rộng kiến thức và hiểu biết về các phương pháp và kỹ thuật phát hiện email spam.

−Góp phần vào việc phát triển các thuật toán và mô hình phát hiện email spam mới hoặc cải tiến các phương pháp hiện có.

−Tìm hiểu các xu hướng mới của email spam và đưa ra các giải pháp phát hiện mới để đối phó với những loại email spam mới này.

6.2. Ý nghĩa thực tiễn

−Việc phát hiện và chặn email spam có thể giúp giảm thiểu sự cố gây hại cho các hệ thống và người dùng, bảo vệ thông tin và dữ liệu quan trọng và tăng cường đáng kể an ninh và bảo mật mạng.

−Nghiên cứu này có thể đưa ra các giải pháp thực tế để phát hiện và chặn email spam trong các hệ thống mạng thực tế và các dịch vụ email, giúp cho doanh nghiệp và cá nhân có thể đảm bảo an toàn và bảo mật thông tin cá nhân của mình.

7. Cấu trúc luận văn

Chương 1 Cơ sở khoa học của đề tài

Chương 2 Cơ sở lý thuyết

Chương 3 Phương pháp nghiên cứu

Chương 4 Kết quả nghiên cứu thực nghiệm

CHƯƠNG 1. CƠ SỞ KHOA HỌC CỦA ĐỀ TÀI

1.1 Cơ sở lý luận của đề tài

Phát hiện email spam là một lĩnh vực nghiên cứu trong lĩnh vực trí tuệ nhân tạo và khoa học máy tính. Các phương pháp phát hiện email spam dựa trên các kỹ thuật xử lý ngôn ngữ tự nhiên, học máy và phân loại dữ liệu.

Xử lý ngôn ngữ tự nhiên: Để phát hiện email spam, ta cần phân tích cú pháp và nội dung của email. Xử lý ngôn ngữ tự nhiên (NLP) cung cấp các công cụ và kỹ thuật để phân tích và hiểu các đoạn văn bản tự nhiên.

Học máy: Học máy là một phương pháp giúp máy tính học và cải thiện hiệu suất của nó trong việc phân loại dữ liệu. Học máy được sử dụng rộng rãi trong phát hiện email spam.

Phân loại dữ liệu: Để phát hiện email spam, ta cần phân loại các email vào hai nhóm: email hợp lệ và email spam. Phân loại dữ liệu là một trong những phương pháp chính được sử dụng để phát hiện email spam.

Mạng nơ-ron: Mạng nơ-ron là một phương pháp học máy được sử dụng rộng rãi trong phát hiện email spam. Mạng nơ-ron có khả năng học và cải thiện hiệu suất của mình theo thời gian.

Kỹ thuật rút trích đặc trưng: Kỹ thuật rút trích đặc trưng là một phương pháp để chuyển đổi dữ liệu đầu vào thành các đặc trưng có ý nghĩa. Kỹ thuật này được sử dụng trong phát hiện email spam để tạo ra các đặc trưng từ nội dung email và tiêu đề.

Kỹ thuật xử lý tín hiệu số: Kỹ thuật xử lý tín hiệu số được sử dụng để phân tích các thuộc tính tín hiệu, chẳng hạn như tần số, biên độ và pha. Kỹ thuật này được sử dụng trong phát hiện email spam để phân tích các thuộc tính của email và xác định liệu email đó có phải là spam hay không.

Những cơ sở lý thuyết trên cùng với các kỹ thuật và công nghệ tương ứng được áp dụng để phát hiện email spam hiệu quả.

1.2 Cơ sở thực tiễn của đề tài

Email spam là vấn đề bảo mật mạng và quản lý email. Email spam là một vấn đề đáng lo ngại trong việc quản lý email và bảo mật mạng, vì nó có thể chứa các thông tin độc hại hoặc gây phiền nhiễu cho người dùng. Điều này đã thúc đẩy sự phát triển của các phương pháp và kỹ thuật phát hiện email spam.

Các doanh nghiệp và tổ chức phải đối mặt với hàng ngàn email spam mỗi ngày, và việc phát hiện và chặn email spam trở thành một nhu cầu thiết yếu. Việc phát hiện và chặn email spam không chỉ giảm thiểu sự cố gây hại cho các hệ thống và người dùng, bảo vệ thông tin và dữ liệu quan trọng, mà còn giúp tăng cường đáng kể an ninh và bảo mật mạng.

1.3.Tổng quan các công trình nghiên cứu liên quan

1.3.1. Các công trình nghiên cứu trên thế giới

Mô hình mới tự động phát hiện và lọc thư rác của các nhà nghiên cứu tại Viện Công nghệ Sinhgad Lonavala ở Ấn Độ. Kỹ thuật này có thể giúp cải thiện tính năng bảo mật người dùng, đồng thời, giúp họ đọc lướt các email không liên quan hoặc không mong đợi. Mô hình mới dựa trên việc lựa chọn tính năng đa mục tiêu và một mạng lưới thích ứng, kỹ thuật học sâu mới có triển vọng cao. Trái ngược với các phương pháp trước đây, mô hình được đào tạo trên cả tập dữ liệu hình ảnh và văn bản.

1.3.2. Các công trình nghiên cứu tại Việt Nam

Nghiên cứu các phương pháp lọc thư rác tại Việt Nam và thế giới, xây dựng và đề xuất phương pháp lọc thư rác tiếng Việt của tác giả Lâm Tăng Doan.

CHƯƠNG 2 CƠ CỞ LÝ THUYẾT

2.1 Tìm hiểu về Email Spam

Khái niệm: Email spam là những email không mong muốn, thường chứa các thông điệp quảng cáo hoặc lừa đảo, gửi đến người dùng mà không được yêu cầu hoặc chấp nhận trước đó.

Mục đích: Mục đích của email spam thường là quảng cáo, tuy nhiên, nó cũng được sử dụng để lừa đảo, phát tán virus hoặc tấn công các hệ thống máy tính.

Đặc trưng:

−Tiêu đề không liên quan hoặc nghi ngờ: Email spam thường có tiêu đề không liên quan hoặc nghi ngờ, nhằm thu hút sự chú ý của người nhận.

−Nội dung không mong muốn: Nội dung của email spam thường là quảng cáo hoặc thông điệp lừa đảo, không có giá trị thực sự cho người nhận.

−Địa chỉ gửi không rõ ràng: Email spam thường được gửi từ địa chỉ email không rõ ràng hoặc không xác định được nguồn gốc.

Hậu quả: Đánh cắp thông tin, tống tiền, gây stress và căng thẳng cho người dùng.

2.2 Tổng quan về khai phá dữ liệu

Khai phá dữ liệu (data mining) là quá trình tìm kiếm thông tin ẩn trong các cơ sở dữ liệu lớn và phức tạp, thông qua việc sử dụng các kỹ thuật và phương pháp phân tích dữ liệu. Mục đích của khai phá dữ liệu là tìm ra các mẫu, quy luật và thông tin hữu ích từ dữ liệu, giúp cho việc ra quyết định và dự đoán trong các lĩnh vực khác nhau.

Các ứng dụng của khai phá dữ liệu bao gồm:

−Phát hiện gian lận và lừa đảo: Khai phá dữ liệu có thể giúp phát hiện các hoạt động gian lận hoặc lừa đảo trong các giao dịch tài chính hoặc thương mại điện tử.

−Dự đoán và tối ưu hóa: Khai phá dữ liệu có thể được sử dụng để dự đoán kết quả hoặc tối ưu hóa các quy trình và quyết định trong các lĩnh vực khác nhau.

−Phát hiện spam và phân loại email: Khai phá dữ liệu có thể giúp phân loại email là spam hoặc không phải spam.

−Phân tích dữ liệu y tế: Khai phá dữ liệu có thể được sử dụng để phân tích dữ liệu y tế và tìm ra các mẫu và quy luật giúp cho việc chẩn đoán và điều trị bệnh.

2.3 Tổng quan về ngôn ngữ python

Python là một ngôn ngữ lập trình đa năng, dễ học và sử dụng, được phát triển bởi Guido van Rossum vào năm 1991. Python được thiết kế để có thể đọc được như các

ngôn ngữ tự nhiên, với cú pháp đơn giản và rõ ràng. Python được sử dụng rộng rãi trong các lĩnh vực như khoa học dữ liệu, trí tuệ nhân tạo, web development, game và hệ điều hành.

2.4.Thuật toán Decision Tree (Gini)

Thuật toán Decision Tree (Gini) được sử dụng để xây dựng cây quyết định dựa trên hàm Gini để tối thiểu hóa sự không đồng nhất trong các nhóm dữ liệu. Thuật toán này có thể được sử dụng cho cả bài toán phân loại và dự đoán. Điều quan trọng là lựa chọn các siêu tham số phù hợp để tránh tình trạng overfitting hoặc underfitting. Decision Tree (Gini) có ưu điểm là dễ hiểu, dễ áp dụng và cho kết quả khá tốt trên các dữ liệu có cấu trúc đơn giản.

2.5. Thuật toán Decision Tree (entropy)

Thuật toán Decision Tree (entropy) cũng là một phương pháp xây dựng cây quyết định, nhưng sử dụng hàm entropy để tối thiểu hóa sự không đồng nhất trong các nhóm dữ liệu. Nó cũng có thể được sử dụng cho cả bài toán phân loại và dự đoán. Decision Tree (entropy) có những ưu điểm tương tự như Decision Tree (Gini), tuy nhiên, nó có thể cho ra kết quả chính xác hơn đối với các dữ liệu phức tạp.

2.6. Thuật toán Logistic Regression

Thuật toán Logistic Regression là một phương pháp học có giám sát phổ biến trong bài toán phân loại. Nó sử dụng hàm sigmoid để dự đoán xác suất cho mỗi lớp và tối ưu hóa các tham số bằng cách sử dụng phương pháp gradient descent. Logistic Regression có ưu điểm là đơn giản, dễ giải thích và độ chính xác cao nếu dữ liệu được xử lý đúng.

2.7. Thuật toán Random Forest

Random Forest là một thuật toán học máy dùng cho các tác vụ phân loại, hồi quy và phát hiện bất thường. Nó được đặc trưng bởi việc sử dụng nhiều cây quyết định độc lập để tạo ra một mô hình phân loại hoặc hồi quy chính xác hơn.

Thuật toán Random Forest hoạt động bằng cách xây dựng nhiều cây quyết định độc lập trên các tập con của dữ liệu huấn luyện được chọn ngẫu nhiên. Mỗi cây quyết định được xây dựng bằng cách chọn ngẫu nhiên một tập con của các đặc trưng đầu vào. Các cây quyết định được xây dựng bằng cách chia tập dữ liệu theo các cách khác nhau và chọn cách tốt nhất để giảm thiểu độ không chính xác.

Khi có một mẫu mới cần được phân loại, Random Forest sử dụng tất cả các cây quyết định để đưa ra dự đoán. Kết quả cuối cùng được chọn dựa trên số phiếu bầu. Nói cách khác, mẫu mới sẽ được phân loại vào lớp nào có số phiếu bầu lớn nhất.

Lợi ích của Random Forest là nó có khả năng giảm thiểu overfitting và cải thiện độ chính xác so với một cây quyết định đơn lẻ. Điều này là do việc sử dụng nhiều cây quyết định độc lập nhau. Ngoài ra, Random Forest có khả năng xử lý các tập dữ liệu lớn và có thể xử lý cả các biến đầu vào liên tục và rời rạc.

Tuy nhiên, Random Forest có một số hạn chế. Nó cần nhiều thời gian để huấn luyện hơn so với một cây quyết định đơn lẻ và đòi hỏi nhiều tài nguyên tính toán. Ngoài ra, một số cây quyết định có thể có kết quả không tốt, dẫn đến kết quả dự đoán không chính xác.

2.8 Đánh giá mô hình

– **Decision Tree:**

Là một thuật toán học có giám sát, được sử dụng để phân loại và dự đoán giá trị của biến mục tiêu bằng cách phân tách tập dữ liệu thành các nhóm con dựa trên các thuộc tính đầu vào.

Decision Tree (Gini) và Decision Tree (entropy) là hai phương pháp chính để tính độ không thuần khiết của các nhóm con (node) trong cây quyết định.

Decision Tree (Gini) sử dụng chỉ số Gini Impurity để tính độ không thuần khiết của các node, trong khi Decision Tree (entropy) sử dụng chỉ số Entropy để tính độ không thuần khiết.

Decision Tree (Gini) và Decision Tree (entropy) có thể được sử dụng để xây dựng cây quyết định cho các bài toán phân loại và dự đoán.

– **Thuật toán Logistic Regression:**

Logistic Regression là một thuật toán học có giám sát dùng để dự đoán giá trị của biến phụ thuộc (có giá trị rời rạc hoặc liên tục) bằng cách sử dụng một hoặc nhiều biến độc lập (có giá trị liên tục hoặc rời rạc) và một hàm logistic.

Logistic Regression là một phương pháp phân loại tuyến tính, nghĩa là nó sử dụng một hàm tuyến tính để tìm ra giá trị xác suất (probability) cho mỗi lớp phân loại, sau đó áp dụng hàm logistic để chuyển đổi giá trị xác suất này thành giá trị dự đoán cho lớp phân loại.

Logistic Regression thường được sử dụng trong các bài toán phân loại nhị phân (binary classification) hoặc đa lớp (multiclass classification).

– **Thuật toán Random Forest:**

Random Forest là một thuật toán học có giám sát dùng để dự đoán giá trị của biến mục tiêu bằng cách sử dụng nhiều cây quyết định (decision tree) độc lập nhau.

Random Forest sử dụng phương pháp bagging (bootstrap aggregating) để xây dựng một tập hợp các cây quyết định, trong đó mỗi cây được huấn luyện trên một tập con của dữ liệu ban đầu (lấy ngẫu nhiên với hoàn lại).

Kết quả dự đoán của Random Forest là sự trung bình hoặc phiếu phiếu bầu (voting) của các cây quyết định trong tập hợp.

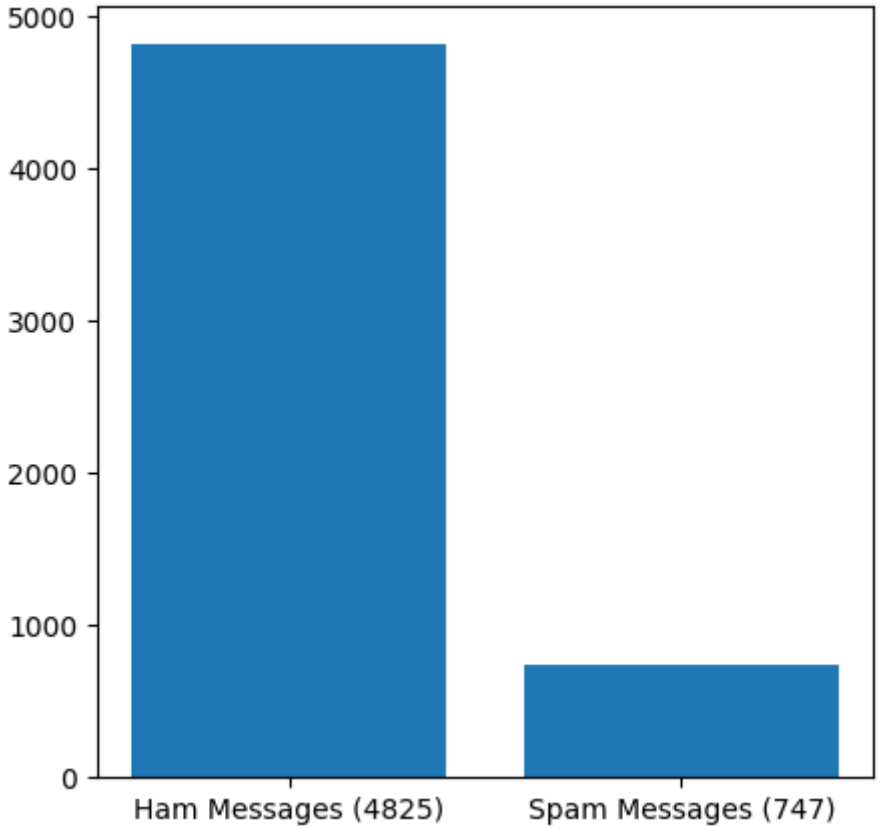
Random Forest thường được sử dụng trong các bài toán phân loại và dự đoán, đặc biệt là trong các bài toán có số lượng thuộc tính (features) lớn hoặc có tính tương quan cao giữa các thuộc tính. Thuật toán này có thể giúp giảm thiểu các vấn đề liên quan đến overfitting và variance của cây quyết định đơn lẻ.

CHƯƠNG 3 PHƯƠNG PHÁP ĐỀ XUẤT

3.1.Đặc điểm dữ liệu

Bộ dữ liệu SPAMtextmessage20170820 được lấy từ Kaggle với 5572 email gồm nhãn ham và spam và cột dữ liệu là email

Trong đó có 4825 email bình thường và 747 email spam



Hình 3.1. Hình ảnh bộ dữ liệu

3.2.Thuật toán đề xuất

- Thuật toán Decision Tree (Gini)
- Thuật toán Decision Tree (entropy)
- Thuật toán Logistic Regression
- Thuật toán Random Forest

3.3 Phương pháp đề xuất

3.3.1. Giai đoạn huấn luyện mô hình

Chia tập dữ liệu 20% test, 80% train. Sau đó áp dụng các thuật toán Decision Tree (Gini), Decision Tree (entropy), Logistic Regression, Random Forest.

Thuật toán Decision Tree (Gini):

Khởi tạo một đối tượng DecisionTreeClassifier với các tham số cấu hình:

−criterion="gini": sử dụng tiêu chí Gini để đánh giá chất lượng phân chia.

−random_state=123: thiết lập giá trị seed cho việc tạo số ngẫu nhiên, để đảm bảo kết quả của mô hình được tái lặp được trên các lần chạy khác nhau.

−max_depth=10: giới hạn độ sâu của cây quyết định, để tránh tình trạng quá khớp.

−min_samples_leaf=8: định nghĩa số lượng điểm dữ liệu tối thiểu cần phải có trong mỗi lá của cây.

−Sử dụng phương thức fit trên đối tượng DecisionTreeClassifier để huấn luyện mô hình trên tập dữ liệu huấn luyện (xv_train và y_train).

−Sau khi huấn luyện xong, mô hình sẽ được lưu vào đối tượng model_dt_gini.

Thuật toán Decision Tree (entropy)

Khởi tạo một đối tượng DecisionTreeClassifier với các tham số cấu hình như sau:

−criterion="entropy": sử dụng tiêu chí entropy để đánh giá chất lượng phân chia.

−random_state=123: thiết lập giá trị seed cho việc tạo số ngẫu nhiên, để đảm bảo kết quả của mô hình được tái lặp được trên các lần chạy khác nhau.

−max_depth=10: giới hạn độ sâu của cây quyết định, để tránh tình trạng quá khớp.

−min_samples_leaf=8: định nghĩa số lượng điểm dữ liệu tối thiểu cần phải có trong mỗi lá của cây.

−Sử dụng phương thức fit trên đối tượng DecisionTreeClassifier để huấn luyện mô hình trên tập dữ liệu huấn luyện (xv_train và y_train).

−Sau khi huấn luyện xong, mô hình sẽ được lưu vào đối tượng model_dt_entropy.

Thuật toán Logistic Regression

−Đầu tiên, một đối tượng LogisticRegression được tạo với tham số random_state được đặt là 0. Đối tượng này sẽ được sử dụng để huấn luyện mô hình hồi quy logistic.

−Tiếp theo, mô hình được huấn luyện bằng cách sử dụng phương thức fit(), trong đó đầu vào là dữ liệu huấn luyện (xv_train) và nhãn tương ứng (y_train).

−Sau khi mô hình được huấn luyện, nó được lưu vào một tệp có tên là "model_lr.pkl" bằng cách sử dụng module pickle.

Thuật toán Random Forest

Mô hình Random Forest được huấn luyện trên tập dữ liệu huấn luyện xv_train và y_train bằng cách sử dụng lớp RandomForestClassifier trong thư viện sklearn. Mô hình được cấu hình với 100 cây quyết định và độ sâu tối đa của mỗi cây là 100. Tham số random_state được sử dụng để đảm bảo rằng kết quả huấn luyện của mô hình sẽ ổn định qua các lần chạy khác nhau.

Sau khi huấn luyện xong, mô hình được lưu vào tệp model_rf.pkl bằng cách sử dụng module pickle

3.3.2. Giai đoạn kiểm thử mô hình

Thuật toán Decision Tree (Gini):

−Tính toán các thông số đánh giá kết quả dự đoán bằng cách so sánh kết quả dự đoán (y_pred_dtg) với nhãn thật (y_test) bằng các hàm accuracy_score, precision_score, recall_score và f1_score.

−Tính ma trận nhầm lẫn (confusion matrix) bằng cách sử dụng hàm confusion_matrix từ module sklearn.metrics.

−Vẽ ma trận nhầm lẫn dưới dạng heatmap bằng cách sử dụng hàm plot_confusion_matrix từ module mlxtend.plotting.

−Hiển thị kết quả đánh giá và ma trận nhầm lẫn bằng cách sử dụng hàm print.

Thuật toán Decision Tree (entropy)

−Dùng mô hình đã huấn luyện để thực hiện dự đoán trên tập dữ liệu kiểm tra (xv_test) bằng cách sử dụng phương thức predict.

−Tính toán độ chính xác của mô hình bằng cách so sánh kết quả dự đoán (y_pred_dte) với nhãn thật (y_test) bằng hàm accuracy_score.

−Tính ma trận nhầm lẫn (confusion matrix) bằng cách sử dụng hàm confusion_matrix từ module sklearn.metrics.

−Vẽ ma trận nhầm lẫn dưới dạng heatmap bằng cách sử dụng hàm plot_confusion_matrix từ module mlxtend.plotting.

−Hiển thị kết quả đánh giá và ma trận nhầm lẫn bằng cách sử dụng hàm print.

Thuật toán Logistic Regression

−Để đánh giá hiệu suất của mô hình, phương thức predict() được sử dụng để dự đoán nhãn của dữ liệu kiểm tra (xv_test). Sau đó, hàm accuracy_score() từ module sklearn.metrics được sử dụng để tính toán độ chính xác của mô hình trên dữ liệu kiểm tra.

−Ma trận nhầm lẫn được tính toán bằng cách sử dụng hàm confusion_matrix() từ module sklearn.metrics. Sau đó, ma trận nhầm lẫn kết quả được vẽ bằng cách sử dụng hàm plot_confusion_matrix().

Thuật toán Random Forest

Mô hình được đánh giá trên tập dữ liệu kiểm tra xv_test và y_test. Kết quả dự đoán của mô hình được tính bằng cách sử dụng phương thức predict() của mô hình. Sau đó, độ chính xác của mô hình được tính bằng cách so sánh kết quả dự đoán với nhãn thực

tế và tính tỷ lệ phần trăm giống nhau. Cuối cùng, ma trận nhầm lẫn được tính toán bằng cách sử dụng phương thức confusion_matrix() trong thư viện sklearn và được hiển thị dưới dạng biểu đồ bằng cách sử dụng hàm plot_confusion_matrix().

3.4.Kịch bản thực nghiệm

–Chuẩn bị dữ liệu: Sử dụng một tập dữ liệu chứa các email, văn bản tin nhắn hoặc bất kỳ loại văn bản nào có nhãn spam/ham (hoặc tương đương) bằng tiếng anh

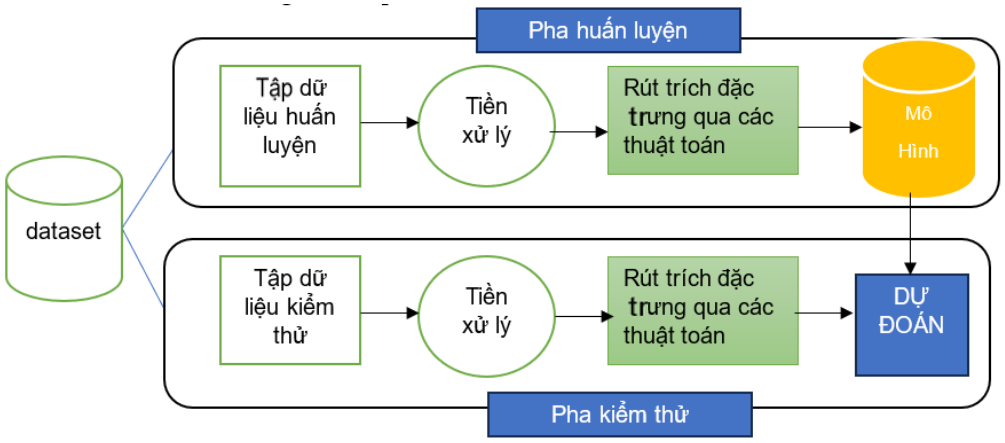
–Tiền xử lý dữ liệu: Thực hiện các bước tiền xử lý dữ liệu

–Huấn luyện mô hình: Sử dụng các mô hình học máy như Decision Tree, Logistic Regression, Random Forest để huấn luyện mô hình trên tập huấn luyện.

–Đánh giá mô hình: Sử dụng các độ đo đánh giá mô hình như precision, recall, F1-score, accuracy, ROC curve, Precision-Recall curve để đánh giá hiệu quả của mô hình trên tập kiểm tra.

–Dự đoán: Sử dụng mô hình đã được huấn luyện để dự đoán nhãn của các Email.

–Đánh giá kết quả: So sánh kết quả dự đoán của mô hình với nhãn thực tế của các Email để đánh giá hiệu quả của mô hình trên dữ liệu thực tế.



Hình 3.4 Mô hình tổng quát

CHƯƠNG 4 KẾT QUẢ NGHIÊN CỨU VÀ THỰC NGHIỆM

4.1 Môi trường cài đặt

Google colab với ngôn ngữ Python

4.2 Các tham số của mô hình

4.2.1 Thuật toán Decision Tree (Gini)

−criterion="gini": sử dụng hàm đo lường sự tinh khiết của node là GINI để phân chia các node.

−random_state=123: đảm bảo kết quả huấn luyện có tính nhất quán, ngẫu nhiên được tạo ra bởi số nguyên này.

−max_depth=10: giới hạn độ sâu của cây quyết định tối đa là 10.

−min_samples_leaf=8: Số lượng mẫu tối thiểu để một node được coi là lá là 8.

4.2.2 Thuật toán Decision Tree (entropy)

−criterion="entropy": sử dụng hàm đo lường sự tinh khiết của node là entropy để phân chia các node.

−random_state=123: đảm bảo kết quả huấn luyện có tính nhất quán, ngẫu nhiên được tạo ra bởi số nguyên này.

−max_depth=10: giới hạn độ sâu của cây quyết định tối đa là 10.

−min_samples_leaf=8: Số lượng mẫu tối thiểu để một node được coi là lá là 8.

4.2.3 Thuật toán Logistic Regression

Random_state=0: đảm bảo kết quả huấn luyện có tính nhất quán, ngẫu nhiên được tạo ra bởi số nguyên này.

4.2.4 Thuật toán Random Forest

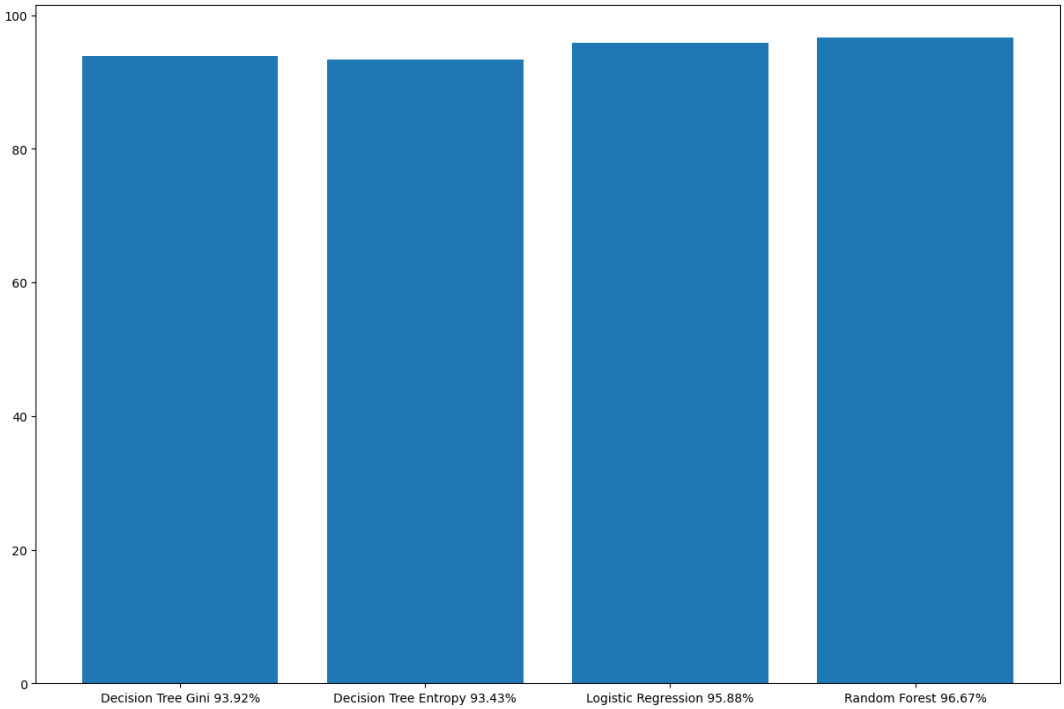
−n_estimators=100: số lượng cây trong rừng là 100.

−max_depth=100: giới hạn độ sâu của mỗi cây trong rừng là 100.

−random_state=0: đảm bảo kết quả huấn luyện có tính nhất quán, ngẫu nhiên.

4.3 Kết quả

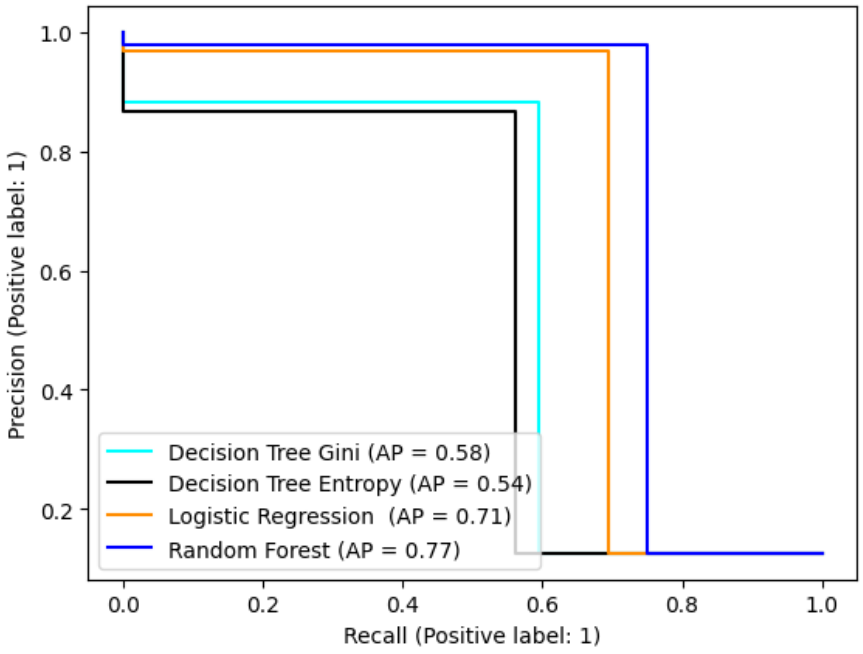
4.3.1 Kết quả nghiên cứu huấn luyện



Hình 4.3.1.1 Kết quả nghiên cứu huấn luyện

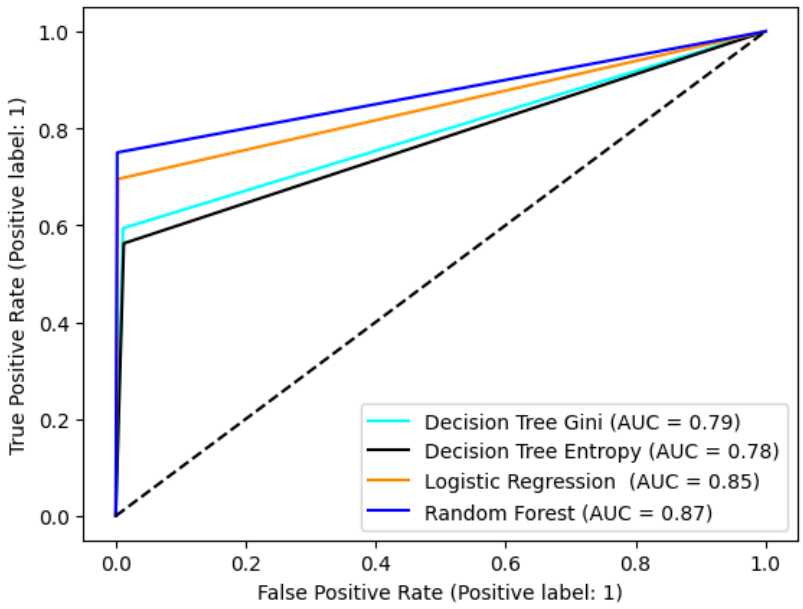
- Thuật toán Decision Tree (Gini) có độ chính xác 93.92%
- Thuật toán Decision Tree (entropy) có độ chính xác 93.43%
- Thuật toán Logistic Regression có độ chính xác 95.88%
- Thuật toán Random Forest 96.67%

Qua đó ta thấy Random Forest có độ chính xác cao nhất trong các thuật toán.



Hình 4.3.1.2 Biểu đồ ROC Precision-Recall

Các giá trị AP cho biết độ chính xác trung bình của bộ phân loại trên tất cả các mức độ recall. Giá trị AP càng cao thì hiệu suất của bộ phân loại càng tốt. Từ kết quả, ta có thể thấy rằng thuật toán Random Forest có giá trị AP cao nhất (0,77), tiếp đến là Logistic Regression (0,71), Decision Tree Gini (0,58) và Decision Tree Entropy (0,54). Do đó, thuật toán Random Forest là thực hiện tốt nhất dựa trên phân tích đường cong Precision-Recall.



Hình 4.3.1.3 Biểu đồ ROC

ROC curve biểu thị mối liên hệ giữa tỷ lệ true positive (TPR) và tỷ lệ false positive (FPR) của một mô hình phân loại, ở các ngưỡng quyết định khác nhau.

Kết quả của mỗi mô hình được đánh giá thông qua AUC (Area Under the Curve) của ROC curve. AUC là một phép đo tổng quát của độ chính xác của mô hình trên tất cả các ngưỡng quyết định có thể. AUC càng gần 1 thì mô hình càng tốt.

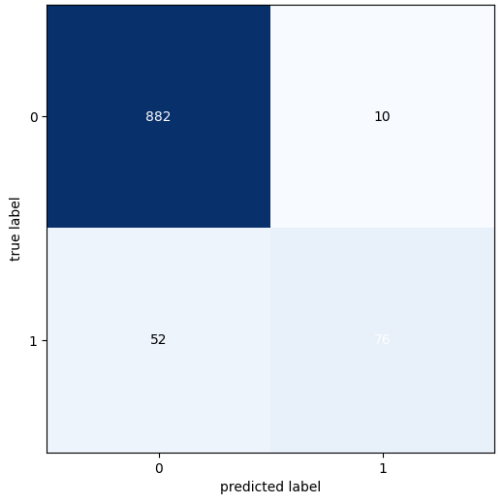
– Decision Tree Gini (AUC = 0.79): Mô hình có AUC đạt 0.79, tức là mô hình có khả năng phân loại tốt hơn so với việc đoán ngẫu nhiên

– Decision Tree Entropy (AUC = 0.78): Mô hình này có AUC hơi thấp hơn so với Decision Tree Gini, nhưng chênh lệch này không quá lớn.

– Logistic Regression (AUC = 0.85): Mô hình này có AUC cao hơn đáng kể so với hai mô hình cây quyết định và tiệm cận tới giá trị 1.0, cho thấy hiệu suất phân loại của mô hình này rất tốt.

– Random Forest (AUC = 0.87): Mô hình này có AUC cao nhất trong số các mô hình được đánh giá, đồng thời cũng gần tiệm cận tới giá trị 1.0. Kết quả này cho thấy mô hình Random Forest có hiệu suất phân loại tốt nhất.

4.3.2 Kết quả kiểm tra thực nghiệm



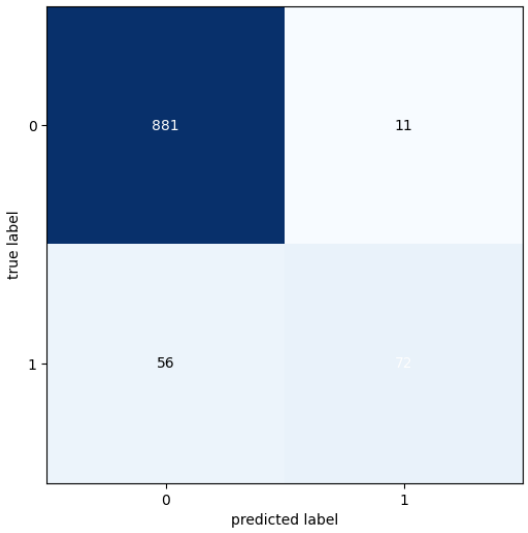
Hình 4.3.2.1 Kết quả thuật toán Decision Tree (Gini)

–Accuracy Score : 93.92 %

–Precision Score : 88.37 %

–Recall Score : 59.38 %

–F1 Score : 71.03 %



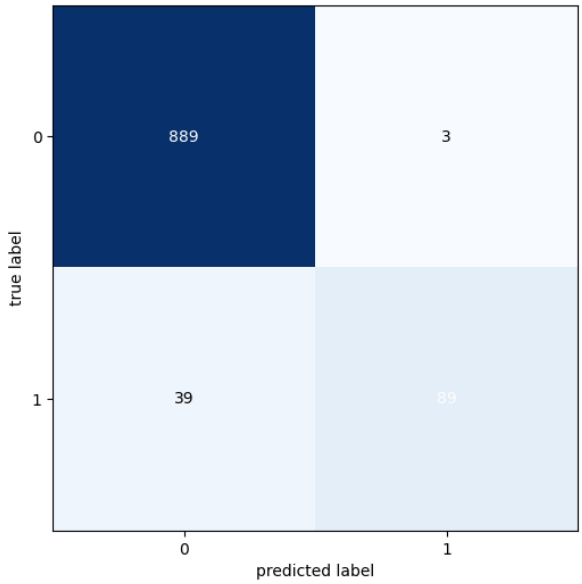
Hình 4.3.2.2 Kết quả thuật toán Decision Tree (entropy)

–Accuracy Score : 93.43 %

–Precision Score : 86.75 %

–Recall Score : 56.25 %

–F1 Score : 68.25 %



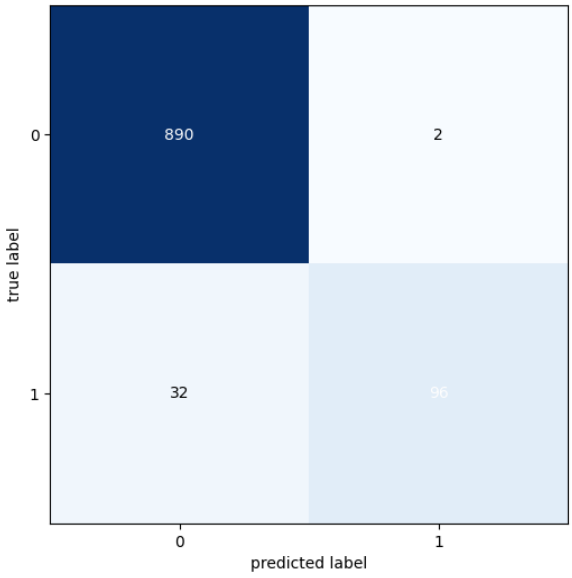
Hình 4.3.2.3 Kết quả thuật toán Logistic Regression

–Accuracy Score : 95.88 %

–Precision Score : 96.74 %

–Recall Score : 69.53 %

–F1 Score : 80.91 %



Hình 4.3.2.4 Kết quả thuật toán Random Forest

–Accuracy Score : 96.67 %

–Precision Score : 97.96 %

–Recall Score : 75.0 %

–F1 Score : 84.96 %

```
# Predict new text
new_text ="I see the letter B on my car"
# Vectorize the new text into numerical form
new_text_vector = vectorization.transform([new_text])
# Model prediction
model_dt_gini_pred = model_dt_gini.predict(new_text_vector)[0]
print('Decision Tree (Gini) model Prediction:')
print('Spam message' if model_dt_gini_pred==1 else 'non-spam message')
model_dt_entropy_pred = model_dt_entropy.predict(new_text_vector)[0]
print('Decision Tree (Entropy) model Prediction:')
print('Spam message' if model_dt_entropy_pred==1 else 'non-spam message')
model_lr_pred = model_lr.predict(new_text_vector)[0]
print('Logistic Regression Model Prediction:')
print('Spam message' if model_lr_pred==1 else 'non-spam message')
model_rf_pred = model_rf.predict(new_text_vector)[0]
print('Random Forest Model Prediction:')
print('Spam message' if model_rf_pred==1 else 'non-spam message')
```

```
Decision Tree (Gini) model Prediction:
non-spam message
Decision Tree (Entropy) model Prediction:
non-spam message
Logistic Regression Model Prediction:
non-spam message
Random Forest Model Prediction:
non-spam message
```

Hình 4.3.2.5 Hình ảnh test kết quả Email bình thường

Hình 4.3.2.6 Hình ảnh test kết quả Spam Email

4.3 Đánh giá

Kết quả cho thấy, thuật toán Random Forest cho kết quả tốt nhất với Accuracy Score là 96.67%, Precision Score là 97.96%, Recall Score là 75.0% và F1 Score là 84.96%. Thuật toán Logistic Regression cũng cho kết quả tốt với Accuracy Score là 95.88%, Precision Score là 96.74%, Recall Score là 69.53% và F1 Score là 80.91%. Các thuật toán Decision Tree (Gini và entropy) cho kết quả thấp hơn so với Random Forest và Logistic Regression.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận:

5.1.1 Những điều đạt được

- Có so sánh đánh giá các thuật toán với nhau
- Áp dụng các phương pháp xử lý dữ liệu vào đề tài
- Xây dựng được mô hình phân loại Email

5.1.2 Những điều chưa đạt được

- Chỉ phân loại được Email tiếng Anh
- Chưa phân loại được email quá dài

5.2 Hướng phát triển:

- Cải tiến các thuật toán để nâng cao tính chính xác
- Hướng đến việc phân loại email với nhiều ngôn ngữ khác nhau
- Hoàn thiện các thiếu xót

TÀI LIỆU THAM KHẢO

[1] Mô hình mới tự động phát hiện và lọc thư rác https://vista.gov.vn/news/cac-linh-vuc-khoa-hoc-va-cong-nghe/mo-hinh-moi-tu-dong-phat-hien-va-loc-thu-rac-4819.html

[2] Tài liệu hướng dẫn thực hành khai phá dữ liệu trường đại học sư phạm Kỹ thuật Vĩnh Long

[3] Khai phá dữ liệu – Tổng quan, ứng dụng và các nền tảng thông dụng để khai phá dữ liệu https://aita.gov.vn/khai-pha-du-lieu-%E2%80%93-tong-quan-ung-dung-va-cac-nen-tang-thong-dung-de-khai-pha-du-lieu