BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NÔNG LÂM THÀNH PHỐ HỒ CHÍ MINH
KHOA MÔI TRƯỜNG VÀ TÀI NGUYÊN

CƠ SỞ VÀ ỨNG DỤNG TRÍ TUỆ NHÂN TẠO

ỨNG DỤNG HỌC MÁY XÂY DỰNG BỘ LỌC THƯ RÁC – EMAIL SPAM FILTER

Nhóm 4:

Họ và tên	**Mã số sinh viên**
Nguyễn Thành Hưng	20166032
Phạm Ngô Phú Khánh	20166035
Võ Tùng Lâm	20166038
Nguyễn Thị Huỳnh Như	20166050
Nguyễn Thị Hồng Nhung	20166051

Thành phố Hồ Chí Minh, ngày 04 tháng 10 năm 2023

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NÔNG LÂM THÀNH PHỐ HỒ CHÍ MINH
KHOA MÔI TRƯỜNG VÀ TÀI NGUYÊN

CƠ SỞ VÀ ỨNG DỤNG TRÍ TUỆ NHÂN TẠO

ỨNG DỤNG HỌC MÁY XÂY DỰNG BỘ LỌC THƯ RÁC – EMAIL SPAM FILTER

Nhóm 4:

Họ và tên	**Mã số sinh viên**
Nguyễn Thành Hưng	20166032
Phạm Ngô Phú Khánh	20166035
Võ Tùng Lâm	20166038
Nguyễn Thị Huỳnh Như	20166050
Nguyễn Thị Hồng Nhung	20166051

Thành phố Hồ Chí Minh, ngày 04 tháng 10 năm 2023

LỜI CAM KẾT

Chúng tôi xin được phép tuyên bố: bài tiểu luận được đưa lên hội đồng, là thành quả của chúng tôi dưới sự chỉ đạo của giáo viên hướng dẫn, tiến hành nghiên cứu đạt được. Ngoại trừ những gì đã được trích dẫn trong bài báo, bài báo này không chứa bất kỳ kết quả nghiên cứu khoa học nào đã được công bố hoặc viết bởi bất kỳ cá nhân hoặc tập thể nào khác. Các cá nhân và tập thể có đóng góp quan trọng cho nghiên cứu này đã được ghi rõ trong văn bản. Trách nhiệm pháp lý của tuyên bố này là của chúng tôi.

Ngày 23 Tháng 12 Năm 2023

LỜI CẢM ƠN

Trong bài tiểu luận về *Ứng dụng học máy xây dựng bộ lọc thư rác – Email Spam Filter*, chúng tôi xin gửi lời cảm ơn chân thành đến giáo viên hướng dẫn của chúng tôi, Dưới sự hướng dẫn tận tình của **TS.Nguyễn Đình Long** , chúng tôi đã hoàn thành bài tiểu luận này. Chúng tôi xin gửi lời cảm ơn đến thầy vì sự giúp đỡ và hỗ trợ của thầy trong việc viết luận văn của chúng tôi.

Chúng tôi cũng xin cảm ơn các tác giả và người cung cấp tài liệu tham khảo trong bài báo đã cung cấp cho chúng tôi những kiến thức quý báu và hỗ trợ thêm để hoàn thành báo cáo này. Những đóng góp của họ đóng một vai trò quan trọng trong việc thúc đẩy nghiên cứu của chúng tôi.

Mặc dù thời gian và kiến thức của chúng tôi khi viết báo còn hạn chế nhưng bản báo cáo không tránh khỏi những thiếu sót. Chúng tôi rất mong các thầy cô và độc giả đưa ra những góp ý, nhận xét quý báu giúp chúng tôi hoàn thiện và hoàn thiện báo cáo này.

Cuối cùng, chúng tôi xin bày tỏ lòng biết ơn chân thành tới tất cả những người đã hỗ trợ, giúp đỡ chúng tôi trong quá trình học tập và nghiên cứu. Nếu không có sự động viên và hỗ trợ của các bạn, chúng tôi sẽ không thể hoàn thành công việc nghiên cứu này.

Một lần nữa bày tỏ lòng biết ơn của chúng tôi!

TÓM TẮT

Bộ lọc thư rác là phương pháp được sử dụng để tự động phát hiện và xóa email không mong muốn, thường được gọi là thư rác, khỏi hộp thư đến của người dùng. Bộ lọc thư rác ngày càng cần thiết khi số lượng email thương mại, lừa đảo hoặc quảng cáo không được yêu cầu ngày càng tăng.

Trong nghiên cứu này, chúng tôi sử dụng thuật toán Naïve Bayes để lọc thư rác trên bộ dữ liệu được thu nhập và kiểm tra hiệu suất của nó, bộ dữ liệu này được phát triển bởi nhà nghiên cứu Venkatesh Garnepudi vào năm 2018 với tên "Spam Email Dataset" với 5171 hàng dữ liệu chứa 1499 email được xem là Spam và 3672 email được xem là không phải thư rác (HAM). Hiệu suất của bộ dữ liệu được chúng tôi đánh giá dựa trên độ chính xác, sai số thống kê.

Mục đích của việc lập trình bộ lọc thư rác là giúp người dùng tiết kiệm thời gian và năng lượng bằng cách tự động lọc ra các email không mong muốn. Tuy nhiên, bộ lọc thư rác không thể hoàn toàn chính xác và có thể gây ra lỗi. Vì vậy, người dùng nên thường xuyên kiểm tra và đảm bảo những tin nhắn quan trọng không bị lẫn lộn hoặc xóa nhầm trong quá trình lọc.

MỤC LỤC

DANH SÁCH HÌNH ẢNH

DANH SÁCH BẢNG BIỂU

CHƯƠNG I MỞ ĐẦU

1.1 Tính cấp thiết của đề tài

Thư rác (hay Spam Email) là một vấn đề cấp thiết và cần được chú trọng trong thời kỳ công nghệp phát triển mạnh như hiện nay. Theo Shafi'I[1], thư rác đã trở thành mối đe dọa nghiêm trọng không chỉ đối với Internet mà còn đối với xã hội. Sự cấp thiết xoay quanh về đề tài nghiên cứu của chúng tôi về việc ứng dụng trí tuệ nhân tạo để xây dựng bộ lọc thư rác sử dụng thuật toán Naive Bayes bắt nguồn từ nhiều yếu tố quan trọng.

Sự tràn ngập không ngừng của thư rác đã trở thành một nỗi phiền toái, làm hỗn loạn hộp thư đến và cản trở các email khác. Điều này không chỉ làm giảm năng suất công việc của cá nhân mà còn mang đến những rủi ro đáng kể về an ninh mạng (Kumar, Sonowal, 2020)[2]. Thư rác thường là cầu nối cho các hoạt động độc hại như lừa đảo, chứa phần mềm độc hại và xâm nhập thông tin cá nhân. Do đó, việc bảo vệ quyền riêng tư và thông tin nhạy cảm của người dùng trở thành vấn đề cấp bách .(Cormack, 2008)[3]

Ngoài ra, sự gia tăng đáng ngại của nội dung độc hại bên trong các tin nhắn rác càng làm tăng sự cần thiết của việc xây dựng bộ lọc thư rác hiệu quả. Những thư rác sử dụng các kỹ thuật ngày càng tinh vi, trong khi các bộ lọc dựa trên quy tắc truyền thống gặp khó khăn khi thích ứng và nhận diện chính xác các thư rác đang được gửi đến hộp thư (Karim, et al. 2019)[4]. Bằng cách tận dụng sức mạnh của trí tuệ nhân tạo và thuật toán Naive Bayes, bài nghiên cứu hướng đến việc phát triển một giải pháp mạnh mẽ và linh hoạt, có khả năng phát hiện và loại bỏ thư rác một cách nhanh chóng, từ đó nâng cao an ninh trực tuyến, tối ưu hóa giao tiếp và tăng cường biện pháp an ninh mạng.

---

[1] Shafi'I, M. A., Abd Latiff, M. S., Chiroma, H., Osho, O., Abdul-Salaam, G., Abubakar, A. I., & Herawan, T. (2017). *A review on mobile SMS spam filtering techniques*. IEEEAccess, 5, 15650-15666.

[2] Kumar, N., & Sonowal, S. (2020, July). *Email spam detection using machine learning algorithms*. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 108-113). IEEE.

[3] Cormack, G. V. (2008). *Email spam filtering: A systematic review*. Foundations and Trends® in Information Retrieval, 1(4), 335-455.

[4] Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., & Alazab, M. (2019). *A comprehensive survey for intelligent spam email detection.* IEEE Access, 7, 168261-168295.

1.2 Lý do chọn đề tài

1.2.1 Lý do tổng quát

Nhu cầu lọc thư rác ngày càng tăng trên toàn thế giới. Theo báo cáo của Radicati Group[5], khoảng 306,4 tỷ email được gửi mỗi ngày vào năm 2020, trong đó khoảng 45% là thư rác. Kết quả là hàng tấn email không cần thiết tràn ngập hộp thư đến của người dùng, lãng phí thời gian và giảm năng suất làm việc.

Bộ lọc thư rác qua email là công cụ quan trọng để giảm thiểu sự xâm nhập của thư rác. Hiện nay, trên thị trường có rất nhiều chương trình phần mềm và dịch vụ cung cấp khả năng lọc thư rác. Tuy nhiên, những cách này không phải lúc nào cũng hiệu quả và có thể dẫn đến những sai sót không mong muốn. Vì vậy, việc nghiên cứu bộ lọc thư rác với tốc độ xử lý và độ chính xác cao sẽ giúp người dùng có thể kiểm soát và quản lý thư rác một cách tốt hơn.

1.2.2 Lý do trọng tâm

Theo Schneider[6], bộ lọc thư rác (Email Spam Filter) có thể xác định và loại bỏ các email không mong muốn bằng cách phân tích nội dung và tiêu đề của email. Các kỹ thuật phổ biến được sử dụng trong bộ lọc thư rác bao gồm sử dụng thuật toán học máy, phân loại dựa trên quy tắc, phân tích từ khóa và kiểm tra danh sách đen. Tuy nhiên, mỗi phương pháp này có những hạn chế và đòi hỏi sự điều chỉnh phù hợp để đảm bảo tính hiệu quả.

Việc xây dựng một bộ lọc thư rác trong email có thể giúp người dùng giảm thiểu sự xâm phạm của thư rác và tăng cường bảo mật thông tin cá nhân. Ngoài ra, việc nghiên cứu và lập trình bộ lọc thư rác còn mang lại những kiến thức, kỹ năng về phân tích dữ liệu, xử lý ngôn ngữ tự nhiên (Natural Language), học máy (Machine Learning).

1.3 Mục tiêu nghiên cứu

1.3.1 Mục tiêu chính

[5] Radicati Group. (2020). *Email statistics report*, 2020-2024: Executive summar

[6] Schneider, K. M. (2003, April). *A comparison of event models for naive bayes anti-spam e-mail filtering*. In 10th Conference of the European Chapter of the Association for Computational Linguistics.

Xây dựng một hệ thống bộ lọc thư rác thông minh và hiệu quả, có khả năng phân loại các email thành hai nhóm: thư rác và thư hợp lệ. Điều này sẽ giúp người dùng tiết kiệm thời gian và tăng cường hiệu suất làm việc bằng cách loại bỏ các thông báo không cần thiết.

1.3.2 Mục tiêu cụ thể

Giảm thiểu các rủi ro an ninh mạng từ thư rác độc hại. Bằng cách phát hiện và cô lập những email chứa nội dung độc hại như thử lừa đảo, virus và phần mềm độc hại, hệ thống bộ lọc sẽ giúp bảo vệ người dùng khỏi các mối đe dọa trực tuyến.

Cuối cùng là tạo ra một bộ lọc thư rác linh hoạt và dễ dàng cập nhật. Với sự phát triển không ngừng của các hình thức thư rác mới, hệ thống cần có khả năng thích ứng và nhận diện các mẫu thư rác tiến triển. Bằng cách sử dụng trí tuệ nhân tạo và thuật toán Naive Bayes, mục tiêu của nghiên cứu là cung cấp một công cụ mạnh mẽ và linh hoạt để ngăn chặn thư rác và nâng cao an ninh thông tin trong giao tiếp trực tuyến.

1.4 Phạm vi nghiên cứu

Phạm vi nghiên cứu của bài tiểu luận này bao gồm các khía cạnh liên quan đến thư rác và việc ứng dụng trí tuệ nhân tạo để xây dựng bộ lọc thư rác sử dụng thuật toán Naive Bayes. Nghiên cứu tập trung vào phân loại và xử lý các email được gửi đến hộp thư đến của người dùng. Dữ liệu nghiên cứu được thu thập từ tập hợp các email, bao gồm cả thư rác và thư hợp lệ, để phát triển và đánh giá hiệu quả của bộ lọc. Phạm vi cũng bao gồm việc xác định và phân loại các đặc điểm của thư rác, ví dụ như từ khóa, cấu trúc, và các thuộc tính khác, để tạo ra một mô hình Naive Bayes đáng tin cậy.

Nghiên cứu cũng có thể xem xét các phương pháp tiền xử lý dữ liệu và kỹ thuật tinh chỉnh tham số để cải thiện hiệu suất của bộ lọc. Mục tiêu là giới hạn phạm vi của nghiên cứu vào việc xây dựng một bộ lọc thư rác hiệu quả sử dụng trí tuệ nhân tạo và thuật toán Naive Bayes, hướng đến việc cải thiện khả năng phân loại và loại bỏ thư rác, đồng thời tối ưu hóa an ninh và hiệu suất giao tiếp cho người dùng.

1.5 Ý nghĩa nghiên cứu

Ý nghĩa nghiên cứu của bài tiểu luận này là xây dựng một bộ lọc thư rác dựa vào việc ứng dụng trí tuệ nhân tạo và sử dụng thuật toán Naive Bayes. Mô hình này mang lại nhiều ý nghĩa quan trọng. Trước hết, việc áp dụng trí tuệ nhân tạo và thuật toán

Naive Bayes giúp cải thiện hiệu quả của bộ lọc thư rác. Nhờ khả năng tự động học và tư duy của trí tuệ nhân tạo, mô hình có khả năng phân loại các email một cách chính xác và nhanh chóng, giúp người dùng tiết kiệm thời gian và tăng cường hiệu suất làm việc.

Thứ hai, kết quả của mô hình cung cấp khả năng loại bỏ thư rác đáng tin cậy. Bằng cách sử dụng thuật toán Naive Bayes và dữ liệu huấn luyện, mô hình có thể xác định các đặc điểm của thư rác và từ đó phân loại chúng một cách chính xác. Điều này giúp giảm thiểu số lượng thư rác trong hộp thư đến và tạo ra một môi trường giao tiếp sạch hơn và hiệu quả hơn.

Cuối cùng, nghiên cứu này có ý nghĩa trong việc cải thiện mô hình so với các mô hình khác. Thuật toán Naive Bayes đã được chứng minh là một phương pháp hiệu quả trong việc phân loại thông tin. Áp dụng thuật toán này vào bộ lọc thư rác giúp tăng cường khả năng phân loại và giảm thiểu sai sót. Ngoài ra, mô hình cũng có khả năng cập nhật linh hoạt để đối phó với các hình thức thư rác mới. Điều này đảm bảo tính hiệu quả và bền vững của mô hình trong việc loại bỏ thư rác.

CHƯƠNG II TỔNG QUAN

2.1 Cơ sở lý thuyết

2.1.1 Tổng quan về thư rác

Có rất nhiều định nghĩa khác nhau cho thư rác. Theo (Rao, Reiley, 2012)[7], thư rác (spam) là hoạt động gửi thư điện tử không được yêu cầu, với số lượng lớn và trong một số trường hợp liên tục gửi đến những cá nhân mà không có liên hệ trước đó và địa chỉ e-mail của họ được thu thập không hợp pháp, hoặc spam thường được hiểu là việc gửi đi hàng loạt các thư thương mại không được yêu cầu bởi một người gửi ngụy trang hoặc giả mạo danh tính, hay được định nghĩa là tin nhắn điện tử không được yêu cầu, bất kể nội dung (Rao, Reiley, 2012)[7] (Granacher , et al. 2020)[8]. Định nghĩa này có tính đến các đặc điểm của thư điện tử hàng loạt (Granacher , et al. 2020)[8].

Các định nghĩa về thư rác đều có những đặc điểm chung như sau:

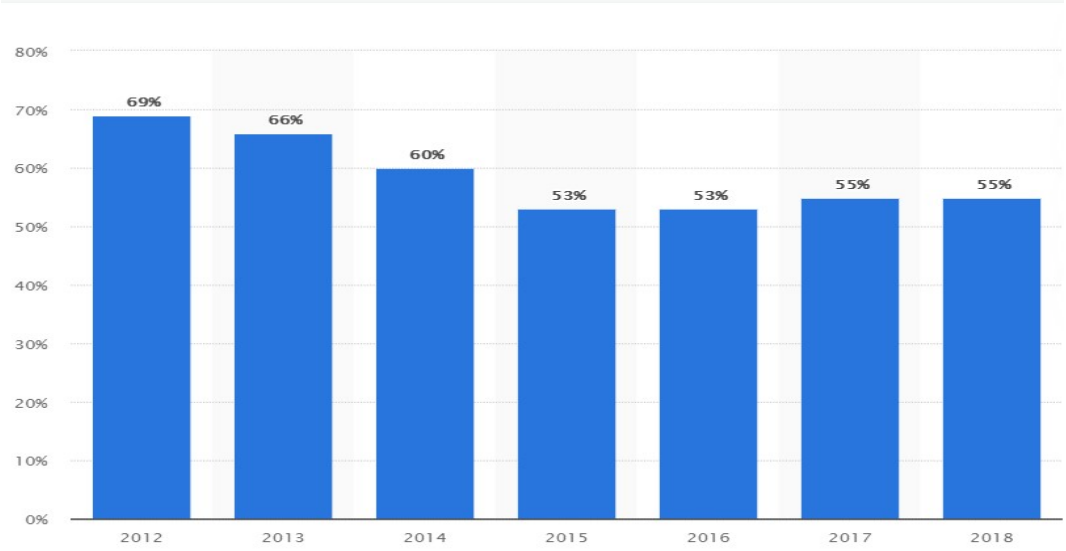
1. Gửi bằng thư điện tử
2. Sử dụng các địa chỉ được thu thập mà không có sự đồng ý
3. Gửi hàng loạt
4. Không mong muốn nhận
5. Lặp đi lặp lại
6. Nhằm mục đích thương mại hoặc tài chính
7. Không có mục tiêu và bừa bãi
8. Không thể ngăn cản
9. Ẩn danh và / hoặc ngụy trang
10. Nội dung bất hợp pháp hoặc xúc phạm
11. Nội dung lừa đảo hoặc gian lận

[7] J., M., Rao, & D.,H., Reiley, "The economics of spam.," Journal of Economic Perspectives, vol. 26, no. 3, pp. 87-110, 2012.

[8] A., Granacher, D. Harz, J., Kader, J., Noll, & M., Usher, Unsolicited bulk email A., Granacher, D. Harz, J., Kader, J., Noll, & M., Usher, Unsolicited bulk email detection using URL tree hashes, Washington, DC: U.S.: Patent and Trademark Office, 2020.

2.1.2. Thống kê và tác hại của thư rác

Thư rác chiếm một phần rất lớn của băng thông mạng, theo thống kê của (Statista, 2023)[9], từ những năm 2012 đến 2018, số lượng thư rác luôn nhiều hơn so với số lượng thư hợp lệ (ham). Mặc dù số lượng thư rác có giảm xuống so với những năm trước đây từ 69% giảm xuống 55%, tuy nhiên, có thể thấy thư rác vẫn chiếm một lượng băng thông mạng lớn.



Bảng 1 Tỉ lệ thư rác điện tử từ năm 2012 đến 2018 (Statista, 2023)[9]

Theo thống kê mới nhất của (Clement, 2020)[10], thư rác chiếm 53,95 phần trăm lưu lượng thư điện tử vào tháng 3 năm 2020. Trong khoảng thời gian gần đây nhất, Nga chiếm phần lớn nhất trong số các thư rác với 20,74 % tổng lượng thư rác toàn cầu. Bất chấp sự phổ biến của nó, tỷ lệ thư rác e-mail toàn cầu thực sựđang giảm: tỷ lệ thư rác hàng năm toàn cầu trong năm 2018 là 55%, giảm so với 69% vào năm 2012 (Clement , 2020)[11].

[9] Statista, [Online]. Available: https://www.statista.com/statistics/270899/global-email-spam-rate/. [Truy cập 24/12/2023].

[10] J. Clement, "Global spam categories 2019," 14 5 2020. [Online]. Available: https://www.statista.com/statistics/263452/most-common-content-of-spam-messages-worldwide-by-category/. [Truy cập 24/12/2023].

[11] J. Clement, "Global spam volume as percentage of total e-mail traffic from January 2014 to March 2020, by month," 24 June 2020. [Online]. Available: https://www.statista.com/statistics/420391/spam-email-traffic-share/. [Truy cập 24/12/2023]

Thư rác là thư điện tử được gửi không mong muốn, gây ra nhiều ảnh hưởng đến nhiều mặt khác nhau:

1. Quá tải băng thông: Thư rác chặn các kênh liên lạc và tạo ra lưu lượng băng thông lớn (chi phí băng thông này công ty/doanh nghiệp phải trả tiền). Ngoài ra, có những máy chủ thư phải xử lý thư rác và những máy chủ này phải được bảo trì bởi các chuyên gia được trả lương cao. Do đó, chi phí vận hành cơ sở hạ tầng tăng đáng kể.
2. Tốn thời gian. Nếu thư rác đến hộp thư đến của người dùng, người nhận phải xóa nó theo cách thủ công. Một người đọc 10-20 thư điện tử mỗi ngày có thể nhận được khoảng 160 -180 tin nhắn rác cùng với thư từ công việc của họ. Điều đó có nghĩa là họ sẽ dành 5-6 giờ mỗi tháng chỉ để xóa thư rác, gây phương hại đến thời gian làm việc hiệu quả của họ.
3. Khó chịu và phiền toái. Bằng cách xóa thủ công thư rác, người dùng trở thành một kỹ thuật viên xử lý rác thải 'điện tử'. Việc buộc phải thực hiện các biện pháp như vậy không thể không gây khó chịu cho người dùng, dẫn đến những cảm xúc tiêu cực không mong muốn. Cũng có thể trong quá trình xóa, người dùng cũng có thể vô tình xóa mất một thư điện tử quan trọng cùng với vô số thư rác. Tất cả những ai đã đối mặt với tình huống như vậy đều cảm thấy khó chịu và phiền toái.

Ảnh hưởng của thư rác đối với cá nhân:

Thư rác gây tốn kém chi phí cho người dùng và cho xã hội nói chung. Cho dù tài khoản doanh nghiệp hay cá nhân, người nhận thường sẽ tốn thời gian để phân loại thư điện tử và xóa các thư không mong muốn và do đó phải chịu chi phí cơ hội về thời gian. Thư rác cũng gia tăng chi phí của các nhà cung cấp dịch vụ Internet (ISP) do tiêu thụ nhiều băng thông. Cuối cùng, thư rác được sử dụng để đạt được các mục tiêu gian lận hoặc tội phạm khác, gây ra các chi phí gián tiếp tiềm ẩn có liên quan đến nó.

Theo ước tính (Rao, Reiley, 2012)[7] rằng các công ty và người tiêu dùng Mỹ phải chịu chi phí gần 20 tỷ đô la mỗi năm do thư rác. Con số của người tiêu dùng phải gánh chịu chi phí gần 20 tỷ đô la hàng năm do thư rác. Ước tính rằng những người gửi thư rác và người bán quảng cáo bằng thư rác thu thập tổng doanh thu trên toàn thế giới

theo đơn đặt hàng là 200 triệu đô la mỗi năm. Do đó, "tỷ lệ" của chi phí bên ngoài cho lợi ích bên trong cho thư rác là khoảng 100:1. (Rao, Reiley, 2012)[7]

Ảnh hưởng của thư rác đối với doanh nghiệp:

Trong môi trường kinh doanh, thư rác gây ra tốn kém cho chi phí bảo mật dịch vụ, phần cứng và phần mềm; chi phí huấn luyện; mất năng suất do mất thời gian xóa các thư điện tử không mong muốn (hoặc tìm kiếm những cái đã xóa) và chi phí mua dung lượng lưu trữ bổ sung. Theo ước tính (Silva, 2022)[12] của về chi phí spam từ các nhà xuất bản và tạp chí lên tới 1,1 tỷ đô la Mỹ mỗi năm. Tính tất cả các loại spam, chi phí tăng lên khoảng 2,6 tỷ đô la Mỹ mỗi năm.

Ngoài ra, các doanh nghiệp cũng bị từ các tác động gián tiếp của thư rác, chẳng hạn như phải trả giá cao hơn cho các dịch vụ ISP. Ở cấp độ nhà cung cấp dịch vụ (ISP), chi phí xử lý thư rác là một phần của ngân sách bảo mật. Cách đây vài năm, ISP coi vẫn coi thư rác là một vấn đề của người dùng cá nhân (Schneider, 2003)[6]. Tuy nhiên, với một lượng lớn thư rác gia tăng, các ISP phải đối mặt với các khoản đầu tư có thể tốn kém vào cơ sở hạ tầng thư và đầu tư mua thêm thiết bị lưu trữ, khiến cho các nhà ISP ngày càng quan tâm vấn đề này và giúp làm sáng tỏ những chi phí tiềm ẩn này.

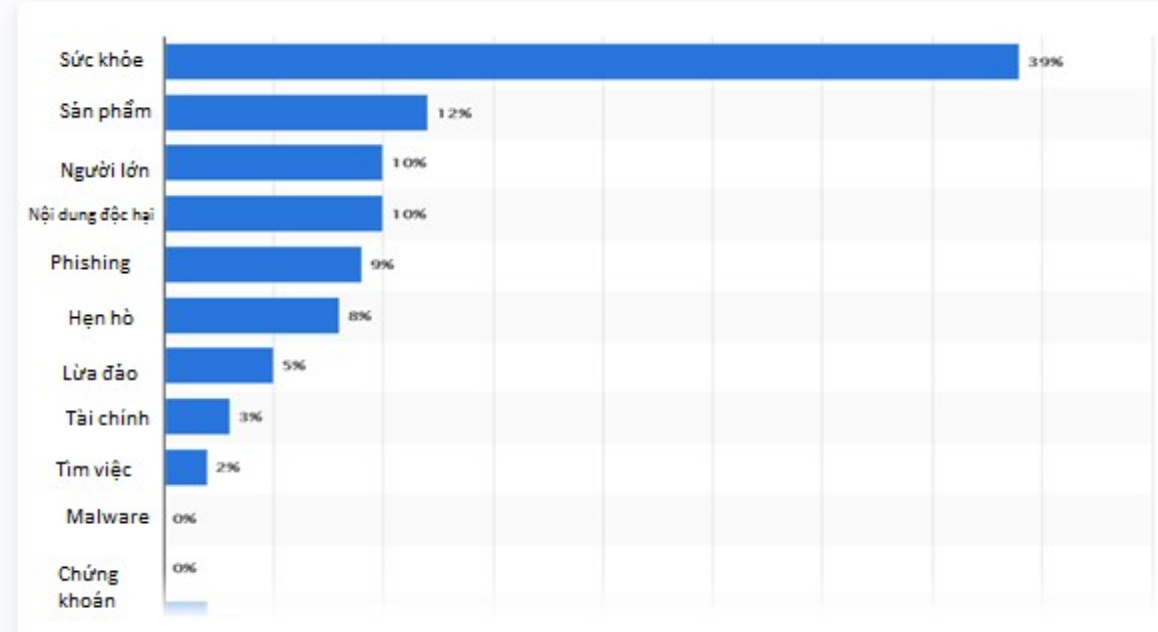
2.1.3 Các loại thư rác

Thư rác có thể được phân loại theo mục tiêu của người gửi thư rác. Nhiều người gửi thư rác gửi e-mail hàng loạt vì lý do quảng cáo như gửi quảng cáo thương mại hoặc mời tham gia vào các chiến dịch chính trị, hoặc nhằm mục đích lừa đảo hay phân phối phần mềm độc hại như virus hay trojan. Phần này trình bày các loại thư rác phổ biến và đưa ra số liệu thống kê, nếu có.

Theo (Statista, 2023)[9], trong năm 2018, 281,1 tỷ thư điện tử được gửi và nhận hàng ngày. Điều này bao gồm hàng tỷ thư quảng cáo được gửi bởi các nhà tiếp thị mỗi ngày. Trong khi nhiều người dùng e-mail tin rằng nội dung như sẽ nằm trong mục thư rác (Spam), hay thư tiếp thị nói chung là vô hại, hoặc chỉ là gây khó chịu cho người dùng. Tính đến quý 2 năm 2018, chỉ 85% thư điện tử tiếp thị nằm trong Hộp thư đến

[12] Silva, J. A. T.d.,A. Al-Khatib, & P.Tsigaris, " Spam emails in academia: issues and costs.," Scientometrics, , vol. 122, no. 2, pp. 1171-1188, 2020.

(Inbox) của người dùng và 7% đã bị bộ lọc thư rác bắt được. Tuy nhiên, mọi thứ đang được cải thiện vào năm 2018, tỷ lệ đặt thư rác thương mại đã giảm xuống còn 9%, giảm từ 14% vào năm 2017.



Bảng 2. Thống kê phân loại các nội dung của thư rác năm 2020 (Clement , 2020)[11].

Theo thống kê (Clement, 2020)[10], các danh mục nội dung thư rác năm 2019, đứng đầu là nội dung về sức khỏe (39%), sau đó là quảng cáo các sản phẩm (12%), và các nội dung khác (người lớn, độc hại, lừa đảo, hẹn hò,…) chiếm từ 2% đến 10%. Nhìn chung, các thư rác được phân thành các loại phổ biến như sau:

2.1.3.1 Thư rác quảng cáo

Thư rác quảng cáo hay còn được gọi là thư rác mục đích thương mại (viết tắt là UCE). Hầu hết, các thư rác quảng cáo được coi là một loại hình tiếp thị trực tiếp và được các công ty coi là một công cụ quan trọng để tiếp cận khách hàng (tiềm năng), vì thư điện tử chi phí rẻ (hầu như miễn phí) và dễ dàng để liên hệ với một nhóm lớn khách hàng. Tuy nhiên, hầu hết thư rác không được gửi bởi chính các công ty quảng cáo, mà bởi những người gửi thư rác (spammer), là những người nhận hoa hồng từ các công ty này (Rao, Reiley, 2012)[7] một nghiên cứu ước tính rằng chi phí gửi một thư điện tử là từ 0,01 đô la Mỹ đến 0,05 đô la Mỹ (Rao, Reiley, 2012)[7]. Vì chi phí gửi thư rác rất thấp nên những người gửi thư rác có thể kiếm được lợi nhuận mặc dù tỷ lệ phản hồi cực kỳ thấp. Theo (Rao, Reiley, 2012)[7] chỉ ra mặc dù chi phí thấp, tỷ lệ phản hồi thấp, tuy nhiên khoảng 8% người trả lời thừa nhận họ đã thực sự mua một sản phẩm

quảng cáo qua thư rác. Miễn là những kẻ gửi thư rác có thể kiếm được nhiều tiền hơn hơn chi phí của họ, họ có thể sẽ tiếp tục gửi thư rác. Đây là hành vi "hợp lý" theo nghĩa kinh tế.

Ngoài các thư rác nhằm mục đích quảng cáo trực tiếp còn có các thư rác gián tiếp sẽ khuyến nghị mua một cổ phiếu cụ thể, nhằm tác động đến giá cổ phiếu nào đó. Một nghiên cứu thực nghiệm (Ohme, Holz, 2006).[13] cho thấy trong ngắn hạn, cổ phiếu thư rác có tác động đáng kể đến cả khối lượng giao dịch và định giá thị trường. Theo thống kê của Symantec, 80% thư rác về lĩnh vực tài chính, sức khỏe, Internet và các sản phẩm dành cho người lớn. Theo một thống kê khác thì thư rác về quảng cáo các loại thuốc chức năng chiếm ưu thế. Hai số liệu thống kê có sự khác biệt rõ ràng, điều này phụ thuộc vào dữ liệu mẫu của hai thống kê thu thập được. Tuy nhiên, không thể phủ nhận rằng thư rác quảng cáo chiếm một thị phần rất lớn của thư rác.

2.1.3.2. Thư rác tuyên truyền, vận động

Các thư quảng cáo không cần thiết phải có mục đích thương mại. Họ cũng có thể tuyên truyền các ý tưởng và/hoặc tổ chức chính trị, văn hóa hoặc tôn giáo. Ví dụ, vào năm 2003, các thành viên của Quốc hội Hoa Kỳ đã gửi hàng trăm nghìn thư không được yêu cầu cho thành viên nhằm vận động các chiến dịch. Các loại thư điện tử này thường nhằm mục đích kêu gọi sự ủng hộ, đồng thuận của những người nhận thư.

2.1.3.3. Thư rác lừa đảo

Lừa đảo qua thư điện tử là các tin nhắn lừa đảo cung cấp số tiền lớn và yêu cầu chi tiết tài khoản ngân hàng hoặc lừa đảo giả mạo các dịch vụ phổ biến và lừa người nhận cung cấp chi tiết thẻ tín dụng /tài khoản của họ [8]. Một trong những trò lừa đảo rửa tiền phổ biến là người dùng nhận được thư điện tử từ một người tự xưng là quan chức chính phủ, thành viên gia đình của một quan chức đã qua đời hoặc luật sư đại diện cho một khách hàng giàu có đã qua đời. Thư điện tử hỏi chi tiết ngân hàng hoặc yêu cầu người nhận thanh toán trước như một cử chỉ thiện chí, với lời hứa sẽ hoàn lại tiền trong tương lai. Nếu người nhận cung cấp thông tin chi tiết, thì tài khoản của

[13] R. B¨ohme and T. Holz, "The Effect of Stock Spam on Financial Markets," in The Workshop on the Economics of Information Security (WEIS) 2006, University of Cambridge, 2006.

khách hàng sẽ bị tấn công và trừ tiền. Ở mỗi quốc gia khác nhau lại có nhiều phiên bản lừa đảo khác nhau tồn tại.

Một dạng lừa đảo qua thư rác nữa là một doanh nghiệp hợp pháp được thành lập, nhằm lừa người dùng cung cấp thông tin cá nhân, thông tin này sẽ được sử dụng để đánh cắp danh tính. Thư gửi đến sẽ hướng dẫn người dùng truy cập một trang web nơi họ được yêu cầu cập nhật thông tin cá nhân, chẳng hạn như mật khẩu và thẻ tín dụng, số an sinh xã hội và số tài khoản ngân hàng mà tổ chức hợp pháp đã có. Tuy nhiên, trang web này là giả mạo và được thiết lập chỉ để lấy cắp thông tin của người dùng. Sự nguy hiểm của các trò gian lận lừa đảo là trang web mà nạn nhân hướng đến thường giống thật, vì đó là trang web giả mạo nhằm sao chép trang web của doanh nghiệp hợp pháp. Lừa đảo qua thư điện tử này được gọi là phishing, là một biến thể của "câu cá", ý tưởng là mồi được ném ra với hy vọng rằng trong khi hầu hết sẽ bỏ qua mồi, một số sẽ bị dụ cắn [8].

2.1.3.4. Thư rác chứa mã độc

Tuy nhiên, không phải tất cả các thư rác đều là thư quảng cáo lành tính. Một phần đáng kể các tin nhắn rác có tính chất độc hại hơn, nhằm mục đích phá hoại hoặc chiếm đoạt hệ thống của người dùng. Các biến thể phổ biến nhất của thư rác độc hại trên toàn thế giới bao gồm vi rút, trojan, phần mềm gián điệp và phần mềm tống tiền [8]. Virus là một chương trình, giống như virus sinh học, có thể sao chép và đôi khi làm hỏng máy tính bị nhiễm. Bằng phương thức này, vi rút là một chương trình hoặc tài liệu được đính kèm với một thư điện tử mà khi mở ra, nó sẽ lây lan bằng cách tự chuyển tiếp hàng loạt người nhận trong danh bạ của người gửi hoặc người dùng sẽ tải xuống và cài đặt giúp kẻ tấn công chiếm quyền điều khiển hệ thống. Đối với thư điện tử từ người lạ, không nên mở bất kỳ tệp đính kèm nào nếu không chắc chắn là nó không gây hại. Thư rác phát tán phần mềm độc hại để lây nhiễm sang máy chủ nhằm chiếm quyền điều khiển từ xa và được sử dụng để gửi nhiều thư rác hơn. Các máy chủ bị nhiễm được gọi là "zombie". Nhiều người tin rằng hầu hết thư rác được gửi qua mạng botnet, là một mạng lưới các máy tính cá nhân bị lây nhiễm mã độc, tuy nhiên giả thuyết này cũng khó được chứng minh.

2.1.3.5. Thư rác bôi nhọ

Thư rác bôi nhọ - "Joe job" là thuật ngữ Internet để chỉ địa chỉ thư giả mạo, nhìn có vẻ đúng là địa chỉ thư của ai đó, nhưng thực sự đó là địa chỉ thư được giả mạo bởi một người khác, với ý định tạo ra các phiền toái, bôi nhọ hoặc làm tổn hại đến danh tiếng của nạn nhân vô tội. Ví dụ, kẻ xấu có thể gửi một thư rác chứa nội dung khiêu dâm trẻ em cho hàng nghìn người sử dụng địa chỉ trả lại giả mạo để khiến người nhận phẫn nộ và kích động. Tên "joe job" lần đầu tiên được sử dụng để mô tả một kế hoạch hướng đến Joe Doll, người đã cung cấp dịch vụ lưu trữ cho các trang web miễn phí. Một người dùng có tài khoản bị xóa vì quảng cáo thông qua thư rác; để trả đũa, anh ta đã gửi một thư rác khác cho vài triệu nạn nhân vô tội, nhưng với tiêu đề "trả lời" được giả mạo Joe Doll.

2.1.4 Tổng quan về Navie Bayes

Naive Bayes là một phương pháp thống kê cơ bản dựa trên xác suất được đề xuất bởi Sahami cùng cộng sự [14][14]. Thuật toán dự đoán và phân loại thư mới là thư rác hay hợp pháp bằng cách sử dụng "training set"- là một tập dữ liệu mẫu và được "huấn luyện". Từ đó, thuật toán sẽ sử dụng thông tin từ dữ liệu mẫu để đưa ra thống kê và so sánh, đối chiếu các đặc điểm của thư mới với các bộ mẫu [8].

Bộ lọc thống kê dựa trên "Định lý xác suất Naïve Bayes" được coi là rất hữu ích trong việc phát hiện thư rác từ những năm 1998-2000 [8]. Hiện nay, nó vẫn rất phổ biến và được triển khai rộng rãi và sử dụng phương pháp thống kê toán học và vectơ đặc trưng– ví dụ như là cụm từ *"ưu đãi"* Công thức Naïve Bayes là:

$$\Pr(S|W) = \frac{\Pr(S|W).\Pr(S)}{\Pr(W|S).\Pr(S) + \Pr(W|H).\Pr(H)}$$

Với:

P(S) là xác suất xuất hiện của tổng số thư rác

P(H) là xác suất xuất hiện của tổng số thư hợp lệ

P(W|H) là xác suất từ, cụm từ đó xuất hiện trong thư hợp lệ

P(W|S) là xác suất từ, cụm từ đó xuất hiện trong các thư rác Với P(S|W) là xác suất thư có chứa từ *"ưu đãi"* là thư rác

[14] A.T. Sabri, A.H. Mohammads, B. Al-Shargabi & M.A Hamdeh, "Developing new continuous learning approach for spam detection using artificial neural network (CLA_ANN).," European Journal of Scientific Research, vol. 42, no. 3, pp. 525-535, 2010.

Cho ví dụ cụ thể hơn như sau:

- Số lượng thư thư rác là 5000. Trong đó 600 thư điện tử chứa từ *"ưu đãi"*
- Số lượng thư hợp lệ là 500, trong đó 9 thư có từ *"ưu đãi"*

Như vậy xác suất một thư chứa từ *"ưu đãi"* có khả năng là spam được tính như sau:

$$P(S|W) = \frac{(600/5000) \times (5000/5500)}{\left(\frac{600}{5000}\right) \times \left(\frac{5000}{5500}\right) + \left(\frac{9}{500}\right) \times \left(\frac{500}{5500}\right)} \approx 88\%$$

Trên thực tế, để xem xét một thư có phải là thư rác hay không, ta phải xem xét rất nhiều từ ngữ, dẫn đến việc tính toán xác suất khả năng của một thư có phải là thư rác hay không như sau:

$$p = \frac{p_1 p_2 \dots P_N}{p_1 p_2 \dots P_N + (1-p_1)(1-p_2) \dots (1-p_N)}$$

Với:

p là xác suất là thư rác của một thư điện tử đang kiểm tra.

p_1 là xác suất $P(S|W_1)$ thư rác có chứa từ thứ nhất p_2 là xác suất $P(S|W_2)$ thư rác có chứa từ thứ hai p_n là xác suất $P(S|W_2)$ của thư rác có chứa từ thứ n.

Thuật toán Naïve Bayes cho phân loại thư rác như sau:

Thuật toán 1. Naïve Bayes

---

1. Nhập tập thư điện tử
2. Trích xuất văn bản thành các token thành phần w_i
3. Xác suất spam cho mỗi token là $P(S/w_i)$, xác suất ham cho mỗi token là $P(H|w_i)$
4. Lưu giá trị xác suất spam cho mỗi token này
5. **for each (M) do** *//chạy vòng lặp duyệt thư từ đầu đến cuối*
6. **while (M not end) do** *//chạy vòng lặp cho đến cuối thư*
7. Chạy thư cho đến token kế tiếp w_i
8. Tìm trong database xác suất spam của $P(S|w_i)$
9. Tính xác suất spam của thư M là P(S|M) và xác suất ham của thư P(H|M)
10. Tính toán khả năng xác suất thư rác của toàn thư điện tử P(M)

11. **if** P(M) > ngưỡng
12. thư được đánh dấu là spam
13. **else if**
14. thư được đánh dấu là ham
15. **end if**
16. **end if**
17. **end while**
18: **end for**
19. **return** kết quả phân loại thư rác.
20. Kết thúc chương trình

Thuật toán trên trải qua những bước sau:

Giai đoạn 1. Huấn luyện

Đầu vào: Thư cần kiểm tra là thư rác hay không.

Phân tích cú pháp mỗi email thành các từ, cụm từ hay còn gọi là token. Tính toán xác suất cho các token này.

Giai đoạn 2. Lọc

Đối với mỗi thư M, duyệt tuần tự từ đầu cho đến cuối thư và tính toán xác suất là spam của từng token w_i, sau đó tính toán xác suất ham của từng token.

Sau đó tính tổng giá xác suất của thư, xác tổng xác suất P(M) lớn hơn ngưỡng, thư sẽ bị đánh dấu là thư rác khác, nếu không sẽ được đánh dấu là thư hợp lệ.

Đầu ra: Trả lại kết quả phân loại là thư rác hay thư hợp lệ.

2.2 Thông tin bối cảnh

2.2.1 Tại Việt Nam

(Quan Dang Dinh và alt, 2014)[15] đã thực hiện đề tài "Automated generation of ham rules for Vietnamese spam filtering" trong bài báo này tác giả đề xuất một phương pháp tạo quy tắc SpamAssassin có thể chỉ ra mức độ quan trọng của một email. Cụ thể là tác giả đã thêm một tỷ lệ quy tắc ham có trọng số âm và HPSOWM

[15] Quan Dang Dinh và alt. (2014, 10). Automated generation of ham rules for Vietnamese spam filtering. doi:10.1109/CISDA.2014.7035628

được điều chỉnh, một thuật toán tiến hóa hiệu quả, để tối ưu hóa điểm quy tắc SpamAssassin. Kết quả bằng cách sử dụng bộ quy tắc mới SpamAssassin có thể đưa ra điểm số mang tính biểu thị cho cả thư rác và nội dung vi phạm. những điểm số này có thể được ứng dụng email khách sử dụng để phân loại thư đến dựa trên tầm quan trọng của tác giả với người dùng.

(vu hoang và alt, 2007)[16] đã thực hiện đề tài "Topic-Based Vietnamese News Document Filtering in the BioCaster Project" Trong bài viết này, tác giả mô tả một hệ thống lọc tài liệu tiếng Việt (VTDF) dựa trên chủ đề trong Dự án BioCaster tự động phân loại các tài liệu tin tức từ nhiều nguồn khác nhau thành các chủ đề có liên quan phù hợp để phát hiện dịch bệnh. Với số lượng rất lớn các báo cáo tin tức phải được phân tích mỗi ngày, VTDF là một bước tiền xử lý quan trọng trong việc giảm gánh nặng của chú thích ngữ nghĩa. Ở đây tác giả trình bày hai cách tiếp cận khác nhau cho vấn đề phân loại tài liệu Việt Nam sẽ được sử dụng trong hệ thống VTDF. Bằng cách sử dụng các phương pháp tiếp cận Bag OfWords - BOW và Statistical N-Gram LanguageModeling - N-Gram, có thể đánh giá hai phương pháp phân loại được sử dụng rộng rãi này cho nhiệm vụ của mình và cho thấy N-Gram có thể đạt được độ chính xác trung bình 95% với thời gian lọc trung bình 79 phút trong khoảng 14.000 tài liệu (3 tài liệu / giây).

2.2.2 Thế giới

Rusland và nhóm nghiên cứu [17][17] đã thực hiện sử dụng thuật toán Naïve Bayes để lọc thư rác trên hai tập dữ liệu để đánh giá độ chính xác, độ thu hồi, độ chính xác và độ F. Thuật toán Naïve Bayes là một bộ phân loại dựa trên xác suất và xác suất được tính theo mật độ xuất hiện và sự kết hợp của các giá trị trong tập dữ liệu. Nghiên cứu này được thực hiện thông qua ba giai đoạn như tiền xử lý, trích xuất tính năng và lọc Naïve Bayes. Đầu tiên họ xóa tất cả các từ liên kết, các từ nối để lấy những từ khóa chính. Sau đó họ tạo cho hai tập dữ liệu thông qua công cụ WEKA; là Spamdata và SpamBase. Sự khác

[16] vu hoang và alt. (2007, 01). Topic-Based Vietnamese News Document Filtering in the BioCaster Project. doi:10.1109/ALPIT.2007.56

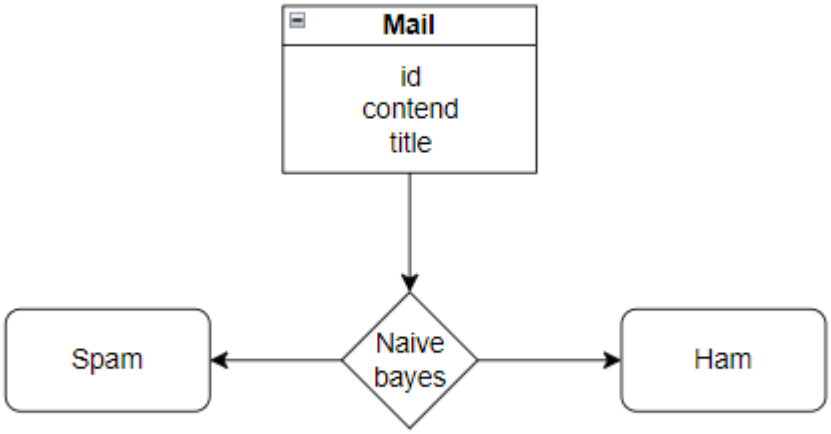
[17] N.F. Rusland, N. Wahid, S., Kasim, S., & H. Hafit, "Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets," IOP Conference Series: Materials Science and Engineering, vol. 226, no. 1, 2017

biệt trung bình là 8,59% trong đó độ chính xác khi lọc Spamdata chiếm 91,13% và khi lọc SpamBase có độ chính xác 82,54%. Trung bình độ chính xác đối với SpamBase là 88% và đối với Spam Data là 83%. Họ đề xuất rằng, bộ lọc Naïve Bayes hoạt động tốt hơn trên dữ liệu SpamBase hơn là dữ liệu Spamdata.

CHƯƠNG III PHƯƠNG PHÁP LUẬN

3.1 Hướng tiếp cận nghiên cứu

Bộ lọc thư rác hoạt động bằng cách phân tích thư trước khi được đưa vào hộp thư đến của người dùng nhằm kiểm tra xem có phải thư rác hay không. Bộ lọc này phân tích nội dung, địa chỉ gửi thư, đầu thư, các tệp đính kèm, ngôn ngữ và các dấu hiệu khả nghi khác



Hình 1 Mô tả hướng tiếp cận

3.2 Thu thập dữ liệu

3.2.1 Đánh giá dữ liệu nghiên cứu

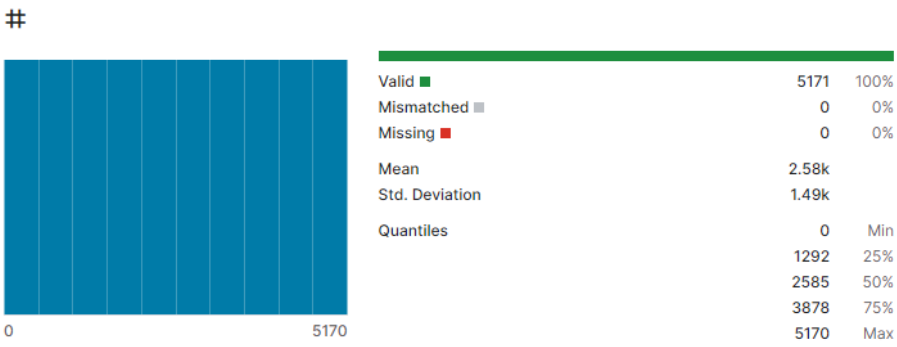
Bộ dữ liệu là bộ dữ liệu thứ cấp: Spam Mails Dataset (Garnepudi, 2018)

Mô tả dữ liệu

- spam_ham_dataset.csv gồm 4 cột và 5170 hàng.
    - Cột 1 trống tên trường: là cột gán giá trị ID cho một mail.
    - label: là nhãn email có thể là Spam hoặc Ham.
    - text: chứa nội dung của mail.
    - label_num: nếu thư rác thì là 1, nếu không thì là 0.

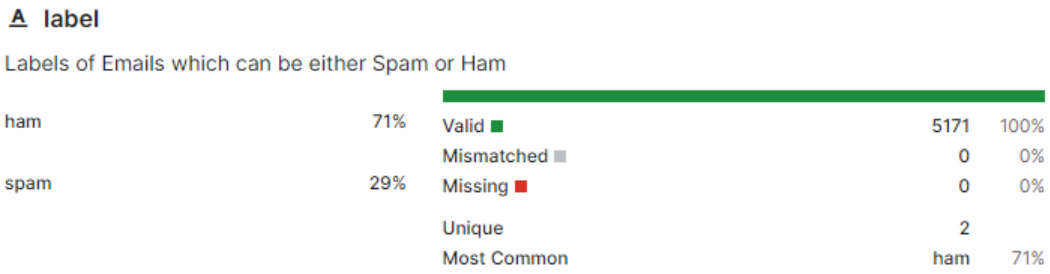
3.2.2 Thống kê dữ liệu

- Cột 1 mang giá trị từ 0 đến 5170.



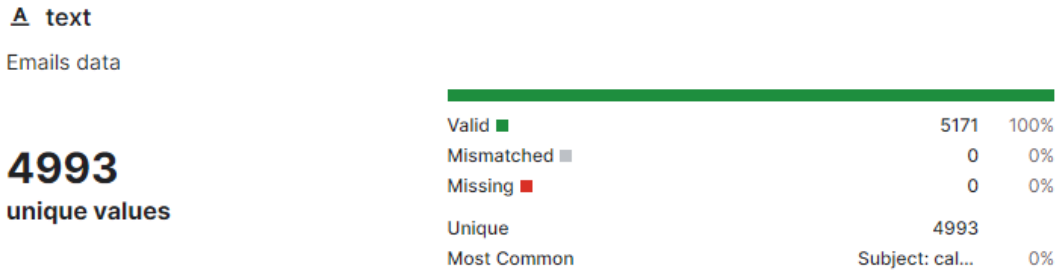
Hình 2 Thống kê cột 1

- **label**: có 71% ham và 29% spam.



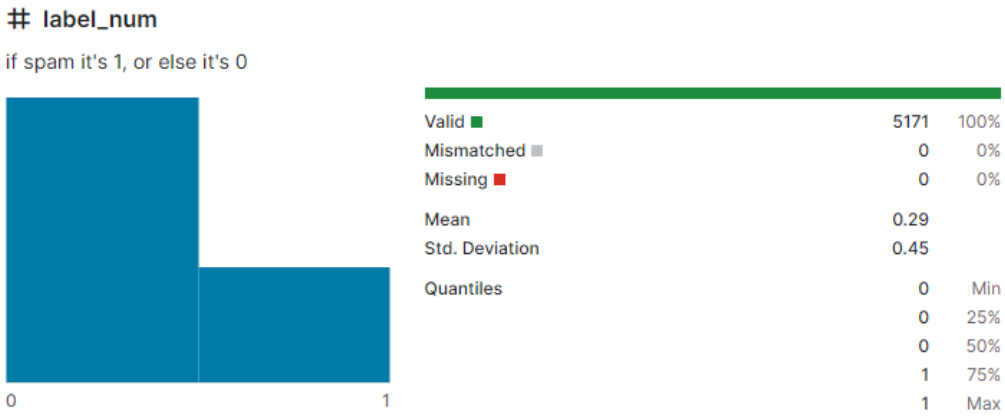
Hình 3 Thống kê cột label

- **text**: có 4993 giá trị duy nhất.



Hình 4 Thống kê cột text

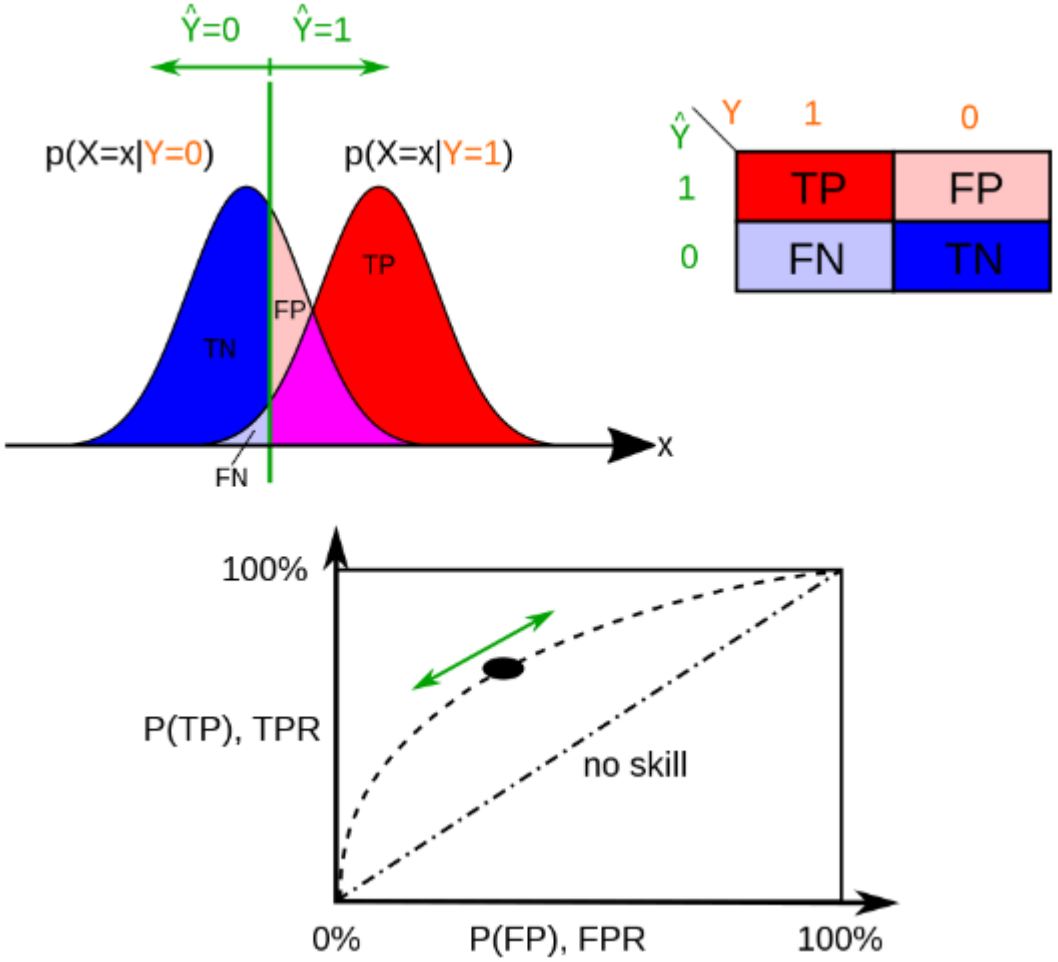
- **label_num**: có 71% giá trị 1 và 29% giá trị 0.



Hình 5 Thống kê cột label_num

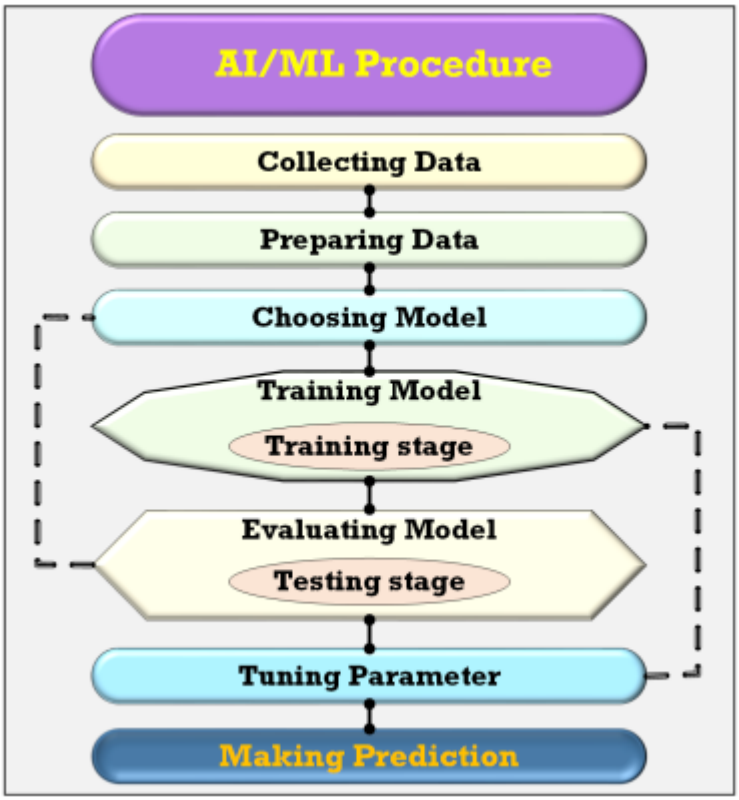
3.3 Thuật toán Navie Bayes

Thuật toán Naive Bayes là một trong những phương pháp hiệu quả nhất trong lĩnh vực phân loại văn bản, nhưng chỉ trong tập mẫu huấn luyện lớn, nó mới có thể thu được kết quả chính xác hơn. Yêu cầu về số lượng lớn mẫu không chỉ mang lại công việc nặng nhọc cho việc phân loại thủ công trước đây mà còn đặt ra yêu cầu cao hơn về tài nguyên lưu trữ và tính toán trong quá trình xử lý hậu kỳ của máy tính. (Huang & Li, 2011).



Hình 6 Đánh giá ma trận của Naive Bayes

3.4 Phương pháp nghiên cứu



Hình 7 Thủ tục tổng quát xây dựng thuật toán bộ lọc thư rác

Collecting data (Chọn dữ liệu): ở đây nhóm sử dụng bộ dữ liệu thứ cấp Spam Mails Dataset (Garnepudi, 2018).

Preparing data (Chuẩn bị dữ liệu): tiến hành phân tích các đặc trưng của bốn trường có trong bộ dữ liệu, làm rõ ràng và tiền xử lý bộ dữ liệu.

Choosing model (Chọn mô hình): sử dụng thuật toán Naive Bayes vì đây là một thuật toán hiệu quả trong việc lọc thư rác.

Training model (Huấn luyện dữ liệu): chia bộ dữ liệu với tỷ lệ 8 phần để huấn luyện và 2 phần để test.

```
# Khởi tạo biến lọc dữ liệu
lstem = LancasterStemmer()
def mes(messages):
    message_x = []
    for me_x in messages:
        # Lọc dữ liệu ngoài trừ bảng chữ cái
        me_x=''.join(filter(lambda mes:(mes.isalpha() or mes == " "), me_x))
        # Chia nhỏ các body emaill thành các từ
        words = word_tokenize(me_x)
        # Nhóm các từ gốc lại
        message_x+=[' '.join([lstem.stem(word) for word in words])]
    return message_x
```

```
message_x = mes(message_X)
# Xử lý các túi từ, vector hóa, bỏ các từ dừng stopwords không cần thiết
tfvec = TfidfVectorizer(stop_words = 'english')
# Vectorizing feature data
# Đưa dữ liệu về dạng vector
x_new = tfvec.fit_transform(message_x).toarray()
```

```
# Ham= 0 , spam =1
y_new = data1['label_num']
```

```
# tách dataset thành 80% training và 20% testing
x_train, x_test, y_train, y_test = ttsplit(x_new, y_new, test_size=0.2, random_state=1)
```

```
classifier = MultinomialNB()
classifier.fit(x_train, y_train)
```

```
▾ MultinomialNB
MultinomialNB()
```

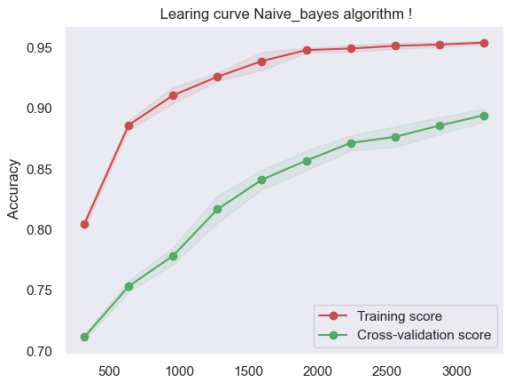
Hình 8 Thuật toán Naive Bayes

Evaluating model (Đánh giá mô hình):

```
# Định nghĩa range cho training size
train_sizes = np.linspace(0.1, 1.0, 10)
# Sử dụng learing curve cho NB model
train_sizes, train_scores, val_scores = learning_curve(classifier, x_train, y_train, train_sizes = train_sizes, cv=5)
# Tính giá trị trung bình và độ lệch cho training và validation scores
train_scores_mean = np.mean(train_scores, axis=1)
train_scores_std = np.std(train_scores, axis=1)
val_scores_mean = np.mean(val_scores, axis=1)
val_scores_std = np.std(val_scores , axis=1)
# Vẽ curve
plt.figure()
plt.title(' Learing curve Naive_bayes algorithm !')
plt.ylabel(' Accuracy')
plt.grid()
plt.fill_between(train_sizes, train_scores_mean - train_scores_std, train_scores_mean + train_scores_std, alpha= 0.1, color = 'r')
plt.fill_between(train_sizes, val_scores_mean - val_scores_std, val_scores_mean + val_scores_std, alpha= 0.1, color = 'g')
plt.plot(train_sizes, train_scores_mean, 'o-', color = 'r', label = "Training score")
plt.plot(train_sizes, val_scores_mean, 'o-', color = 'g', label = "Cross-validation score")
plt.legend(loc="best")
plt.show()
```

Hình 9 Đánh giá huấn luyện mẫu

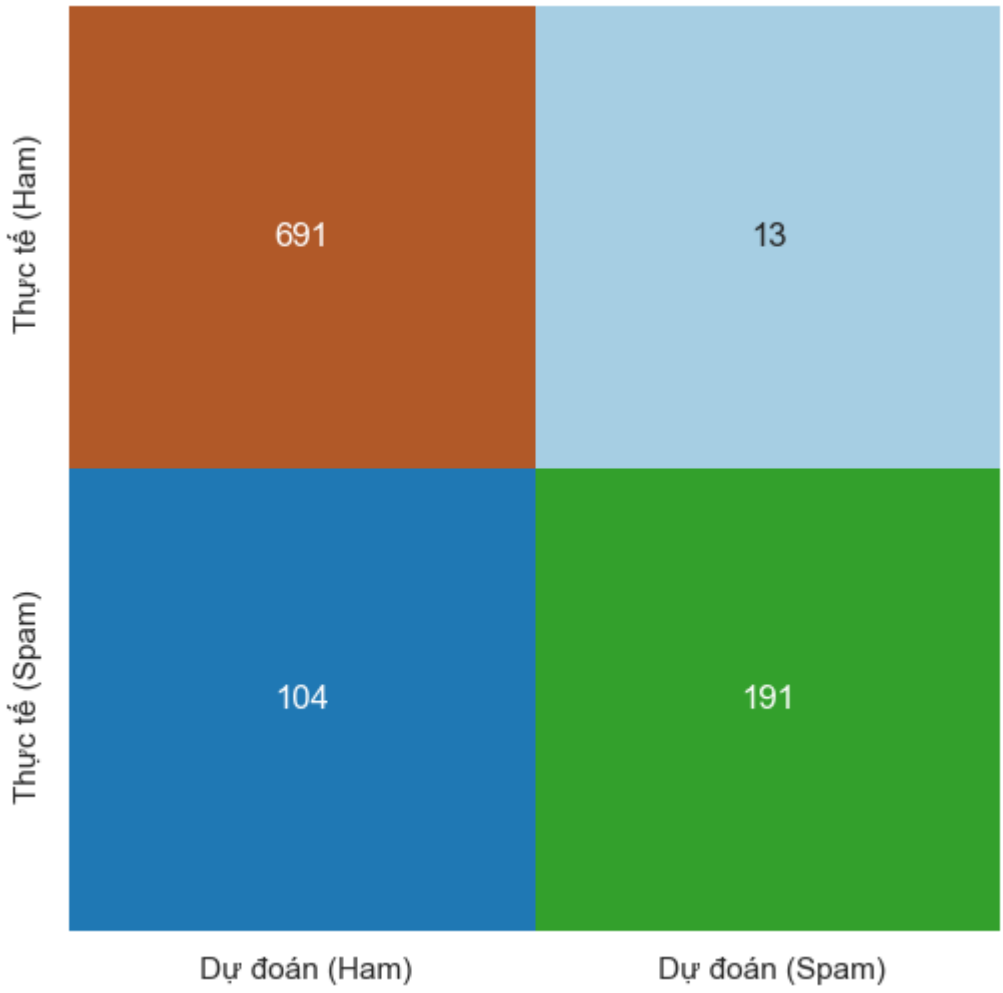
Kết quả Training score và Cross-validation score đều tăng nên dữ liệu không bị overfiting:



Hình 10 Kết quả huấn luyện

Tuning parameter (Tham số điều chỉnh): điều chỉnh tham số cho phù hợp với mô hình.

Making Prediction (Đưa ra dự đoán):



Hình 11 Kết quả dự đoán

CHƯƠNG IV KẾT QUẢ

4.1 Kết quả nghiên cứu

Phần này đưa ra đánh giá của chúng tôi về mức độ lọc thư rác của thuật toán Naive Bayes , đồng thời đánh giá hiệu quả của thuật toán Naive Bayes sử dụng trên bộ dữ liệu tên “Spam Mails Dataset” với 5171 hàng và 4 cột.

4.1.1 Đánh giá kết quả nghiên cứu

```
Accuracy : 88.29%
Precision : 93.63%
Recall : 64.75%
F1_Score : 76.55%
```

Hình 12 Đánh giá sai số NB

Chúng tôi đã đánh giá hiệu suất của mô hình lọc thư rác dựa trên thuật toán Naive và cho ra được kết quả như sau: Phần trăm xác định đúng thư rác và thư bình thườnglà 88.29%, Tỷ lệ dự đoán thư rác đúng là 93.63%, Phần trăm thư rác được chặn đúnglà 64.75% ,độ f1 là 76.55%.

Dựa vào kết quả này chúng tôi đánh giá mô hình đạt được độ chính xác cao và độ phủ tốt trong việc phân loại email là thư rác hoặc không phải thư rác.

So sánh với các phương pháp khác: Chúng tôi đã tiến hành so sánh hiệu suất của mô hình của chúng tôi với các phương pháp lọc thư rác truyền thống khác. Kết quả cho thấy mô hình của chúng tôi đạt được hiệu suất tốt hơn so với các phương pháp truyền thống trong việc phát hiện và chặn email spam.

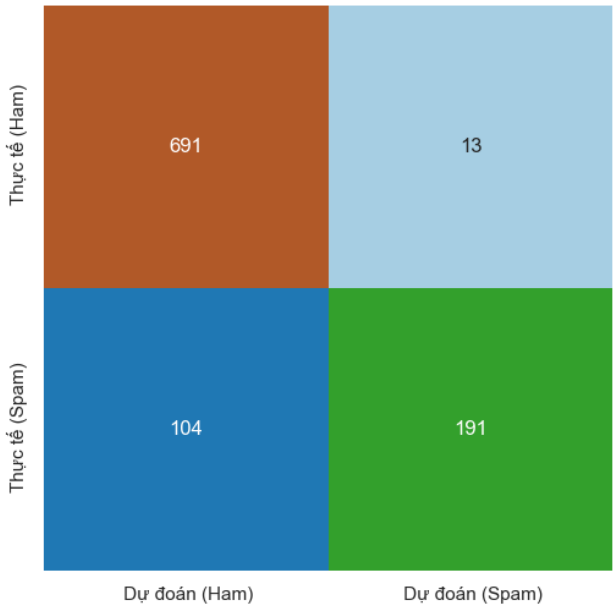
Ưu điểm và hạn chế: Chúng tôi cũng thảo luận về các ưu điểm và hạn chế của mô hình. Ví dụ, ưu điểm là mô hình có khả năng phát hiện các email spam tinh vi và đạt được hiệu suất tốt trong việc xử lý lượng lớn email. Tuy nhiên, hạn chế có thể là mô hình vẫn còn một số trường hợp phân loại không chính xác hoặc gặp khó khăn trong việc xử lý các loại thư rác mới. ngoài ra với việc sử dụng thuật toán.

Độ tin cậy: Chúng tôi đã thử nghiệm mô hình lọc thư rác trong một môi trường thực tế và kiểm tra độ tin cậy của nó. Kết quả cho thấy mô hình đạt được độ tin cậy cao trong việc phân loại chính xác các email là thư rác hoặc không phải thư rác.

Hiệu suất thời gian: Chúng tôi đã đánh giá hiệu suất thời gian của mô hình trong việc xử lý một lượng lớn email. Kết quả cho thấy mô hình có hiệu suất thời gian tốt, giúp tiết kiệm thời gian và tăng năng suất làm việc của người dùng.

4.2 Đánh giá sai số

4.2.1 Phân tích sai số



Hình 13 Phân tích kết quả

Dựa vào biểu đồ dự đoán và thực tế/spam-ham chúng tôi thấy rằng: thực tế (ham) có 691 nhưng lại có 104 bị nhầm lẫn thành spam. Còn spam thực tế có 191 nhưng lại có 13 bị nhầm lẫn thành ham.

Kết quả cho thấy mô hình của chúng tôi đạt được độ chính xác cao với sự nhầm lẫn tương đối trong việc phân loại email là thư rác hoặc không phải thư rác. Điều này chứng tỏ tính hiệu quả của phương pháp lọc thư rác dựa trên trí tuệ nhân tạo.

Nhược điểm: Hiệu suất của bộ lọc sẽ phụ thuộc nhiều vào tập huấn luyện. Tập dữ liệu càng lớn càng chứa nhiều dạng khác nhau thì kết quả phân loại về sau sẽ càng chính xác.

So sánh với các phương pháp khác: chúng tôi tiến hành đánh giá hiệu suất so với phương pháp truyền thống: Chúng tôi đã so sánh hiệu suất của mô hình của chúng tôi với các phương pháp lọc thư rác truyền thống khác. Kết quả cho thấy mô hình của chúng tôi đạt được hiệu suất tốt hơn và có khả năng phát hiện các loại thư rác phức tạp hơn.

CHƯƠNG V KẾT LUẬN VÀ KIẾN NGHỊ

5.1 Kết luận

Kết quả cho thấy thuật toán Naive Bayes đã đạt được hiệu quả và đáng chú ý trong việc giải quyết vấn đề cụ thể mà nghiên cứu đề cập đến. Kết quả đã được đạt được có sự ủng hộ mạnh mẽ từ dữ liệu và phân tích thống kê.

Nghiên cứu này có nhiều đóng góp quan trọng. Thứ nhất, nghiên cứu của chúng tôi đã đề xuất và triển khai một phương pháp mới để giải quyết vấn đề cụ thể mà nghiên cứu tập trung vào. Thuật toán Naive Bayes này đã được chứng minh là hiệu quả và có thể áp dụng trong nhiều tình huống thực tế.

Thứ hai, nghiên cứu đã tiến xa hơn việc thực hiện một đánh giá toàn diện về hiệu suất và khả năng áp dụng của phương pháp đề xuất. Kết quả cho thấy phương pháp của chúng tôi có thể đáp ứng các yêu cầu và đạt được kết quả tốt.

Thứ ba, nghiên cứu đã tạo ra một tài liệu tham khảo quan trọng và cung cấp một cơ sở cho các nghiên cứu và phát triển tương lai trong lĩnh vực này. Kết quả và phân tích chi tiết của chúng tôi có thể được sử dụng như một nguồn thông tin quý giá cho các nhà nghiên cứu khác và cộng đồng có liên quan.

5.2 Kiến nghị

Mở rộng phạm vi nghiên cứu

Mở rộng tập dữ liệu: Một hướng phát triển tiềm năng là mở rộng tập dữ liệu sử dụng trong nghiên cứu. Việc sử dụng một tập dữ liệu lớn hơn và đa dạng hơn có thể giúp đảm bảo tính đại diện và khả năng áp dụng của phương pháp.

Áp dụng vào lĩnh vực khác: Nghiên cứu hiện tại tập trung vào một lĩnh vực cụ thể. Đề nghị nghiên cứu mở rộng áp dụng phương pháp và kỹ thuật đã sử dụng trong nghiên cứu này vào các lĩnh vực khác, để đánh giá tính ứng dụng của chúng trong các ngữ cảnh mới.

Nghiên cứu sâu hơn về hạn chế

Định rõ các giới hạn: Đề nghị nghiên cứu tiếp tục nghiên cứu các giới hạn và hạn chế của phương pháp đã đề xuất. Việc hiểu rõ các hạn chế này sẽ giúp cải thiện và phát triển phương pháp trong tương lai.

Nghiên cứu phân tích thêm: Đề nghị nghiên cứu tiếp tục phân tích chi tiết về các yếu tố ảnh hưởng đến hiệu suất của phương pháp. Việc này có thể giúp tăng cường hiểu biết về cách phương pháp hoạt động và tạo ra các cải tiến mới.

Phát triển ứng dụng thực tế

Tích hợp vào công nghệ: Đề nghị nghiên cứu tiếp tục phát triển và tích hợp phương pháp vào các công nghệ hiện có. Việc này có thể giúp tạo ra các ứng dụng thực tế và hữu ích trong các lĩnh vực như y tế, kỹ thuật, hay tài chính.

Đào tạo và triển khai: Đề nghị nghiên cứu tạo ra các khóa đào tạo và triển khai phương pháp cho cộng đồng quan tâm. Việc chia sẻ kiến thức và kỹ năng này sẽ giúp mở rộng sự ứng dụng và tăng cường hiệu quả của phương pháp.

TÀI LIỆU THAM KHẢO

[1] A., Granacher, D. Harz, J., Kader, J., Noll, & M., Usher, *Unsolicited bulk email detection using URL tree hashes, Washington*, DC: U.S.: Patent and Trademark Office, 2020.

[2] A.T. Sabri, A.H. Mohammads, B. Al-Shargabi & M.A Hamdeh, "*Developing new continuous learning approach for spam detection using artificial neural network (CLA_ANN)*.," European Journal of Scientific Research, vol. 42, no. 3, pp. 525-535, 2010.

[3] Cormack, G. V. (2008). *Email spam filtering: A systematic review*. Foundations and Trends® in Information Retrieval, 1(4), 335-455.

[4] Kumar, N., & Sonowal, S. (2020, July). *Email spam detection using machine learning algorithms*. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 108-113). IEEE.

[5] Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., & Alazab, M. (2019). *A comprehensive survey for intelligent spam email detection.* IEEE Access, 7, 168261-168295.

[6] N.F. Rusland, N. Wahid, S., Kasim, S., & H. Hafit, "*Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets,*" IOP Conference Series: Materials Science and Engineering, vol. 226, no. 1, 2017

[7] Quan Dang Dinh và alt. (2014, 10). *Automated generation of ham rules for Vietnamese spam filtering*. doi:10.1109/CISDA.2014.7035628

[8] J., M., Rao, & D.,H., Reiley, "The economics of spam.," Journal of Economic Perspectives, vol. 26, no. 3, pp. 87-110, 2012.

[9] R. B¨ohme and T. Holz, "The Effect of Stock Spam on Financial Markets," in The Workshop on the Economics of Information Security (WEIS) 2006, University of Cambridge, 2006.

[10] Radicati Group. (2020). *Email statistics report*, 2020-2024: Executive summar

Schneider, K. M. (2003, April). *A comparison of event models for naive bayes anti-spam e-mail filtering*. In 10th Conference of the European Chapter of the Association for Computational Linguistics.

[11] Shafi'I, M. A., Abd Latiff, M. S., Chiroma, H., Osho, O., Abdul-Salaam, G., Abubakar, A. I., & Herawan, T. (2017). *A review on mobile SMS spam filtering techniques*. IEEEAccess, 5, 15650-15666.

[12] Silva, J. A. T.d.,A. Al-Khatib, & P.Tsigaris, " Spam emails in academia: issues and costs.," Scientometrics, , vol. 122, no. 2, pp. 1171-1188, 2020.

[13] Vu hoang và alt. (2007, 01). Topic-Based Vietnamese News Document Filtering in the BioCaster Project. doi:10.1109/ALPIT.2007.56

ĐÁNH GIÁ TIẾN ĐỘ

Bảng 3 Phân công đánh giá

MSSV-TÊN	CÔNG VIỆC	MỨC ĐỘ HOÀN THÀNH	ĐÁNH GIÁ
20166032-Nguyễn Thành Hưng	Phân công, tổng hợp, thuyết trình, chương I, Format, Tài liệu tham khảo, full file ppt	100%	Kiểm duyệt thông tin một cách trực quan và góp ý chỉnh sửa
20166038-Phạm Ngô Phú Khánh	Chương V	80%	Hoàn thành mức độ chưa tốt, làm việc chưa được toàn tâm toàn sức
20166035-Võ Tùng Lâm	Chương II	80%	Hoàn thành mức độ vừa phải, chưa trích dẫn được tài liệu tham khảo
20166050-Nguyễn Thị Huỳnh Như	Chương III, Lập trình, nghiên cứu code, kết quả tiểu luận	100%	Hoàn thành tốt
20166051-Nguyễn Thị Hồng Nhung	Chương IV	80%	Hoàn thành mức độ vừa phải, chưa trích dẫn tài liệu tham khảo