**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

**TÓM TẮT**

Lập trình bộ lọc thư rác qua email 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.

Bộ lọc thư rác hoạt động bằng cách sử dụng các quy tắc hoặc thuật toán phân loại email để xác định xem một email cụ thể có phải là thư rác hay không. Các quy tắc này có thể dựa trên nhiều yếu tố, chẳng hạn như từ khóa, địa chỉ email nguồn, địa chỉ IP hoặc bất kỳ thuộc tính nào của nội dung email. Các thuật toán phân loại thường được đào tạo bằng cách sử dụng dữ liệu được gắn nhãn trước để xác định các mẫu bằng thuật toán học máy. Bộ lọc thư rác có thể được triển khai ở nhiều cấp độ khác nhau, từ máy khách (được cài đặt trong phần mềm email của người dùng) đến máy chủ email (được cài đặt trong máy chủ email của nhà cung cấp email). Người dùng có thể đặt bộ lọc thư rác cho mục đích sử dụng cá nhân.

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**

[**TÓM TẮT** 3](#_Toc147325133)

[**MỤC LỤC** 4](#_Toc147325134)

[**DANH SÁCH CÁC CHỮ VIẾT TẮT, BẢNG SỐ, HÌNH VÀ BIỂU ĐỒ** 6](#_Toc147325135)

[**CHƯƠNG 1: MỞ ĐẦU** 7](#_Toc147325136)

[**1.** **Tính cấp thiết của đề tài** 7](#_Toc147325137)

[**2.** **Ý nghĩa khoa học và thực tiễn** 7](#_Toc147325138)

[**3.** **Mục đích/mục tiêu hoặc yêu cầu nghiên cứu** 9](#_Toc147325139)

[**4.** **Đối tượng nghiên cứu** 10](#_Toc147325140)

[**CHƯƠNG 2: TỔNG QUAN** 11](#_Toc147325141)

[**1.** **Cơ sở lý thuyết** 11](#_Toc147325142)

[**2.** **Thông tin bối cảnh** 21](#_Toc147325143)

[**a.** **Trên thế giới** 21](#_Toc147325144)

[**b.** **Tại Việt Nam** 22](#_Toc147325145)

[**CHƯƠNG 3: NỘI DUNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU** 23](#_Toc147325146)

[**1.** **Vật liệu nghiên cứu** 23](#_Toc147325147)

[**a.** **Dữ liệu** 23](#_Toc147325148)

[**b.** **Phần mềm** 23](#_Toc147325149)

[**2.** **Phương pháp nghiên cứu** 23](#_Toc147325150)

[**a.** **Phạm vi nghiên cứu** 23](#_Toc147325151)

[**b.** **Thời gian thực hiện** 24](#_Toc147325152)

[**c.** **Kết quả kỳ vọng** 24](#_Toc147325153)

[**d.** **Cách tiếp cận** 24](#_Toc147325154)

[**e.** **Quy trình thực hiện** 25](#_Toc147325155)

[**3.** **Ưu điểm và nhược điểm của phương pháp nghiên cứu** 26](#_Toc147325156)

[**a.** **Ưu điểm** 34](#_Toc147325157)

[**b.** **Nhược điểm** 34](#_Toc147325158)

[**CHƯƠNG 4: KẾT QUẢ VÀ THẢO LUẬN** 35](#_Toc147325159)

[**1.** **Kết quả** 35](#_Toc147325160)

[**2.** **Thảo luận** 35](#_Toc147325161)

[**CHƯƠNG 5: KẾT LUẬN VÀ ĐỀ NGHỊ** 37](#_Toc147325162)

[**1.** **Kết luận** 37](#_Toc147325163)

[**2.** **Đề nghị** 37](#_Toc147325164)

[**TÀI LIỆU THAM KHẢO** 39](#_Toc147325165)

**DANH SÁCH CÁC CHỮ VIẾT TẮT, BẢNG SỐ, HÌNH VÀ BIỂU ĐỒ**

Support Vector Machines (SVM)

exploratory-data-analysis (EDA)

# **CHƯƠNG 1: MỞ ĐẦU**

1. **Tính cấp thiết của đề tài**

Đề tài "Ứng dụng Học máy xây dựng bộ lọc thư rác" là một vấn đề cấp thiết trong thời đại công nghệ thông tin hiện nay. Theo tác giả Nazirova, trong bài báo " Survey on Spam Filtering Techniques " (Communications and Network, 2011, 3, 153-160), tình trạng email spam đã trở thành một vấn đề nghiêm trọng trong việc quản lý và sử dụng email.

Email spam, hay còn được gọi là thư rác, là những email không mong muốn được gửi đến người dùng mà không được yêu cầu. Thư rác không chỉ gây phiền phức cho người nhận mà còn có thể gây hại bằng cách chứa mã độc hoặc liên kết đến các trang web độc hại. Theo Tạp chí Khoa học và Công nghệ Việt Nam (2019), các email spam chiếm khoảng 50-70% tổng lượng email được gửi hàng ngày trên toàn cầu.

Với sự phát triển không ngừng của công nghệ thông tin, các phương pháp lọc thư rác cũng đã được nghiên cứu và đưa ra. Theo Marcus Costa Sampaio trong bài báo "Machine Learning Techniques Applied to Email Spam Filtering" (Journal of Information and Data Management, 2017), các phương pháp lọc thư rác truyền thống như dựa vào danh sách đen hoặc quy tắc đã không còn hiệu quả trong việc chặn những email spam ngày càng tinh vi.

Để giải quyết vấn đề này, “Ứng dụng Học máy xây dựng bộ lọc thư rác” là một nhiệm vụ cực kỳ cấp thiết. Theo Yiming Ying và Hui Luo trong bài báo "A personal spam filter using sender time patterns" (International Conference on Control Engineering and Artificial Intelligence, 2017), việc áp dụng các kỹ thuật học máy như mạng neural, học tăng cường, hoặc phân lớp để lọc email spam đã cho thấy hiệu quả cao.

Bên cạnh đó, việc nghiên cứu và phát triển các phương pháp lọc thư rác trong email không chỉ giúp người dùng tiết kiệm thời gian và tăng năng suất làm việc mà còn đảm bảo sự an toàn và bảo mật thông tin cá nhân. “Ứng dụng Học máy xây dựng bộ lọc thư rác” là một đề tài cực kỳ cấp thiết nhằm đáp ứng nhu cầu ngày càng cao của người dùng và hỗ trợ trong việc quản lý và sử dụng email một cách hiệu quả và an toàn. Từ thực tế trên đề tài “Ứng dụng Học máy xây dựng bộ lọc thư rác” được tiến hành.

1. **Ý nghĩa khoa học và thực tiễn**

Theo tác giả Saadat của bài báo " Survey on Spam Filtering Techniques ", việc lập trình bộ lọc thư rác là một phương pháp hữu hiệu để giảm tải công việc và tăng cường an ninh thông tin. Điều này đặc biệt quan trọng trong thế giới kỹ thuật số hiện nay, khi mà lượng email thư rác không ngừng gia tăng và có thể gây ánh hưởng tiêu cực đến sự tập trung và năng suất làm việc của người dùng.

Ý nghĩa khoa học của việc lập trình bộ lọc thư rác nằm ở việc áp dụng các thuật toán và kỹ thuật máy học để phân loại các email. Theo tác giả Ricardo Baeza-Yates và Berthier Ribeiro-Neto trong sách "Modern Information Retrieval", việc phát hiện và loại bỏ thư rác dựa trên các tiêu chí như từ khóa, địa chỉ email gửi, đánh giá tin tưởng của người gửi và nội dung email giúp tăng cường khả năng xác định thư rác và tránh nhầm lẫn với các email quan trọng.

Ứng dụng thực tiễn của bộ lọc thư rác giúp người dùng tiết kiệm thời gian bằng cách tự động loại bỏ các email không mong muốn, giúp dễ dàng tìm kiếm các email quan trọng hơn. Nhờ vào việc lập trình bộ lọc thư rác, người dùng không cần phải xem qua hàng trăm hoặc thậm chí hàng ngàn email mỗi ngày để tìm kiếm những email quan trọng. Điều này giúp tăng cường hiệu suất làm việc và cải thiện trải nghiệm người dùng.

Trong tổng quan, việc lập trình bộ lọc thư rác trong email không chỉ đơn giản là một công cụ tiện ích giúp tổ chức và sắp xếp email hiệu quả, mà còn là một ứng dụng của khoa học và công nghệ thông tin trong việc tự động hoá quản lý thông tin và tăng cường hiệu suất làm việc.

1. **Lý do chọn đề tài**

Chủ đề “ Ứng dụng Học máy xây dựng bộ lọc thư rác” được chọn vì nhu cầu ngăn chặn thư rác ngày càng tăng cũng như những vấn đề mà nó gây ra cho người sử dụng email. Theo báo cáo của Radicati Group, 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 và lập trình một bộ lọc thư rác trong email tùy chỉnh sẽ giúp người dùng có thể kiểm soát và quản lý thư rác một cách tốt hơn.

Theo tài liệu của Deitel, bộ lọc thư rác (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, machine learning.

1. **Mục đích nghiên cứu**

Mục đích của “Ứng dụng Học máy xây dựng bộ lọc thư rác” là để phát triển các thuật toán và phương pháp giúp tự động nhận diện và loại bỏ các email không mong muốn, gửi thư rác hay email lừa đảo khỏi hộp thư đến của người dùng.

* Giảm tiêu hao thời gian: Bộ lọc thư rác giúp người dùng tiết kiệm thời gian bằng cách tự động tách riêng các email hữu ích từ email thư rác, giúp người dùng dễ dàng quản lý hộp thư và truy cập vào những email quan trọng.
* Bảo mật thông tin: Các email lừa đảo hoặc email thư rác có thể chứa các đường dẫn độc hại, mã độc, hoặc yêu cầu thông tin cá nhân. Bộ lọc thư rác giúp ngăn chặn những email có khả năng gây hại này để bảo vệ thông tin cá nhân và máy tính cá nhân.
* Tăng hiệu suất làm việc: Bằng cách tự động loại bỏ email thư rác, người dùng sẽ không phải mất thời gian xử lý và xóa chúng thủ công, giúp tăng hiệu suất và năng suất làm việc.
* Cải thiện trải nghiệm người dùng: Nhận nhiều email thư rác có thể làm cho hộp thư đến trở nên lộn xộn và khó quản lý. Bộ lọc thư rác giúp trải nghiệm người dùng trở nên dễ dàng hơn và tăng cường sự hài lòng của người dùng với dịch vụ email.
* Phòng ngừa các hoạt động lừa đảo: Các email lừa đảo có thể gây thiệt hại lớn cho người dùng bằng cách lừa đảo và lấy cắp thông tin cá nhân hoặc tài sản. Bộ lọc thư rác giúp ngăn chặn các email lừa đảo trước khi người dùng cung cấp thông tin nhạy cảm hoặc thực hiện các hành động không an toàn.

1. **Đối tượng nghiên cứu**

Đối tượng nghiên cứu trong chủ bao gồm các phương pháp và thuật toán để phát hiện và loại bỏ các email không mong muốn (spam) từ hộp thư đến của người dùng. Các đối tượng trong quá trình nghiên cứu bao gồm:

* Tập dữ liệu: Đối tượng nghiên cứu cần tạo và sử dụng tập dữ liệu chứa các email gồm cả thư rác và thư thường để huấn luyện và đánh giá hiệu quả của bộ lọc.
* Thuật toán học máy: Đối tượng nghiên cứu cần tìm hiểu và áp dụng các thuật toán học máy như học có giám sát (supervised learning), học không giám sát (unsupervised learning), mạng neuron nhân tạo (artificial neural networks) và học sâu (deep learning) để xây dựng mô hình dự đoán xác suất email là thư rác.
* Kỹ thuật rút trích đặc trưng: Đối tượng nghiên cứu cần phân tích và chọn ra những đặc trưng quan trọng từ nội dung và tiêu đề của email như từ khóa, từ ngữ, độ dài, địa chỉ gửi, đính kèm, và các thông tin liên quan khác để sử dụng trong quá trình dự đoán xác suất là thư rác.
* Phương pháp đánh giá: Đối tượng nghiên cứu cần đánh giá hiệu quả của bộ lọc thư rác bằng cách sử dụng các phương pháp như chia dữ liệu thành tập huấn luyện và tập kiểm tra, sử dụng các chỉ số đánh giá như độ chính xác (accuracy), độ chính xác dự đoán thư rác (spam precision), độ chính xác dự đoán thư thường (ham precision), độ nhạy (recall), và điểm F1 (F1 score).
* Các vấn đề liên quan: Đối tượng nghiên cứu cần tìm hiểu và đưa ra giải pháp cho các vấn đề liên quan trong lĩnh vực Bộ lọc Thư rác như phát hiện các email lừa đảo (phishing), xử lý email bất hợp pháp, và cải thiện độ chính xác của mô hình qua thời gian với dữ liệu mới.

# **CHƯƠNG 2: TỔNG QUAN**

1. **Cơ sở lý thuyết**
   1. **Tổng quan về thư rác**

**Định nghĩa**

Có rất nhiều định nghĩa khác nhau cho thư rác. 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 email 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. Định nghĩa này có tính đến các đặc điểm của thư điện tử hàng loạt. Các định nghĩa về thư rác đều có những đặc điểm chung như sau:

* Gửi bằng thư điện tử
* Sử dụng các địa chỉ được thu thập mà không có sự đồng ý
* Gửi hàng loạt
* Không mong muốn nhận
* Lặp đi lặp lại
* Nhầm mục đích thương mại hoặc tài chính
* Không có mục tiêu và bừa bãi
* Không thể ngăn cản
* Ẩn danh hoặc ngụy trang
* Nội dung bất hợp pháp hoặc xúc phạm
* Nội dung lừa đảo hoặc gian lận

**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 email 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.

* 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 hoặc tài khoản của họ. 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 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.

* 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. 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. 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. 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.

* 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. Vi rút là một chương trình, giống như vi rút 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 10 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.

* 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.

Tác hại của thư rác

Thư rác hiện chiếm một tỷ lệ rất lớn trong tổng số thư điện tử được gửi qua Internet. Theo các thống kê khác nhau, thư rác chiếm từ 71% đến 87% lượng thư gửi qua các máy chủ thư điện tử. Lượng thư rác quá lớn gây ra nhiều tác hại đối với sự phát triển Internet nói chung và người sử dụng thư điện tử nói riêng. Có thể kể ra một số tác hại của thư rác:

* Thư rác gây thiệt hại về kinh tế cho người nhận thư trong trường hợp người nhận thư phải trả tiền cho lượng thông tin truyền qua mạng.
* Thư rác có thể làm đầy hộp thư người nhận và do vậy làm thất lạc những thư bình thường đến sau.
* Thư rác làm tốn thời gian do người nhận phải mở thư và xoá thư khỏi hộp thư của mình.
* Thư rác gây tâm lý xấu đối với người sử dụng thư điện tử. Theo thống kê tại “pewinternet” , 71% người sử dụng thư điện tử coi thư rác là trở ngại lớn khi sử dụng dịch vụ Internet này.
* Thư rác chiếm một phần đường truyền Internet và làm tốn thời gian xử lý của máy chủ.
  1. **Tổng quan về phương pháp nghiên cứu**

1.2.1 Lọc thư sử dụng phân loại Bayes đơn giản

Để sử dụng phân loại Bayes đơn giản, mỗi thư (phần nội dung) được biểu diễn bởi một vectơ ****= (*x*1, *x*2, …, *x*n), trong đó *x*1, *x*2, …, *x*nlà giá trị của đặc trưng *X1*, *X2*,…, *Xn*. Mỗi đặc trưng có thể là một từ hoặc một cụm từ. Ở đây, *n* là số lượng đặc trưng được xác định từ toàn bộ tập dữ liệu huấn luyện, tức là số lượng từ/cụm từ khác nhau trong tập dữ liệu huấn luyện. Cách xác định *n* sẽ được đề cập trong một phần sau của bài báo. Mỗi thư được gán một nhãn phân loại *Y* có thể nhận một trong hai giá trị: *Y* = 1 cho trường hợp thư rác và *Y* = 0 cho trường hợp thư bình thường.

Để xác định nhãn phân loại cho thư, bộ phân loại Bayes tính xác suất điều kiện

*P* (*Y* = *y* | *X*1 = *x*1,…, *Xn* = *xn*)

tức là xác suất một thư với nội dung (*x*1, *x*2, …, *x*n) nhận nhãn phân loại *y*, *y* ∈ {1,0}. Sử dụng công thức Bayes, xác suất trên được tính như sau

 (1)

Trong công thức (1), giá trị mẫu số không phụ thuộc vào nhãn phân loại và do vậy có thể bỏ qua. Nhãn phân loại *Y* là nhãn tương ứng với giá trị lớn nhất của tử số. Cụ thể, trong trường hợp phân loại thư rác, nhãn của thư được xác định bằng cách tính giá trị biểu thức:

 (2)

Giá trị biểu thức (2) lớn hơn 1 có nghĩa xác suất thư là thư rác lớn hơn xác suất thư bình thường và thư sẽ được gán nhãn thư rác. Giá trị biểu thức (2) nhỏ hơn 1 cho kết quả ngược lại. Tuy nhiên, đối với bài toán lọc thư, lỗi phân loại thư rác thành thư thường dễ chấp nhận hơn nhiều so với phân loại nhầm thư thường thành thư rác. Do vậy, không thể chỉ đơn thuần gán cho thư nhãn có xác suất điều kiện lớn hơn. Giả sử việc phân loại thư rác thành thư thường dễ chấp nhận hơn phân loại thư thường thành thư rác *T* lần, theo lý thuyết phân loại bayes tối ưu trình bày trong [Duda &Hard], thư sẽ được phân loại là rác nếu

 (3)

Trên thực tế, giá trị *T* là tham số và có thể do người sử dụng xác định.

Xác suất *P*(*Y* = *y*) trên tập dữ liệu huấn luyện có thể tính dễ dàng bằng cách đếm tần suất xuất hiện của thư có nhãn “rác” và “bình thường”. Việc xác định *P* ( = ****| *Y* = *y*) phức tạp hơn nhiều do phải tính tất cả các tổ hợp giá trị của vectơ  và đòi hỏi lượng dữ liệu huấn luyện lớn tương ứng. Để khắc phục vấn đề này, phương pháp Bayes đơn giản sử dụng một số giả thiết về tính độc lập xác suất của các đặc trưng nếu đã biết nhãn phân loại. Có một số cách tính giá trị *P* ( = ****| *Y* = *y*) khác nhau tương ứng với các phiên bản khác nhau của phương pháp phân loại văn bản sử dụng Bayes đơn giản.

**Phân loại Bayes đơn giản với mô hình Bernoulli đa trị**

Giả sử *F* = (*f*1, *f*2, …,*fn*) là tập hợp toàn bộ *n* đặc trưng. Trong mô hình đang xét, mỗi thuộc tính *Xi* được coi là một biến nhị phân ngẫu nhiên có thể nhận một trong hai giá trị 0 và 1. *Xi* = 1 nếu đặc trưng *fi* xuất hiện trong thư và *Xi* = 0 trong trường hợp ngược lại (không quan tâm đặc trưng đó xuất hiện chính xác bao nhiêu lần). Mô hình Bernoulli đa trị coi nội dung của một thư bất kỳ với nhãn *y* là kết quả của *n* thực nghiệm Bernoulli (tung đồng xu sấp ngửa), mỗi thực nghiệm cho phép xác định đặc trưng *Xi* nhận giá trị 1 hay 0 với xác suất *P*(*fi* | *y*) và 1 − *P*(*fi* | *y*) tương ứng. Mô hình Bayes đơn giản giả thiết rằng kết quả các thực nghiệm độc lập với nhau nếu biết nhãn phân loại *Y*. Rõ ràng, giả thiết này là không hợp lý vì các từ trong một thư không xuất hiện độc lập với nhau, từ đây mà có tên gọi Bayes “đơn giản”. Tuy nhiên, sử dụng giả thiết về độc lập xác suất cho phép đơn giản hoá việc tính xác suất trong *P* ( = ****| *Y* = *y*) và trên thực tế vẫn cho kết quả phân loại tốt. Giả thiết này cũng được sử dụng trong tất cả các phiên bản của phân loại Bayes đơn giản.

Với mô hình Bernoulli đa trị ở trên, ta có:

 (4)

Trong đó, xác suất *P*(*fi*| *Y* = *y*) là tỷ lệ thư với nhãn *y* đồng thời có chứa *fi* trong số thư có nhãn *y*. Tỷ lệ này được tính trên tập dữ liệu huấn luyện.

Trong trường hợp dữ liệu huấn luyện không đủ lớn, một vấn đề có thể nảy sinh khi tính *P*(*fi* | *Y* = *y*) là thư với nhãn *y* và chứa *fi* không xuất hiện trong tập huấn luyện. Khi đó, *P*(*fi* | *Y* = *y*) = 0 và toàn bộ biểu thức (4) sẽ bằng 0. Để giải quyết vấn đề này, xác suất *P*(*fi* | *Y* = *y*) được tính như sau:

 (5)

Trong đó, *Ny*,*fi* là số lượng thư với nhãn *y* và có chứa *fi*, *Ny* là toàn bộ số lượng thư với nhãn *y*. 1 và 2 là các hằng số được thêm vào sao cho xác suất tiền nghiệm của mỗi phân loại bằng nhau và bằng 1/2 khi chưa có dữ liệu.

Thay thế (4) vào (3), ta có thể phân loại thư là thư rác nếu:

 (6)

**Phân loại Bayes đơn giản với mô hình đa thức**

Trong mô hình đa thức, mỗi đặc trưng *Xi* có thể nhận giá trị *xi* trong đó *xi* = {0,1,2,3…}là số lần xuất hiện của *fi* trong thư *d* tương ứng. Mô hình đa thức coi nội dung thư *d* sinh ra bằng cách lấy ngẫu nhiên có lặp |*d*| đặc trưng từ tập đặc trưng chung *F* với xác suất *P*(*fi*| *y*) cho mỗi *fi* (|*d*| là số lượng đặc trưng trong thư *d*). Sử dụng thêm giả thiết là |d| không phụ thuộc vào nhãn *Y* cho phép tính *P* ( = ****| *Y* = *y*) theo xác suất đa thức như sau:

 (7)

Xác suất *P*(*fi* | *Y* = *y*) được tính từ dữ liệu huấn luyện theo công thức

 (8)

Trong đó *Ny* là số lượng thư với nhãn phân loại *y* và *Ny*,*fi* là tổng số lần đặc trưng *fi* xuất hiện trong các thư có nhãn y.

Tiêu chuẩn phân loại thư rác (3) khi đó trở thành:

 (9)

Một phiên bản đơn giản của mô hình đa thức tiến hành phân loại theo công thức (9) tuy nhiên các đặc trưng chỉ nhận giá trị 0 hoặc 1, tức là *xi* = {0,1} thay vì *xi* = {0,1,2,3…}. Kết quả thử nghiệm trình bày trong [] cho thấy mặc dù đơn giản, mô hình đa thức với các đặc trưng là biến nhị phân cho kết quả phân loại tương đương hoặc tốt hơn mô hình đa thức thông thường. Do vậy, trong phần thử nghiệm ở dưới, chúng tôi chỉ trình bày kết quả cho mô hình đa thức với đặc trưng là biến nhị phân.

1.2.2 Lọc thư sử dụng Support Vector Machines (SVM)

SVM là phương pháp phân loại tự động tương đối mới , có cơ sở lý thuyết rõ ràng dựa trên khái niệm chiều VC về khả năng khái quát hoá cho phân loại mẫu mới. Ngoài ra, nhiều thử nghiệm với các ứng dụng khác nhau cho thấy SVM là một trong những phương pháp phân loại có độ chính xác cao nhất.

SVM cũng là phương pháp phân loại được sử dụng nhiều trong phân loại văn bản. Ưu điểm của SVM đối với phân loại văn bản là phương pháp này có khả năng làm việc với dữ liệu nhiều thuộc tính trong đó có những thuộc tính không liên quan đến nhãn phân loại .

Giả sử dữ liệu huấn luyện bao gồm *n* mẫu được cho dưới dạng , i=1…*n*, trong đó  là véctơ bao gồm *m* phần tử chứa giá trị của *m* thuộc tính hay đặc trưng và *yi* là nhãn phân loại có thể nhận giá trị +1 hoặc -1. Có thể hình dung dữ liệu như các điểm trong không gian ơclit *m* chiều và được gán nhãn. SVM được xây dựng trên cơ sở hai ý tưởng chính.

*Ý tưởng thứ nhất* là ánh xạ dữ liệu gốc sang một không gian mới gọi là *không gian đặc trưng* với số chiều lớn hơn sao cho trong không gian mới có thể xây dựng một siêu phẳng cho phép phân chia dữ liệu thành hai phần riêng biệt, mỗi phần bao gồm các điểm có cùng nhãn phân loại. Ý tưởng ánh xạ sang không gian đặc trưng được minh hoạ trên hình 3.1.

Không gian gốc

Không gian đặc trưng

**Hình 1.** Ánh xạ dữ liệu từ không gian gốc sang không gian đặc trưng cho phép phân chia dữ liệu bởi siêu phẳng

*Ý tưởng thứ hai* là trong số những siêu phẳng như vậy cần lựa chọn siêu phẳng có lề lớn nhất. Lề ở đây là khoảng cách từ siêu phẳng tới các điểm gần nhất nằm ở hai phía của siêu phẳng (mỗi phía tương ứng với một nhãn phân loại). Lưu ý rằng siêu phẳng nằm cách đều các điểm gần nhất với nhãn khác nhau. Trên hình 3.2. là minh hoạ siêu phẳng (đường liền nét) với lề cực đại tới các điểm dữ liệu biểu diễn bởi các hình tròn và hình vuông.

**Hình 2.** Siêu phẳng với lề cực đại cho phép phân chia các hình vuông khỏi các hình tròn trong không gian đặc trưng

Để tránh việc tính toán trực tiếp với dữ liệu trong không gian mới, ta sử dụng một phương pháp gọi là *thủ thuật nhân* bằng cách tìm một *hàm nhân* (kernel function) *K* sao cho:

 (10)

Sử dụng phương pháp nhân tử Lagrăng và thay thế tích vô hướng của hai vectơ bằng giá trị hàm nhân theo công thức (10), bài toán tìm lề cực đại của SVM được đưa về bài toán quy hoạch toán học bậc hai như sau:

Tìm vectơ hệ số  cho phép cực tiểu hoá hàm mục tiêu

 (11)

đồng thời thoả mãn các điều kiện

 (12)

Và 0 ≤ *αi* ≤ *C* (13)

Trong (11), (12), (13),  và *yi* tương ứng là dữ liệu và nhãn phân loại của ví dụ huấn luyện thứ *i*, *αi* là hệ số cần xác định. Trong ràng buộc (13), *C* là số lượng tối đa các điểm dữ liệu có phân loại sai, tức là các điểm nằm ở phía này của siêu phẳng nhưng lại có nhãn của các điểm nằm ở bên kia. Việc sử dụng *C* cho phép khắc phục tình trạng dữ liệu huấn luyện có các ví dụ bị gán nhãn không chính xác.

Quá trình huấn luyện SVM là quá trình xác định *αi* . Có nhiều phương pháp cho phép tính *αi* từ dữ liệu huấn luyện một cách hiệu quả trong đó thông dụng nhất là SMO .

Sau khi huấn luyện xong, giá trị nhãn phân loại cho một ví dụ mới  sẽ được tính bởi

 (14)

Ở đây, *b* được tính trong giai đoạn huấn luyện theo công thức sau

 (15)

trong đó *i* là một hệ số thoả mãn điều kiện 0 < *αi* < *C*.

Đối với bài toán phân loại thư điện tử,  là vectơ đặc trưng biểu diễn cho nội dung thư như trong phần phân loại Bayes và *yi* là nhãn phân loại đối với dữ liệu huấn luyện. Tương tự như phần phân loại Bayes, giá trị *xi* có thể là 0 hoặc 1. Thư mới được phân loại theo công thức (7): giá trị âm là thư bình thường, trong khi giá trị dương tương ứng với thư rác.

**1.2.3 Phân tích dữ liệu thăm dò (exploratory-data-analysis)**

**Khái niệm**

Phân tích dữ liệu khám phá (EDA) được các nhà khoa học dữ liệu sử dụng để phân tích và điều tra các tập dữ liệu cũng như tóm tắt các đặc điểm chính của chúng, thường sử dụng các phương pháp trực quan hóa dữ liệu. EDA chủ yếu được sử dụng để xem dữ liệu nào có thể tiết lộ ngoài nhiệm vụ kiểm tra giả thuyết hoặc lập mô hình chính thức và cung cấp sự hiểu biết tốt hơn về các biến của tập dữ liệu và mối quan hệ giữa chúng. Nó cũng có thể giúp xác định xem các kỹ thuật thống kê bạn đang xem xét để phân tích dữ liệu có phù hợp hay không. Được phát triển đầu tiên bởi nhà toán học người Mỹ John Tukey vào những năm 1970, kỹ thuật EDA tiếp tục là phương pháp được sử dụng rộng rãi trong quá trình khám phá dữ liệu ngày nay.

**Mục đích của EDA**

Các nhà khoa học dữ liệu có thể sử dụng phân tích thăm dò để đảm bảo kết quả họ tạo ra là hợp lệ và có thể áp dụng cho mọi kết quả và mục tiêu kinh doanh mong muốn. EDA cũng giúp đỡ các bên liên quan bằng cách xác nhận rằng họ đang đặt câu hỏi phù hợp. EDA có thể giúp trả lời các câu hỏi về độ lệch chuẩn, các biến phân loại và khoảng tin cậy. Sau khi EDA hoàn tất và rút ra được thông tin chi tiết, các tính năng của nó có thể được sử dụng để phân tích hoặc lập mô hình dữ liệu phức tạp hơn, bao gồm cả học máy

**Các loại phân tích dữ liệu thăm dò**

Có bốn loại EDA chính:

**Đơn biến phi đồ họa.**Đây là hình thức phân tích dữ liệu đơn giản nhất, trong đó dữ liệu được phân tích chỉ bao gồm một biến. Mục đích chính của phân tích đơn biến là mô tả dữ liệu và tìm ra các mẫu tồn tại trong đó.

**Đồ họa đơn biến.** Các phương pháp phi đồ họa không cung cấp bức tranh đầy đủ về dữ liệu. Do đó cần phải sử dụng các phương pháp đồ họa. Các loại đồ họa đơn biến phổ biến bao gồm:

* Biểu đồ thân và lá, hiển thị tất cả các giá trị dữ liệu và hình dạng của phân bố.
* Biểu đồ, một biểu đồ thanh trong đó mỗi thanh biểu thị tần số (số lượng) hoặc tỷ lệ (số lượng/tổng ​​số) các trường hợp trong một phạm vi giá trị.
* Các sơ đồ hình hộp, mô tả bằng đồ họa bản tóm tắt gồm năm số tối thiểu, tứ phân vị thứ nhất, trung vị, tứ phân vị thứ ba và tối đa.

Đa biến phi đồ họa: Dữ liệu đa biến phát sinh từ nhiều hơn một biến. Các kỹ thuật EDA phi đồ họa đa biến thường cho thấy mối quan hệ giữa hai hoặc nhiều biến của dữ liệu thông qua bảng chéo hoặc thống kê.

Đồ họa đa biến: Dữ liệu đa biến sử dụng đồ họa để hiển thị mối quan hệ giữa hai hoặc nhiều bộ dữ liệu. Đồ họa được sử dụng nhiều nhất là biểu đồ thanh hoặc biểu đồ thanh được nhóm lại với mỗi nhóm đại diện cho một cấp độ của một trong các biến và mỗi thanh trong một nhóm đại diện cho cấp độ của biến kia.

**Một số công cụ khoa học dữ liệu phổ biến nhất được sử dụng để tạo EDA bao gồm:**

**Python:** Ngôn ngữ lập trình hướng đối tượng, thông dịch với ngữ nghĩa động. Python và EDA có thể được sử dụng cùng nhau để xác định các giá trị bị thiếu trong tập dữ liệu. Điều này rất quan trọng để bạn có thể quyết định cách xử lý các giá trị bị thiếu cho machine learning.

R: Ngôn ngữ R được các nhà thống kê trong khoa học dữ liệu sử dụng rộng rãi trong việc phát triển các quan sát thống kê và phân tích dữ liệu.

1. **Thông tin bối cảnh**
   1. **Trên thế giới**

(Rusland và alt, 2017) đã thực hiện đề tài “Analysis Of Naive Bayes Algorithm For Email Spam Filtering Across Multiple Datasets” sử dụng thuật toán Naive 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 Naive 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 Naive 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 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 Naive Bayes hoạt động tốt hơn trên dữ liệu SpamBase hơn là dữ liệu Spamdata.

(Yuksel và alt, 2017) đã thực hiện đề tài “Design of a Machine Learning Based Predictive Analytics System for Spam Problem “ sử dụng thuật toán SVM và cây quyết định để lọc thư rác. Các cây quyết định được dùng để khai phá dữ liệu và SVM được sử dụng làm mô hình học tập có giám sát để phân tích dữ liệu nhằm phân loại thư rác. Dữ liệu đầu tiên được chia thành hai phần, để huấn luyện và để thử nghiệm, sau đó thuật toán được huấn luyện và được đánh giá thông qua nền tảng Microsoft Azure. Nền tảng này cung cấp các công cụ để học máy và so sánh kết quả với cây quyết định và SVM. Kết quả của phương pháp SVM là 97,6% và cây quyết định là 82,6%. Kết quả cho thấy rằng, trình phân loại SVM hoạt động tốt hơn cây quyết định.

(Hanif Bhuiyan và alt, 2018) thực hiện đề tài “A Survey Of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques” Bài viết này minh họa một cuộc khảo sát về các hệ thống lọc thư rác email hiện có khác nhau liên quan đến Kỹ thuật học máy (MLT) như Naive Bayes, SVM, K-Nearest Neighbor, Bayes Additive Regression, KNN Tree và các quy tắc. Tuy nhiên, ở đây chúng tôi trình bày việc phân loại, đánh giá và so sánh các hệ thống lọc thư rác khác nhau và tóm tắt kịch bản tổng thể về tỷ lệ chính xác của các phương pháp hiện có khác nhau.

* 1. **Tại Việt Nam**

(Quan Dang Dinh và alt, 2014) 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 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) 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).

**CHƯƠNG 3: NỘI DUNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU**

1. **Vật liệu nghiên cứu**
   1. **Dữ liệu**

Bộ dữ liệu được sử dụng trong nghiên cứu được lấy từ web kaggle.com với tên Spam Mails Dataset.

Dữ liệu được cập nhật vào năm 2018 và tác giả của bộ dữ liệu là Venkatesh Garnepudi.

Bộ dữ liệu mô tả có 5171 hàng và 4 cột:

* Cột 1: cột tên trống chứa các số được sắp xếp để dữ liệu tệp là duy nhất.
* Cột 2: tabel là nhãn đánh dấu các ham-spam.
* Cột 3: text là ký tự chuỗi.
* Cột 4: tabel\_num là cột mã hóa giá trị của ham thành 0 và spam thành 1.
  1. **Phần mềm**

**Nghiên cứu sử dụng các phần mềm:**

* Jupyter Notebook là một ứng dụng soạn thảo sổ ghi chép, dưới sự bảo trợ của Project Jupyter. Được xây dựng dựa trên sức mạnh của định dạng sổ ghi chép tính toán, Jupyter Notebook cung cấp các cách thức mới, nhanh chóng và tương tác để tạo nguyên mẫu và giải thích mã, khám phá và trực quan hóa dữ liệu cũng như chia sẻ ý tưởng với người khác. (Jupyter Team, 2015)
* Notebook mở rộng cách tiếp cận dựa trên bảng điều khiển tới điện toán tương tác theo hướng mới về chất lượng, cung cấp ứng dụng dựa trên web phù hợp để nắm bắt toàn bộ quá trình tính toán: phát triển, ghi lại và thực thi mã cũng như truyền đạt kết quả. (Jupyter Team, 2015)
* Python 3.11: Python 3.11.1 là bản phát hành chính mới nhất của ngôn ngữ lập trình Python và nó chứa nhiều tính năng và tối ưu hóa mới. (PFS, 2022)

1. **Phương pháp nghiên cứu**
   1. **Phạm vi nghiên cứu**

Nghiên cứu được thực hiện ở phạm vi nhỏ, nhằm đáp ứng nhu cầu môn học.

* 1. **Thời gian thực hiện**

Nghiên cứu được bắt dầu vào vào tháng 9 và kết thúc vào tháng 12 năm 2023.

* 1. **Kết quả kỳ vọng**

Áp dụng các thuật toán đểphân loại được đâu là thư rác và đâu không phải thư rác.

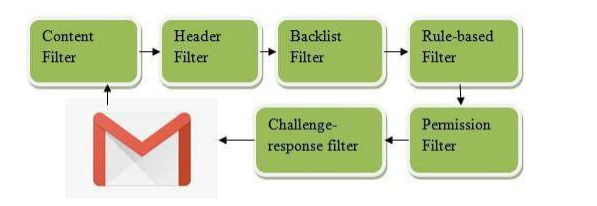
Tự động loại bỏ các loại thư nghi ngờ dính kèm mã độc, quảng cáo, lừa đảo…

Tiết kiệm thời gian cho người dùng thao tác tìm kiếm, không bị bỏ lỡ các thư quan trọng.

* 1. **Cách tiếp cận**

Có nhiều hướng tiếp cận phân loại thư rác, trong đó quy trình lọc thư rác tiêu chuẩn bao gồm những thành phần sau:

* + Lọc nội dung (content filter): phân loại thư rác dựa trên nội dung thư bằng các phương pháp như: thống kê truyền thống, học máy....
  + Lọc đầu thư (header filter): trích xuất thông tin từ đầu thư để lọc.
  + Lọc theo danh sách đen (blacklist filter): xác định thư rác và địa chỉ thư gửi thư rác từ danh sách đen.
  + Lọc dựa trên quy luật (rule-based filter): nhận diện người gửi cụ thể qua đầu thư sử dụng các tiêu chí do người dùng đặt ra.
  + Lọc bằng cấp quyền (permission filter): gửi thư bằng cách có sự đồng ý của người nhận từ trước.
  + Lọc nhờ xác thực bằng cách trả lời (challenge-response filter): kiểm tra thư được gửi tự động hay do người trực tiếp gửi (thông thường thư rác đều được tạo và gửi tự động thông qua phần mềm) qua việc người gửi trả lời để xác thực.



**Hình 3:** Chuẩn quá trình lọc thư rác.

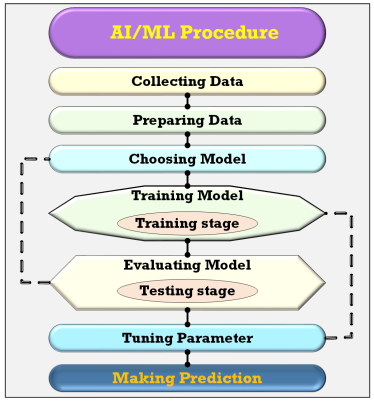
Quá trình lọc thư rác qua email hoạt động thông qua một tập hợp các giao thức để xác định thông điệp là thư rác hay không. Hiện nay có rất nhiều thư rác quá trình lọc đã tồn tại. Trong đó, chuẩn quá trình lọc thư rác tuân theo một số quy tắc và hoạt động như một phân loại với các bộ giao thức. Hình.1 cho thấy rằng, một quy trình lọc thư rác tiêu chuẩn đã thực hiện phân tích bằng cách làm theo một số bước (Christina, V., Karpagavalli, S., & Suganya, G, 2010).

Đầu tiên là bộ lọc nội dung xác định thư rác bằng cách áp dụng một số Kỹ thuật học máy như Naïve Bayes, KNN, SVM, Decision tree,….

Thứ hai, bộ lọc tiêu đề hoạt động bằng cách trích xuất thông tin từ email tiêu đề. Sau đó, bộ lọc danh sách ngược sẽ xác định thư rác nhắn tin và dừng tất cả các email đến từ danh sách ngược tài liệu. Sau đó, “Bộ lọc dựa trên quy tắc” nhận dạng người gửi thông qua dòng chủ đề bằng cách sử dụng tiêu chí do người dùng xác định (Christina, V., Karpagavalli, S., & Suganya, G, 2010).

Tiếp theo, “Bộ lọc quyền” gửi tin nhắn bằng cách nhận sự chấp thuận trước của người nhận. Cuối cùng, “Bộ lọc phản hồi thách thức” được thực hiện bằng cách áp dụng thuật toán cho nhận được sự cho phép từ người gửi để gửi thư.

* 1. **Quy trình thực hiện**



**Hình 4:** Thủ tục thực hiện tổng quát.

Bước 1: Chọn dữ liệu (Collecting data)

Gọi các thư viện cần thiết ( Import required libraries)

Bước 2: Chuẩn bị dữ liệu (Preparing data)

Mở tệp chứa dữ liệu spam\_ham\_dataset.csv ( load file )

Bước 3: Chọn mô hình (Choosing model)

Sử dụng các phương pháp Máy học (Machine Leaning):

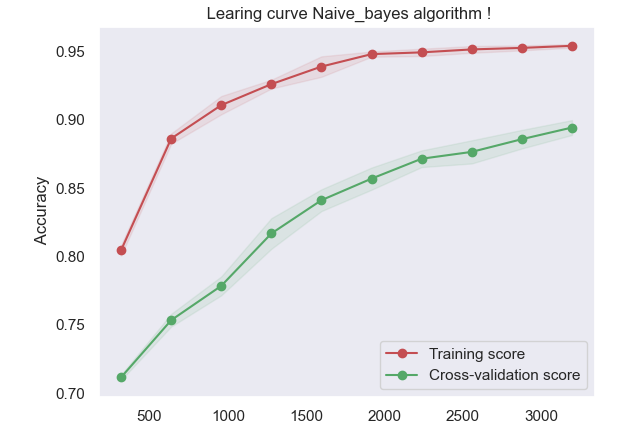
• EDA.

• Naive Bayes.

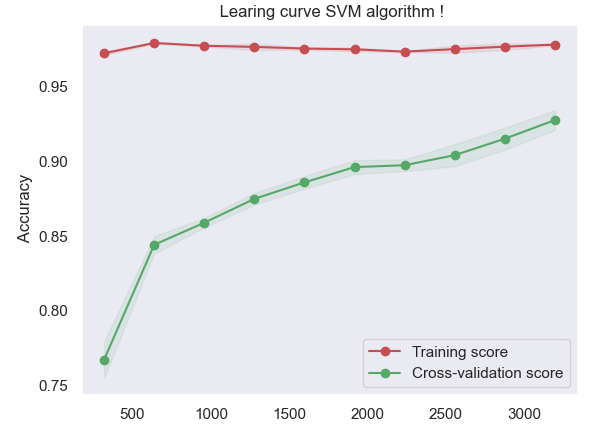
• Support Vector Machine.

Bước 4: Huấn luyện mô hình (Training model)

Sử dụng 80% bộ dữ liệu ban đầu để huấn luyện và 20% để kiểm tra.



**Hình 5:** Kết quả học của thuật toán Naïve Bayes.

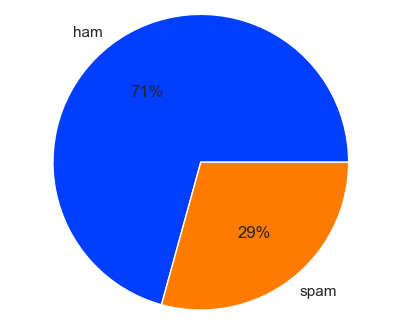


**Hình 6:** Kết quả học của thuật toán SVM

Nhận xét: Kết quả học của hai thuật toán Naive Bayes và SVM không bị overfitting vì hai đường biểu diễn Training score và Cross-validation score đều tăng.

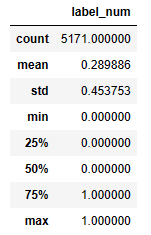
Bước 5: Đánh giá mô hình (Evaluation model)

EDA: Một số thống kê cơ bản.



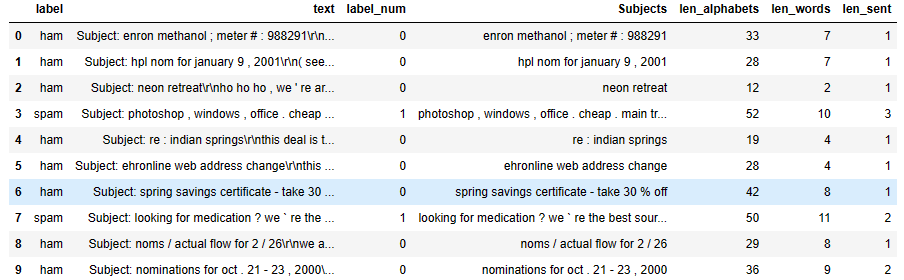
**Hình 7:** Tỉ lệ spam và ham trong bộ dữ liệu.

Một số thống kê cột label\_num:



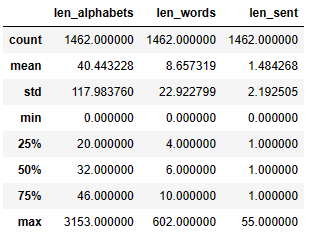
**Hình 8:** Một số thống kê cơ bản của label\_num.

Tạo các cột đo độ dài của cột Subjects: Tạo cột xử lý ngôn ngữ tự nhiên thống NLTK đặt tên ba cột là len\_alphabets, len\_words, len\_sent dùng để thống kê ngôn ngữ tự nhiên.



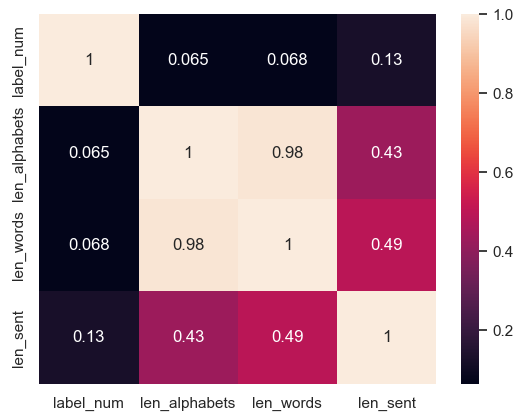
**Hình 9:** Bảng thể hiện thêm các cột độ dài vừa tạo.

Kết quả thống kê xử lý ngôn ngữ tự nhiên.



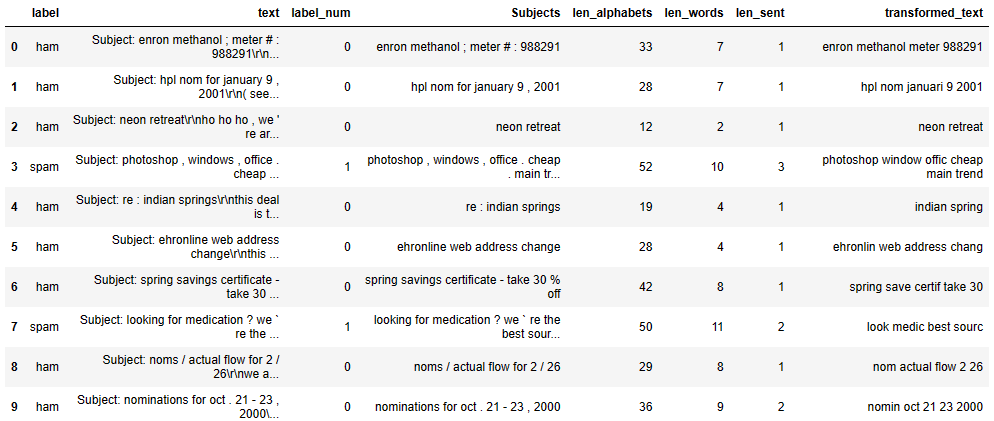
**Hình 10:** Bảng kết quả thống kê ngôn ngữ tự nhiên của cột Subjects.

Biểu diễn mức độ tương quan của kết quả thống kê ngôn ngữ tự nhiên của cột Subjects.



**Hình 11:** Bảng mức độ tương quan của kết quả thống kê ngôn ngữ tự nhiên

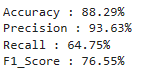
Tạo cột transformed\_text chứa giá trị của cột Subjects sau khi loại bỏ các ký tự đặc biệt.



**Hình 12:** Tạo cột transformed\_text.

NB:

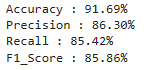
Ma trận đánh giá của thuật toán NB:



**Hình 13:** ma trận đánh giá của NB.

SVM:

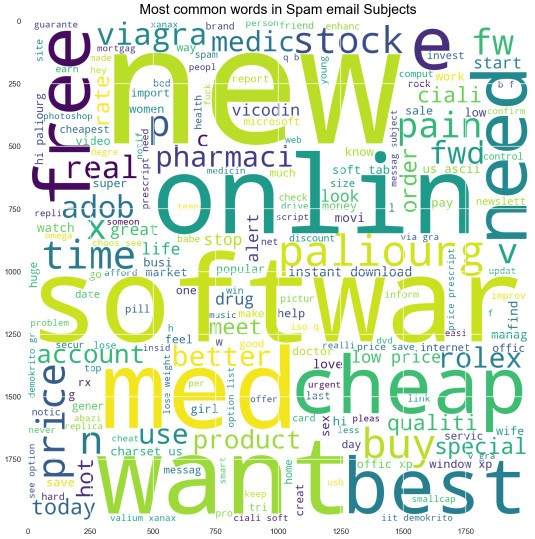
Ma trận đánh giá của thuật toán SVM:



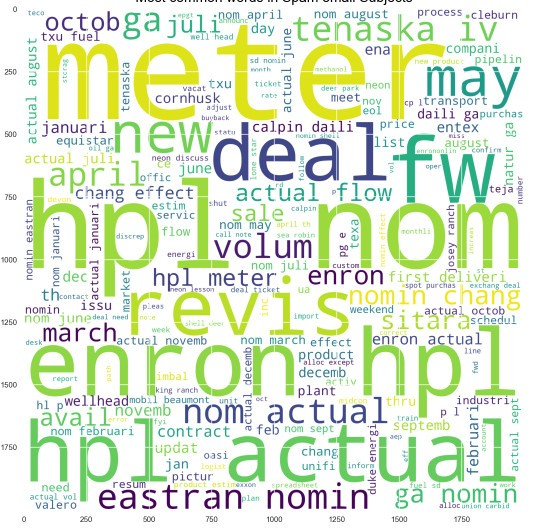
**Hình 14:** Ma trận đánh giá của SVM.

Bước 6: Quyết định (Making desicion)

EDA: các từ thường được xem là spam và ham trong bộ dữ liệu.



**Hình 15:** Các ký tự thường được xem là spam trong bộ dữ liệu.



**Hình 16:** Các ký tự thường được xem là ham trong bộ dữ liệu.

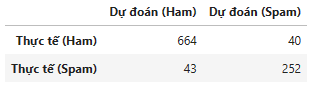
**NB:**



**Hình 17:** Mức độ nhầm lẫn của NB.

Nhận xét: Dự đoán thực tế của ham bị nhầm lẫn là 104. Còn mức độ nhầm lẫn của spam là 13.

SVM:



**Hình 18:** Dự đoán mức độ nhầm của SVM.

Nhận xét: Dự đoán thực tế của ham bị nhầm lẫn là 43. Còn mức độ nhầm lẫn của spam là 40.

1. **Ưu điểm và nhược điểm của phương pháp nghiên cứu**
   1. **Ưu điểm**

Việc sử dụng thuật toán của ML giúp tiết kiệm thời gian, công sức và tăng cường hiệu quả giúp đưa ra phân loại tốt hơn, cải thiện hiệu suất giảm chi phí.

* 1. **Nhược điểm**

Yêu cầu dữ liệu lớn: ML cần một lượng lớn dữ liệu để học hỏi và cải thiện. Điều này có thể là một thách thức đối với các ứng dụng ML mới, nơi dữ liệu có thể không có sẵn.

Dễ bị nhầm lẫn: ML có thể bị thiên vị nếu dữ liệu đào tạo không đại diện cho thế giới thực. Điều này có thể dẫn đến các mô hình ML đưa ra các quyết định không chính xác hoặc không công bằng.

Có thể khó giải thích: Các mô hình ML thường rất phức tạp và có thể khó giải thích lý do tại sao chúng đưa ra các quyết định nhất định. Điều này có thể khiến người dùng khó tin tưởng vào các mô hình ML.

# **CHƯƠNG 4: KẾT QUẢ VÀ THẢO LUẬN**

1. **Kết quả**

Trong phần này, chúng tôi trình bày các kết quả chính thu được từ nghiên cứu về lập trình bộ lọc thư rác trong email. Các kết quả này được trình bày dưới đây:

*Đánh giá hiệu suất*

Đánh giá mô hình: Chúng tôi đã đánh giá hiệu suất của mô hình lọc thư rác thông qua các phép đo như độ chính xác, độ tương quan, độ lệch chuẩn và tỷ lệ chêch lệch. Kết quả cho thấy 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.

*Đánh giá ứng dụng trong thực tế*

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.

1. **Thảo luận**

Trong phần này, chúng tôi tiến hành thảo luận về các kết quả thu được từ nghiên cứu và đưa ra các nhận xét, phân tích và suy luận. Các điểm quan trọng trong phần thảo luận bao gồm:

*Hiệu suất của mô hình*

Độ chính xác và độ tương quan: Kết quả cho thấy mô hình của chúng tôi đạt được độ chính xác cao và độ tương quan tốt 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.

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 trongviệc 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.

*So sánh với các phương pháp khác*

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.

Ưu điểm và hạn chế so với các phương pháp khác: Chúng tôi cũng thảo luận về ưu điểm và hạn chế của mô hình so với các phương pháp khác. Ví dụ, mô hình dựa trên trí tuệ nhân tạo có khả năng tự động học và cải thiện theo thời gian, trong khi các phương pháp truyền thống có thể đòi hỏi cập nhật thủ công và dễ bị lạc hậu.

*Ứng dụng thực tế và tiềm năng phát triển*

Đóng góp và ứng dụng trong thực tế: Chúng tôi thảo luận về đóng góp của mô hình lọc thư rác trong việc giảm thiểu tiếp xúc với thư rác và tăng cường an ninh cho hộp thư điện tử. Chúng tôi cũng đề cập đến tiềm năng ứng dụng của mô hình trong các lĩnh vực khác như phòng chống spam trong tin nhắn ngắn hoặc ứng dụng cho những ngôn ngữ khác nhau.

Hướng phát triển tiếp theo: Cuối cùng, chúng tôi đề xuất các hướng phát triển tiếp theo cho nghiên cứu này. Ví dụ, có thể nâng cao mô hình bằng cách sử dụng các phương pháp học sâu hoặc kết hợp nhiều mô hình để cải thiện hiệu suất và độ tin cậy.

**CHƯƠNG 5: KẾT LUẬN VÀ ĐỀ NGHỊ**

1. **Kết luận**

Trong chương 3, chúng tôi đã trình bày chi tiết về các phương pháp và kỹ thuật đã được sử dụng trong nghiên cứu này. Kết quả cho thấy các phương pháp này đã đạt được hiệu quả và đáng chú ý trong việc giải quyết vấn đề cụ thể mà nghiên cứu đề cập đến. Các 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. Phương pháp 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.

1. **Đề 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 dụngcủa 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**

* "Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No. 04EX929)," Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology, 2004., Rome, Italy, 2004, pp. , doi: 10.1109/ISSPIT.2004.1433664.
* Emmanuel Gbenga Dada et al., Machine learning for email spam filtering: review, approaches and open research problems,Heliyon, Volume 5, Issue 6, 2019, e01802, ISSN 2405-8440, <https://doi.org/10.1016/j.heliyon.2019.e01802>.
* Radicati Group. "Email Statistics Report, 2020-2024". Địa chỉ: <https://www.radicati.com/wp/wp-content/uploads/2020/01/Email-Statistics-Report-2020-2024-Executive-Summary.pdf>. [Truy cập ngày 13/11/2023]
* Deitel. "Java How to Program, Early Objects, 11th Edition"
* Nazirova, Saadat. (2011). Survey on Spam Filtering Techniques. Communications and Network. 3. 153-160. 10.4236/cn.2011.33019.
* Baeza-Yates, Ricardo & Ribeiro-neto, Berthier & Mills, Don & Bonn, Ontario & Juan, San & Mexico, Milan & Taipei, City & Wesley, Addison & Limited, Longman. (1999). Modern Information Retrieval
* Hanif Bhuiyan và alt. (2018, 01). A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques. Được truy lục từ A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques
* machinelearningcoban. (không ngày tháng). Được truy lục từ https://machinelearningcoban.com/2017/08/08/nbc/
* Quan Dang Dinh và alt. (2014, 10). Automated generation of ham rules for Vietnamese spam filtering. doi:10.1109/CISDA.2014.7035628
* Rusland và alt. (2017, 08). Analysis of Naive Bayes Algorithm for Email Spam Filtering across Multiple Datasets. doi:10.1088/1757-899X/226/1/012091
* vu hoang và alt. (2007, 01). Topic-Based Vietnamese News Document Filtering in the BioCaster Project. doi:10.1109/ALPIT.2007.56
* What is exploratory data analysis? (không ngày tháng). Được truy lục từ https://www.ibm.com/topics/exploratory-data-analysis
* Yuksel và alt. (2017, 09). Design of a Machine Learning Based Predictive Analytics System for Spam Problem. doi:10.12693/APhysPolA.132.500