TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

Báo cáo đề tài

Áp dụng máy học vào phân loại thư điện tử

HỌC PHẦN: PHƯƠNG PHÁP NGHIÊN CỨU KHOA HỌC

THUỘC NHÓM NGÀNH: CÔNG NGHỆ THÔNG TIN

GIẢNG VIÊN HƯỚNG DẪN: TS. ĐỖ NHƯ TÀI

THÀNH VIÊN NHÓM 77:

3123410024 - NGÔ THƯỢNG BẢO

3123410242 - NGUYỄN TRỌNG NHÂN

3123410174 - NGUYỄN PHẠM TUẤN KHÔI

3123410387 - NGUYỄN HỮU TRI

Ngày nộp báo cáo: 20/05/2025

Mục lục

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI 6
1.1. Đặt vấn đề 6
1.2. Lý do chọn đề tài 7
1.3. Mục tiêu nghiên cứu 7
1.4. Đối tượng và phạm vi nghiên cứu 8
1.4.1. Đối tượng nghiên cứu 8
1.4.2. Phạm vi nghiên cứu 8
CHƯƠNG 2. LƯỢC KHẢO TÀI LIỆU 9
2.1. Các tài liệu liên quan 9
2.2. Tập dữ liệu 9
2.3. Cơ sở lý thuyết 10
2.3.1. Natural Language Processing 10
2.3.2. Text Vectorization 10
2.3.3. Naïve Bayes 11
2.3.4. Support Vector Machine 12
2.3.5. K nearest neighbor 13
2.3.6. Neural Network 13
CHƯƠNG 3. PHƯƠNG PHÁP NGHIÊN CỨU 15
3.1. Thiết kế nghiên cứu: 15
3.2. Đối tượng và mẫu nghiên cứu: 15
3.3. Cách thu thập dữ liệu: 16
3.4. Phân tích dữ liệu: 17
CHƯƠNG 4. THỰC NGHIỆM VÀ THẢO LUẬN 18
4.1. Môi trường thực nghiệm 18
4.2. Tải dữ liệu vào mô hình 19
4.3. Phân tích dữ liệu 19
4.3.1. Phân bổ số lượng thư rác và thư hợp lệ 19
4.3.2. Những từ xuất hiện nhiều nhất trong ham và spam 21

4.4. Huấn luyện mô hình 22
4.5. Đánh giá và so sánh kết quả 23
CHƯƠNG 5. KẾT LUẬN 25
5.1. Tổng kết 25
5.2. Hướng nghiên cứu tiếp theo 26
TÀI LIỆU THAM KHẢO 26

Lời cảm ơn

Trước hết chúng tôi xin cảm ơn thầy Đỗ Như Tài – người đã tận tình hướng dẫn và chia sẻ những kiến thức quý báu giúp chúng tôi hoàn thiện đề tài "Áp dụng máy học vào phân loại thư rác điện tử" một cách hiệu quả và khoa học

Xin cảm ơn các thành viên trong nhóm, bạn bè đã cùng nhau góp ý, phát triển và hỗ trợ nhau trong quá trình nghiên cứu và hoàn thành báo cáo. Mong rằng đề tài sẽ góp phần nhỏ bé vào việc phát triển công nghệ bảo mật thông tin và nâng cao nhận thức cộng đồng về vấn đề thư rác và lừa đảo trong thời đại số.

Xin chân thành cảm ơn!

Tóm tắt

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

Đề tài "Áp dụng máy học vào phân loại thư rác điện tử" được thực hiện nhằm xây dựng một hệ thống thông minh, ứng dụng công nghệ học máy (Machine Learning) để phân loại và phát hiện email spam hiệu quả. Nghiên cứu sử dụng các thuật toán học có giám sát như Naïve Bayes, SVM, Decision Tree và K-NN, kết hợp với kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) để trích xuất đặc trưng từ nội dung và siêu dữ liệu email.

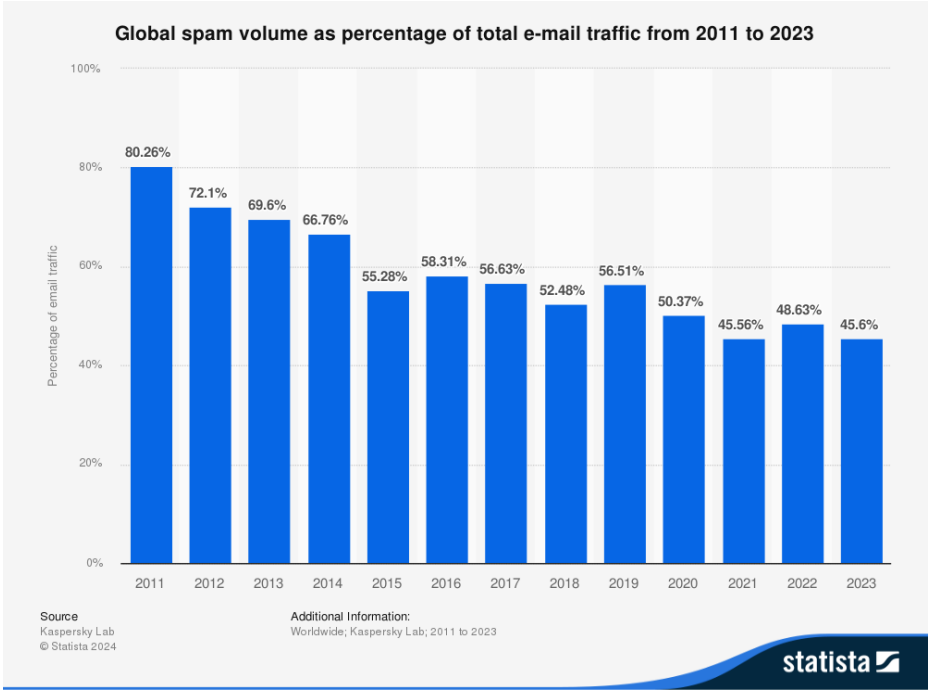
Hệ thống được đánh giá qua các chỉ số độ chính xác, độ nhạy và khả năng phát hiện thời gian thực. Kết quả cho thấy mô hình có khả năng phát hiện chính xác các email gian lận, thích ứng với các chiến thuật spam thay đổi liên tục, góp phần nâng cao hiệu quả phòng chống thư rác và bảo vệ an toàn thông tin trong môi trường số hiện đại.

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI

1.1. Đặt vấn đề

Trong thời đại công nghệ số bùng nổ như hiện nay, email đã trở thành một công cụ giao tiếp phổ biến và không thể thiếu trong mọi lĩnh vực của đời sống xã hội – từ học thuật, công việc hành chính, thương mại điện tử, đến trao đổi cá nhân. Tuy nhiên thì đi cùng với phương tiện liên lạc tiện lợi và hiện dại này là mối quan ngại về lừa đảo, bảo mật và an toàn thông tin cá nhân và doanh nghiệp mang tên thư rác, hay còn được gọi là spam.

Theo một thống kê được thực hiện bởi Kasper Lab vào năm 2023 và được công bố vào tháng 3 năm 2024 bởi Securelist [1], 45,60% của tổng lượng email được gửi đi trên thế giới là thư rác, với 31,45% lượng thư rác được gửi từ Nga, các thư rác này không chỉ chiếm lấy phần lớn dung lượng dự trữ, tốn thời gian xử lý mà còn làm gián đoạn hoạt động giao tiếp, gây ra các mối đe dọa về an ninh mạng dưới dạng email trúng thưởng, hoàn tiền, lừa đảo liên quan tới tiền điện tử,…



Hình 1.1. Thống kê thư rác trên toàn cầu từ năm 2011 tới năm 2023

1.2. Lý do chọn đề tài

Để đối phó với vấn đề thư rác điện tử, các hệ thống phân loại thư từ đã được nghiên cứu và áp dụng vào thực tiễn trong quá khứ để giải quyết vấn đề trên với đa phần bằng phương pháp thủ công hoặc quy tắc tĩnh (rule-based) như lọc theo từ ngữ, địa chỉ người gửi, blacklist, whitelist và dựa vào phản hồi từ cộng đồng để phân biệt, tuy nhiên thì các phương pháp này có phần lỗi thời khi cùng với sự phát triển của công nghệ, các thủ đoạn lan truyền thư rác cũng tinh vi hơn, các biện pháp nói trên trở nên kém hiệu quả và có phần không chính xác.

Cùng chính vì vậy mà nhóm quyết định chọn đề tài "Áp dụng máy học vào phân loại thư điện tử" để nghiên cứu, nhằm tìm ra được giải pháp tốt nhất để giải quyết vấn đề phân loại thư rác điện tử.

1.3. Mục tiêu nghiên cứu

Nghiên cứu hướng đến việc ứng dụng công nghệ máy học để phát hiện và phân loại email, tin nhắn gian lận (spam/lừa đảo) một cách hiệu quả và chính xác. Hệ thống này không chỉ giúp người dùng tránh được các rủi ro về bảo mật thông tin mà còn góp phần giảm thiểu thời gian, công sức lọc thủ công và tăng cường an toàn trong môi trường mạng.

- Xây dựng mô hình phát hiện gian lận: Phát triển mô hình máy học có khả năng phân loại chính xác giữa email hợp lệ và email gian lận và tối ưu để đạt được hiệu suất cao về độ chính xác, độ nhạy và độ đặc hiệu, giảm thiểu lỗi phân loại.
- Lựa chọn và trích xuất đặc trưng: Áp dụng kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) để trích xuất thông tin từ nội dung email cùng với đó là sụ kết hợp với siêu dữ liệu để xây dựng tập đặc trưng có giá trị cho mô hình.

- So sánh và lựa chọn thuật toán tối ưu: Thử nghiệm và đánh giá nhiều thuật toán máy học (Naïve Bayes, SVM, Decision tree, K-nearest neighbor). Cùng với đó là đề xuất ra phương pháp cải tiến nhằm cải thiện hiệu năng, độ chính xác của mô hình

1.4. Đối tượng và phạm vi nghiên cứu

1.4.1. Đối tượng nghiên cứu

Nghiên cứu sẽ tập trung vào các đối tượng như sau:

- Dữ liệu văn bản đầu vào: Sử dụng bộ dữ liệu công khai như Email Spam Classification Dataset CSV[2] từ đó thu thập thêm dữ liệu thực tế từ các hệ thống email với sự đồng ý của người dùng, đảm bảo tuân thủ các quy định bảo mật và quyền riêng tư.
- Kỹ thuật xử lý ngôn ngữ tự nhiên: Nghiên cứu về các phương pháp như Text Vectorization, Tokenization nhằm trích xuất đặc trưng từ email.
- Thuật toán máy học: Tập trung vào các phương pháp học có giám sát, từ truyền thống đến hiện đại như: Naïve Bayes, SVM, Decision tree, K-nearest neighbor,…

1.4.2. Phạm vi nghiên cứu

Nghiên cứu có phạm vi nội dung như sau:

- Phân loại email/văn bản: Nghiên cứu tập trung vào việc xây dựng và đánh giá mô hình máy học vào việc phân loại giữ thư hợp lệ và thư rác điện tử (spam) dựa vào các đặc trưng được trích xuất.
- Xử lý dữ liệu: Làm sạch, chuẩn hóa và trích xuất đặc trưng từ email, giải quyết vấn đề mất cân bằng dữ liệu giữa các lớp (spam và hợp lệ), từ đó nghiên cứu bố cục thường gặp trong email, tin nhắn, hoặc nội dung trực tuyến chứa gian lận: tiêu đề gây chú ý, phần kêu gọi hành động, đường dẫn giả mạo,...

- Thuật toán: So sánh và đánh giá hiệu suất của các thuật toán học máy (Naïve Bayes, SVM, Decision Tree, K-Nearest Neighbors) và mạng nơ-ron học sâu.
- Sử dụng ngôn ngữ python để triển khai mô hình, sử dụng mô hình học sâu TensorFlow.

CHƯƠNG 2. LƯỢC KHẢO TÀI LIỆU

2.1. Các tài liệu liên quan

Vấn đề phân loại thư điện tử không phải là một vấn đề khó để bắt gặp, kể cả trong quá khứ. Cùng vì đó mà đã có các bài nghiên cứu đề ra giải pháp cho vấn đề này, ví dụ như K. Agarwal, T. Kumar. [3] ứng dụng thuật toán Naïve Bayes vào phân loại thư điện tử hay Harisinghaney et al. (2014) [4] với thuật toán K-nearest neighbor. Tuy nhiên thì các bài báo trước chỉ áp dụng các thuật toán máy học thông thường mà chưa mở rộng sang mô hình học sâu.

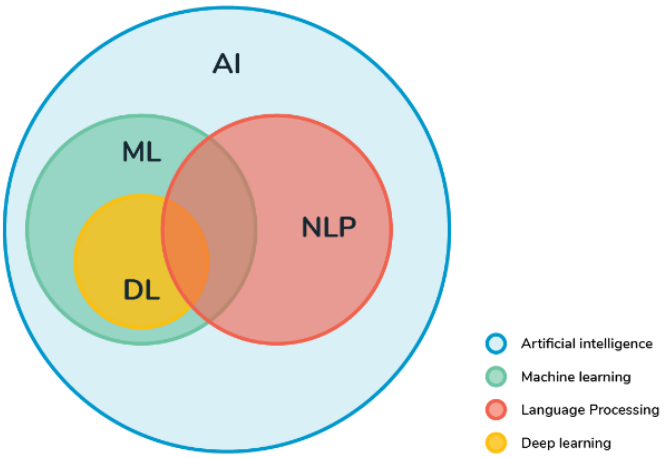
2.2. Tập dữ liệu

Nghiên cứu sử dụng tập tài liệu được tham khảo từ trang web thứ 3 mang tên Kaggle, tập tài liệu này sẽ được sử dụng để huấn luyện mô hình phân loại thư điện tử. Tập dữ liệu “emails.csv” bao gồm thông tin của 5172 emails được chọn ngẫu nhiên và được đánh dấu phân loại thư hợp lệ hoặc thư rác điện tử, với 5172 dòng là mỗi email tương ứng, 3002 cột với cột đầu tiên là tên email, cột cuối cùng là đánh dấu phân loại và 3000 cột còn lại bao gồm các từ xuất hiện nhiều nhất trong các emails. Cùng với đó là một tập dữ liệu tự tạo để kiểm thử mô hình và đánh giá kết quả.

2.3. Cơ sở lý thuyết

2.3.1. Natural Language Processing

Xử lý ngôn ngữ tự nhiên (NLP) là một lĩnh vực thuộc ngành trí tuệ nhân tạo (AI) liên quan đến việc cho phép máy tính hiểu, phân tích, diễn giải và tạo ra ngôn ngữ của con người. NLP kết hợp kiến thức từ khoa học máy tính, ngôn ngữ học và thống kê để xây dựng các mô hình và thuật toán có khả năng xử lý văn bản và giọng nói. Các nhiệm vụ phổ biến trong NLP bao gồm phân tích cú pháp, phân tích ngữ nghĩa, nhận dạng thực thể có tên, dịch máy và tóm tắt văn bản.



Hình 2.1. Sơ đồ biểu hiệu mối liên hệ giữa NLP và AI

2.3.2. Text Vectorization

Biểu diễn văn bản thành vector (text vectorization) là quá trình chuyển đổi văn bản thành các vector số, cho phép máy tính hiểu và xử lý thông tin văn bản. Các phương pháp phổ biến bao gồm:

- Bag of Words (BoW): Đếm tần suất xuất hiện của mỗi từ trong văn bản.

- TF-IDF (Term Frequency-Inverse Document Frequency): Tính trọng số của mỗi từ dựa trên tần suất xuất hiện trong văn bản và độ phổ biến của từ đó trong toàn bộ tập văn bản.
- Word Embeddings (Word2Vec, GloVe, FastText): Biểu diễn mỗi từ thành một vector trong không gian nhiều chiều, thể hiện mối quan hệ ngữ nghĩa giữa các từ.

Review 1: This movie is very scary and long

Review 2: This movie is not scary and is slow

Review 3: This movie is spooky and good

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

Hình 2.3.2. Bag of Words

2.3.3. Naïve Bayes

Thuật toán phân loại Naive Bayes được sử dụng từ năm 1998 để nhận diện thư rác. Đây là một thuật toán học có giám sát. Bộ phân loại Bayesian hoạt động dựa trên các sự kiện phụ thuộc và xác suất của sự kiện sẽ xảy ra trong tương lai, có thể được phát hiện từ các sự kiện đã xảy ra trước đó. Naive Bayes được xây dựng dựa trên định lý Bayes, giả định rằng các đặc trưng là độc lập với nhau. Kỹ thuật phân loại Naive Bayes có thể được sử dụng để phân loại email rác, vì xác suất của từ đóng vai trò chính ở đây. Nếu có bất kỳ từ nào xuất hiện thường xuyên trong thư rác nhưng không có trong thư thường, thì đó là thư rác. Thuật toán phân loại Naive Bayes đã trở thành một kỹ thuật tốt nhất để lọc email. Để đạt hiệu quả, mô hình được huấn luyện bằng bộ lọc Naive Bayes rất tốt. Naive Bayes luôn tính toán xác suất

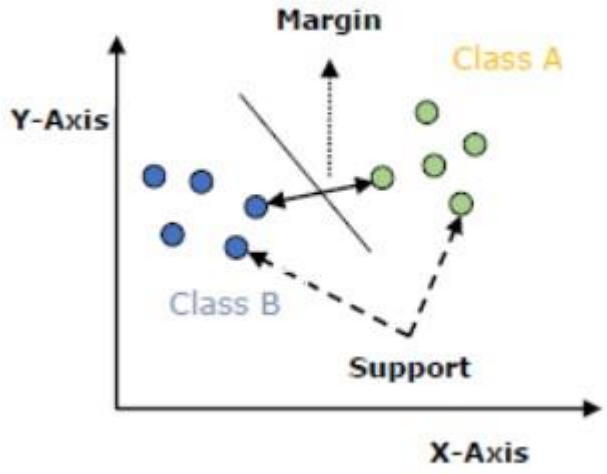
của mỗi lớp và lớp có xác suất lớn nhất sau đó được chọn làm đầu ra. Naive Bayes luôn cung cấp một kết quả chính xác. Nó được sử dụng trong nhiều lĩnh vực như lọc thư rác.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(B) = \sum_y P(B|A)P(A)$$

2.3.4. Support Vector Machine

Máy học vector hỗ trợ (SVM) là một thuật toán học có giám sát phổ biến, mô hình vector hỗ trợ được sử dụng cho các bài toán phân loại trong các kỹ thuật học máy. Các máy vector hỗ trợ hoàn toàn dựa trên ý tưởng về các điểm quyết định. Độ phân giải chính của thuật toán máy vector hỗ trợ là tạo ra đường hoặc ranh giới quyết định. Thuật toán máy vector hỗ trợ cho kết quả là siêu phẳng để phân loại các mẫu mới. Trong không gian 2 chiều, "siêu phẳng là đường thẳng chia mặt phẳng thành 2 phần, mỗi lớp nằm ở một phía.



Hình 2.3.4 Support Vector Machine

2.3.5. K nearest neighbor

K-nearest neighbors là một thuật toán phân loại có giám sát. Thuật toán này có một số điểm dữ liệu và vector dữ liệu được phân tách thành nhiều lớp để dự đoán phân loại của điểm mẫu mới.

K- Nearest neighbor là một thuật toán LAZY, nghĩa là nó chỉ cố gắng ghi nhớ quy trình, nó không tự học. Nó không tự đưa ra quyết định. Thuật toán K- Nearest neighbor phân loại điểm mới dựa trên thước đo độ tương đồng có thể là khoảng cách Euclide.

Thước đo khoảng cách Euclide và xác định ai là hàng xóm của nó.

- Entropy bằng bảng tần số của một thuộc tính:

$$E(S) = \sum_{t=1}^{e} - p_i log_2 p_i$$

- Entropy bằng bảng tần số của hai thuộc tính:

$$E(T, X) = \sum_{t=1}^{e} P(c)E(c)$$

2.3.6. Neural Network

Mạng nơ-ron nhân tạo (Neural Network - NN) là một mô hình tính toán được lấy cảm hứng từ cấu trúc và hoạt động của não bộ con người, đặc biệt là cách các nơ-ron sinh học kết nối và xử lý thông tin. NN là một thành phần quan trọng trong học sâu (Deep Learning), một nhánh của học máy (Machine Learning), và thuộc lĩnh vực trí tuệ nhân tạo (Artificial Intelligence). Mục tiêu chính của NN là mô phỏng khả năng học hỏi từ dữ liệu để thực hiện các nhiệm vụ như phân loại, hồi quy, nhận diện hình ảnh, hay xử lý ngôn ngữ tự nhiên.

Một mạng nơ-ron bao gồm các lớp (layers) nơ-ron được kết nối với nhau:

- Lớp đầu vào (Input Layer): Nhận dữ liệu đầu vào (ví dụ: các đặc trưng như pixel của hình ảnh hoặc từ trong văn bản).
- Lớp ẩn (Hidden Layers): Xử lý thông tin thông qua các phép tính toán. Số lượng lớp ẩn và nơ-ron trong mỗi lớp quyết định độ phức tạp của mô hình.
- Lớp đầu ra (Output Layer): Tạo ra kết quả dự đoán (ví dụ: xác suất một email là spam hay không).

Mỗi nơ-ron trong mạng nhận đầu vào từ các nơ-ron ở lớp trước, áp dụng một trọng số (weight) và bias để tính toán, sau đó truyền qua một hàm kích hoạt (activation function) để tạo đầu ra. Các hàm kích hoạt phổ biến bao gồm ReLU (Rectified Linear Unit), Sigmoid, và Tanh, giúp mô hình học được các mối quan hệ phi tuyến trong dữ liệu.

Quá trình học của NN diễn ra qua hai giai đoạn chính: lan truyền xuôi (forward propagation) và lan truyền ngược (backpropagation).

Trong lan truyền xuôi, dữ liệu đi qua các lớp, được biến đổi thông qua trọng số, bias và hàm kích hoạt, để tạo ra dự đoán. Kết quả này được so sánh với giá trị thực tế bằng một hàm mất mát (loss function), ví dụ: Mean Squared Error cho bài toán hồi quy hoặc Cross-Entropy Loss cho bài toán phân loại.

Trong lan truyền ngược, mô hình sử dụng thuật toán tối ưu hóa (thường là Gradient Descent) để điều chỉnh trọng số và bias, sao cho hàm mất mát giảm dần. Quá trình này lặp lại qua nhiều epoch (vòng lặp huấn luyện) cho đến khi mô hình đạt hiệu suất mong muốn.

CHƯƠNG 3. PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Thiết kế nghiên cứu:

Nghiên cứu được thực hiện theo phương pháp định lượng với thiết kế thực nghiệm, tập trung vào phân tích dữ liệu và đánh giá mô hình bằng các chỉ số thống kê. Cụ thể, nhóm nghiên cứu áp dụng mô hình học máy giám sát để phân loại thư điện tử thành hai nhóm spam (thư rác) và ham (thư bình thường, không phải spam). Quy trình thực hiện bao gồm việc thu thập tập dữ liệu email đã gán nhãn, tiền xử lý và trích xuất đặc trưng từ dữ liệu, sau đó huấn luyện mô hình máy học trên dữ liệu này và đánh giá hiệu quả phân loại. Phương pháp định lượng cho phép định tính mức độ chính xác của mô hình bằng các con số cụ thể (như tỷ lệ phân loại đúng), qua đó đưa ra kết luận về khả năng áp dụng của thuật toán máy học trong bài toán phát hiện thư rác. Với cách tiếp cận thực nghiệm, nghiên cứu tiến hành thử nghiệm trên bộ dữ liệu thực tế để kiểm chứng giả thuyết rằng mô hình học máy có thể nhận diện thư spam tự động với độ chính xác cao. Do không có sự tham gia của đối tượng con người hay các yếu tố định tính, phương pháp định lượng thuần túy được lựa chọn là phù hợp nhất nhằm đáp ứng mục tiêu nghiên cứu.

3.2. Đối tượng và mẫu nghiên cứu:

Đối tượng nghiên cứu trong đề tài này là tập hợp các email và đặc trưng nội dung của chúng phục vụ cho việc phân loại thư rác bằng thuật toán máy học. Nói cụ thể hơn, đối tượng được phân tích là các thư điện tử đã được gán nhãn spam/ham cùng với các đặc trưng từ khóa xuất hiện trong nội dung những email đó. Mỗi email được coi như một đơn vị quan sát trong nghiên cứu, kèm theo thuộc tính đầu ra (nhãn) cho biết đó có phải thư rác hay không. Thông qua các email này, nghiên cứu xem xét đặc điểm phân biệt giữa thư rác và thư thường, từ đó xây dựng mô hình phân loại tự động.

Mẫu nghiên cứu được sử dụng là một tập dữ liệu gồm 5172 email đã được gán nhãn phân loại (spam hoặc không spam). Trong số này có 1500 thư Spam (tương đương ~29% mẫu) và 3672 thư Ham (~71%). Tập dữ liệu được lựa chọn đảm bảo gồm cả hai loại thư với tỷ lệ chênh lệch nhất định, phản ánh hiện tượng thực tế rằng thư rác chiếm một phần nhỏ hơn so với thư hợp lệ trong hộp thư của người dùng. Mẫu 5172 email này đủ lớn để huấn luyện mô hình học máy và cũng đảm bảo tính đa dạng về nội dung email. Các email được thu thập một cách ngẫu nhiên từ nhiều nguồn khác nhau, giúp giảm thiểu thiên lệch do nguồn gốc dữ liệu. Nhờ có sẵn nhãn Spam/Ham cho từng email, mẫu dữ liệu này rất phù hợp cho phương pháp học máy có giám sát, cho phép mô hình học được ranh giới phân biệt giữa thư rác và thư thường. Trong nghiên cứu, mẫu dữ liệu nói trên được coi là đại diện cho bài toán phân loại thư rác nói chung trong môi trường email, làm cơ sở để huấn luyện và kiểm chứng mô hình đề xuất.

3.3. Cách thu thập dữ liệu:

Dữ liệu sử dụng cho nghiên cứu được thu thập từ nguồn dữ liệu mở trực tuyến. Cụ thể, nhóm nghiên cứu đã sử dụng bộ dữ liệu công khai từ trang Kaggle – một nền tảng chia sẻ dữ liệu và bài toán máy học. Bộ dữ liệu Email Spam Classification (Phân loại thư rác) trên Kaggle được cộng đồng xây dựng sẵn, chứa thông tin của 5172 email cùng nhãn tương ứng cho biết đó là “Spam” hay “Ham”. Việc sử dụng dữ liệu có sẵn giúp đảm bảo độ tin cậy về nhãn (do dữ liệu đã được kiểm chứng bởi cộng đồng) và tiết kiệm thời gian so với tự thu thập thủ công. Tập dữ liệu được cung cấp dưới dạng tệp CSV (Comma-Separated Values), trong đó mỗi dòng tương ứng với một email và bao gồm các trường thông tin đặc trưng cũng như nhãn phân loại.

Quá trình thu thập dữ liệu cho nghiên cứu này không yêu cầu phát phiếu khảo sát hay phỏng vấn, thay vào đó dữ liệu thứ cấp được lấy trực tiếp từ kho dữ liệu trực tuyến. Nhóm nghiên cứu tiến hành tải tập tin CSV từ nguồn Kaggle về và sử dụng nó làm đầu vào cho các bước phân tích tiếp theo. Trước khi sử dụng, dữ liệu được kiểm tra nhanh về tính đầy đủ và nhất quán: đảm bảo không có giá trị khuyết thiếu nghiêm trọng và các nhãn spam/ham được đánh dấu rõ ràng. Do dữ liệu đã qua xử lý ban đầu bởi nguồn cung cấp (các email có định dạng văn bản đã được chuyển thành các đặc trưng số), nhóm nghiên cứu không cần thực hiện việc crawl email trực tiếp hay tự gán nhãn thủ công. Điều này giúp đảm bảo tính khách quan của nguồn dữ liệu và cho phép tập trung vào bước xây dựng mô hình máy học. Tóm lại, phương pháp thu thập dữ liệu trong đề tài là sử dụng dữ liệu công khai có sẵn trên internet (Kaggle), một cách tiếp cận phổ biến trong nghiên cứu ứng dụng học máy hiện nay.

3.4. Phân tích dữ liệu:

Trong bài toán phân loại, đặc biệt là phát hiện spam dựa trên dữ liệu ta có được (Naive Bayes, SVM, KNN, Neural Network), các phần mềm và công cụ phổ biến được sử dụng bao gồm các thư viện lập trình mạnh mẽ hỗ trợ học máy và học sâu.Với Python là ngôn ngữ chính, nhờ tính linh hoạt và cộng đồng hỗ trợ rộng lớn. Thư viện scikit-learn được sử dụng rộng rãi để triển khai các thuật toán như Naive Bayes (Gaussian NB, Multinomial NB), SVM, và KNN. Scikit-learn cung cấp các công cụ tiền xử lý dữ liệu (chuẩn hóa, mã hóa), huấn luyện mô hình, và đánh giá hiệu suất (Accuracy, Precision, Recall) một cách dễ dàng.

Đối với Neural Network, TensorFlow hoặc PyTorch là các công cụ chính. TensorFlow, với sự hỗ trợ của Keras, cho phép xây dựng và huấn luyện mạng nơ-ron với các lớp tùy chỉnh, hàm kích hoạt (ReLU, Sigmoid), và tối ưu hóa (Gradient Descent).

Ngoài ra, NumPy và Pandas hỗ trợ xử lý dữ liệu thô (như tập email), trong khi Matplotlib được dùng để trực quan hóa kết quả.Các công cụ này kết hợp tạo ra một quy trình hiệu quả từ tiền xử lý đến đánh giá mô hình, tối ưu hóa hiệu suất.

CHƯƠNG 4. THỰC NGHIỆM VÀ THẢO LUẬN

Chương này sẽ mô tả quá trình nghiên cứu, các bước áp dụng Neural Network vào mô hình, kết quả đạt được cũng như thảo luận và so sánh kết quả thực nghiệm được tiến hành trên tập dữ liệu đã được nêu trên.

4.1. Môi trường thực nghiệm

Thực nghiệm của nghiên cứu sẽ được thực hiện với ngôn ngữ lập trình python cùng với các thư viện tương ứng:

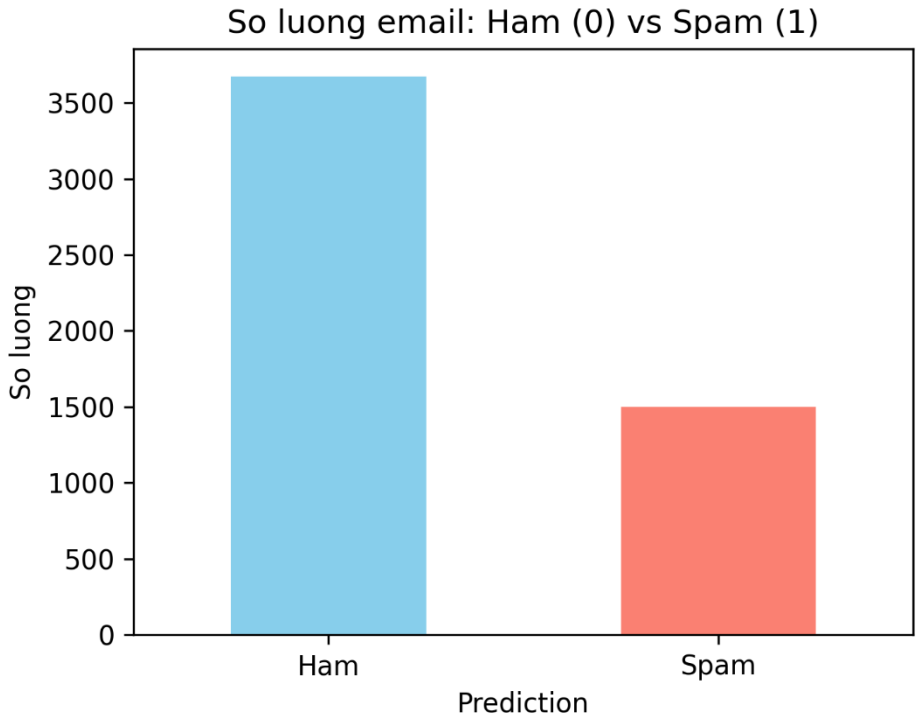
- Pandas: Thư viện hỗ trợ phân tích và thao tác dữ liệu.
- Matplotlib: Thư viện vẽ đồ thị, giúp trực quan hóa các kết quả, thuận tiện cho khảo sát và thống kê dữ liệu.
- Sklearn: Scikit-learn, thư viện chứ nhiều thuật toán máy học như Naïve Bayes, Random Forest, Decision Tree,… thư viện này sẽ được sử dụng để áp dụng các thuật toán trên vào mô hình máy học, từ đó lấy được kết quả, phục vụ cho việc so sánh hiệu suất.
- Tensorflow: Thư viện học sâu, được sử dụng cho thuật toán Neural Network mà ta sẽ áp dụng để phân loại văn bản/thư điện tử trong nghiên cứu này.
- Os: Thư viện cho phép thao tác với hệ thống, hỗ trợ việc lưu các hình ảnh đã tạo về thư mục, là một thư viện không bắt buộc phải có trong nghiên cứu này.

4.2. Tải dữ liệu vào mô hình

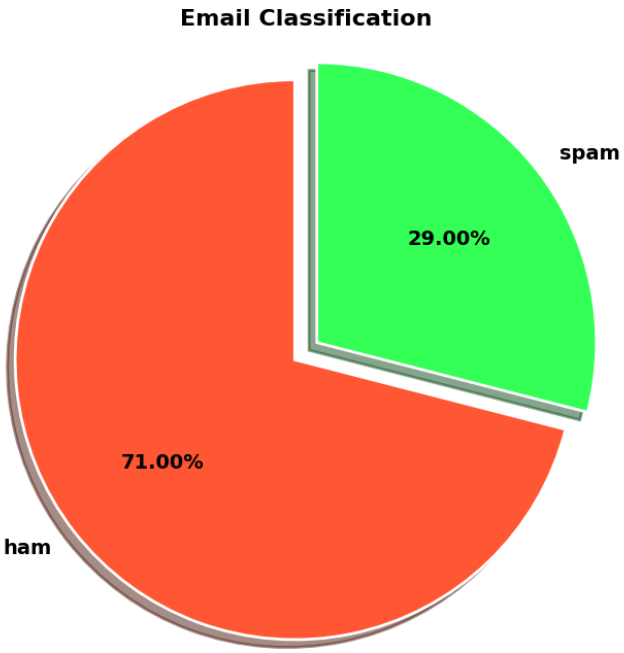
Dữ liệu mà nghiên cứu sẽ sử dụng để đào tạo mô hình được lấy từ Kaggle, với các thông tin về tập dữ liệu đã được cung cấp ở CHƯƠNG 2: LƯỢC KHẢO TÀI LIỆU. Dữ liệu sau khi được đọc vào mô hình sẽ thông qua quá trự làm sạch dữ liệu, bỏ đi các cột không phải là từ khóa và giữ lại các cột là từ khóa, từ khóa sẽ bao gồm chữ cái và các từ thậm chí là chưa hoàn chỉnh. Kết quả sau khi tải và làm sạch dữ liệu sẽ là bộ dữ liệu gồm 3000 từ được phân loại thuộc thư rác điện tử hoặc không.

4.3. Phân tích dữ liệu

4.3.1. Phân bổ số lượng thư rác và thư hợp lệ



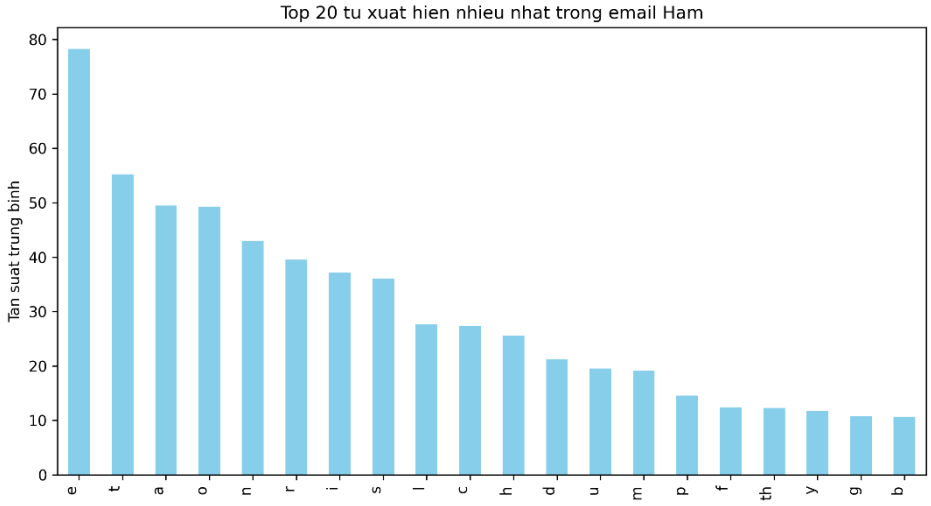
Hình 4.1. Số lượng thư hợp lệ và thư rác trong tập dữ liệu



Hình 4.2. Phần trăm giữ thư hợp lệ và thư rác

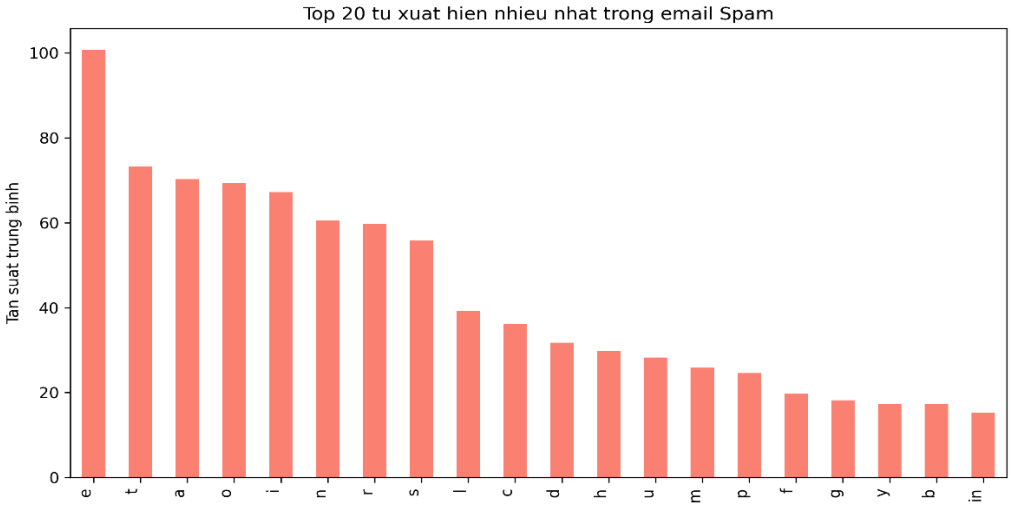
Có thể thấy được tập dữ liệu có phần không cân bằng giữa thư hợp lệ và thư rác, nhưng vì tỉ lệ cân bằng là tương đối chấp nhận được nên chúng ta sẽ không phải điều chỉnh tập dữ liệu trước khi đưa vào huấn luyện mô hình, cũng như tỉ lệ này sẽ không ảnh hưởng tới hiệu suất của mô hình vì sẽ phải thông qua bước chuẩn hóa dữ liệu.

4.3.2. Những từ xuất hiện nhiều nhất trong ham và spam



Hình 4.3. Những từ xuất hiện nhiều nhất trong thư hợp lệ

Sử dụng thư viện vẽ biểu đồ, nhóm nghiên cứu đã trực quan hóa Top 20 từ xuất hiện nhiều nhất trong hai loại email. Tuy không ra được kết quả là từ ngữ cụ thể hay hoàn chỉnh, biểu đồ của 2 loại vẫn giúp ta khám phá được những từ ngữ đặc trưng: ví dụ, các email spam có thể chứa thường xuyên các từ như “free”, “win” (miễn phí, trúng thưởng) trong khi email bình thường có thể hay chứa những từ thông dụng như “meeting”, “report” (phục vụ ngữ cảnh công việc.



Hình 4.4. Những từ xuất hiện nhiều nhất trong thư rác

Thông qua bước EDA, chúng tôi có cái nhìn định tính ban đầu về dữ liệu, phát hiện được những đặc trưng nổi bật và xác định rằng việc xử lý thêm (như cân bằng lại dữ liệu hoặc chọn lọc đặc trưng) có thể cần thiết hay không. Kết quả EDA cho thấy tập dữ liệu khá phong phú và các từ khóa phân biệt giữa spam/ham tồn tại rõ rệt, tạo tiền đề thuận lợi cho mô hình học máy học được quy luật phân loại.

4.4. Huấn luyện mô hình

- **Tải và chuẩn bị dữ liệu:** Lấy dữ liệu đã được xử lý và làm sạch từ các bước trên để sử dụng cho việc huấn luyện mô hình
- **Chuẩn hóa dữ liệu:** Dữ liệu sau khi được tải lên sẽ thông qua quá trình chuẩn hóa, quá trình này là cần thiết trong việc đào tạo mô hình mạng nơ ron, đảm bảo không có đặc trưng nào thống trị do thang đo giá trị đó lớn hơn
- **Xây dựng mô hình mạng nơ-ron sâu:** Sử dụng API chức năng của Keras với đầu vào là vector đăc trưng hóa của văn bản, chuỗi văn bản sau khi được vector và chuẩn hóa cùng với đó là các tầng ẩn, cụ thể như:
    - Tầng Dense với 128 nơ-ron, hàm kích hoạt ReLU, và regularization L2 (hệ số 0.001).
    - Tầng Dropout (tỷ lệ 0.3) để giảm overfitting.
    - Tầng Dense với 64 nơ-ron, hàm kích hoạt ReLU, và regularization L2.
    - Tầng Dropout thứ hai (tỷ lệ 0.3).

  Và đầu ra là tầng Dense với 1 nơ-ron, hàm kích hoạt sigmoid, dự đoán xác suất email là spam (0 hoặc1).
- **Chia dữ liệu thành 2 phần:** tập huấn luyện và tập kiểm thử, với 80% thư được sử dụng để huấn luyện mô hình và 20% còn lại được dùng để kiểm thử, việc tách riêng tập kiểm thử nhằm mục đích đánh

giá khách quan mô hình trên những dữ liệu mà mô hình chưa thấy trong quá trình học, qua đó phản ánh đúng khả năng tổng quát hóa của mô hình.

- **Huấn luyện mô hình:** dữ liệu đặc trưng đầu vào của mỗi email đi qua các lớp mạng nơ-ron, mô hình dự đoán nhãn và so sánh với nhãn thực tế, sau đó cập nhật các trọng số bên trong thông qua thuật toán tối ưu. Mô hình được thiết lập chạy trong nhiều epoch (vòng lặp qua toàn bộ dữ liệu huấn luyện) cho đến khi lỗi trên tập huấn luyện đạt mức ổn định hoặc cải thiện rất nhỏ. Trong quá trình huấn luyện, mô hình tự động điều chỉnh tham số để mô phỏng hàm ánh xạ từ nội dung email (đặc trưng) tới nhãn spam/ham. Để đảm bảo không xảy ra hiện tượng quá khớp (overfitting), nhóm nghiên cứu theo dõi độ lỗi trên tập huấn luyện và sử dụng tập kiểm thử (hoặc một phần của tập huấn luyện làm tập phát triển/kiểm định nếu cần) để quan sát khả năng tổng quát hoá sau mỗi vài epoch. Khi mô hình bắt đầu có dấu hiệu quá khớp (ví dụ độ chính xác trên tập huấn luyện tiếp tục tăng nhưng trên tập kiểm dừng cải thiện hoặc giảm), quá trình huấn luyện sẽ được dừng lại nhằm tránh việc mô hình ghi nhớ dữ liệu huấn luyện thay vì học quy luật tổng quát.

4.5. Đánh giá và so sánh kết quả

Sau khi hoàn thành việc huấn luyện, tập kiểm thử được sử dụng để đánh giá kết quả, hiệu năng của mô hình. Mô hình sẽ được đánh giá thông qua các thông số như Accuracy, Precision, Recall:

- Accuracy: Độ chính xác tổng thể của mô hình
- Precision: Độ chính xác của mô hình trong việc dự đoán thử rác điện tử (spam), hay trong số những thư điện tử được đánh dấu, bao nhiêu là đúng

- Recall: Tỉ lệ phát hiện thư điện tử rác trong tập dữ liệu kiểm thử

Có được kết quả từ mô hình mạng nơ ron học sâu, tiến hành so sánh với kết quả đó với kết quả có được khi áp dụng các thuật toán máy học khác như Naïve Bayes, Support Vector Machine, K-nearest neighbor vào phân loại thư điện tử, từ đó rút ra kết luận về hiệu năng, tốc độ của từng thuật toán. Sau đây là tổng hợp kết quả và thang đo của từng thuật toán được sử dụng:

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

Mặc dù là thuật toán kém phước tạp và giả định đơn giản nhưng thuật toán Naïve Bayes lại là thuật toán có chỉ số ấn tượng nhất, với các độ đo như Recall, Precision và Accuracy cao vượt trội, cao nhất so với 4 thuật toán được sử dụng trong bài toán phân loại thư điện tử. Cùng với đó thì các thuật toán khác như Support Vector Machine và K-nearest neighbor và Neural Network cũng có chỉ số ở mức hơn 95% với ngoại trừ duy nhất là chỉ số precision của thuật toán K-nearest neighbor.

Theo số liệu trên thì độ hiệu quả của thuật toán K-Nearest Neighbor có vẻ là tệ nhất trong các thuật toán, với chỉ số Precision ở mức thất vọng. Thuật toán Naïve Bayes có vẻ như là thuật toán thích hợp nhất để áp dụng giải bài toán phân loại này. Mặc dù Neural Network có chỉ số thua so với Naïve Bayes nhưng thuật toán này lại có tốc độ xử lý nhanh nhất với chỉ số không phải là quá tồi. Lý do mà thuật toán Neural Network có phần kém hiệu quả hơn so với Naïve Bayes có vẻ là do mô hình Neural Network không thể thể

hiện toàn bộ điểm mạnh của mình do độ phức tạp của bộ dữ liệu còn thấp cũng như không đủ lớn và đa dạng.

CHƯƠNG 5. KẾT LUẬN

5.1. Tổng kết

Qua quá trình nghiên cứu và thực nghiệm, đề tài đã hoàn thành mục tiêu đề ra khi xây dựng thành công một mô hình học máy phân loại thư điện tử thành spam hoặc ham. Bộ dữ liệu gần 5.200 email (spam/ham) từ Kaggle đã được thu thập và tiền xử lý kỹ lưỡng, bao gồm loại bỏ các thuộc tính không cần thiết và chuẩn hóa đặc trưng để đảm bảo dữ liệu đầu vào nhất quán. Trên cơ sở phân tích dữ liệu khám phá (EDA), đề tài nhận thấy sự khác biệt rõ rệt giữa email spam và ham về tần suất từ khóa và nội dung: các email spam thường chứa nhiều từ khóa chào mời, quảng cáo (“free”, “win”, “offer”, v.v.), trong khi email ham tập trung vào thông tin công việc, trao đổi cá nhân. Những hiểu biết này đã hỗ trợ lựa chọn phương pháp phù hợp cho mô hình.

Về phương pháp, đề tài thử sử dụng mô hình mạng nơ ron làm bộ phân loại chính. Mô hình được thiết kế với kiến trúc 3 lớp (có các hàm kích hoạt ReLU ở tầng ẩn và sigmoid ở đầu ra) và được huấn luyện bằng thuật toán tối ưu Adam kết hợp cơ chế EarlyStopping để tránh overfitting. Kết quả thực nghiệm cho thấy mô hình có tốc độ xử lý nhanh và đạt độ chính xác cao (xấp xỉ 96–97% trên tập kiểm thử), cùng với các chỉ số Precision, Accuracy, Recall đều ở mức tốt nhưng vẫn thua kém hơn so với thuật toán Naïve Bayes. Điều này chứng tỏ việc áp dụng Machine Learning ở các mức cao hơn cho các vấn đề đơn giản không phải luôn là lựa chọn tốt nhất khi mà bộ dữ liệu không đủ đa dạng và phức tạp. Nhìn chung, đề tài đã ứng dụng được mô hình máy học, hay cụ thể là học sâu vào bài toán phát hiện

thư rác, có được kết quả so sánh và rút ra được điểm mạnh của các thuật toán.

5.2. Hướng nghiên cứu tiếp theo

Mặc dù kết quả trong việc áp dụng mô hình máy học cao vào việc giải quyết bài toán không ra được kết quả như mong đợi, nhóm vẫn hướng tới ứng dụng các mô hình máy học cao hơn vào việc giải quyết bài toán phân loại này để tìm ra thuật toán phù hợp nhất cho vấn đề. Cũng như hướng theo phát triển về mặt thực tiễn, phát triển mô hình phát hiện thư rác theo thời gian thực, hỗ trợ lọc và quản lý thư từ cho người dùng, góp phần nâng cao an toàn thông tin người dùng và giảm thiểu tác hại từ thư rác trong môi trường số.

TÀI LIỆU THAM KHẢO

[1] Svistunova, O., Kulikova, T., Kovtun, A., Shimko, I., & Dedenok, R. (2024, June 7). Spam and phishing in 2023. Spam and Phishing in 2023. https://securelist.com/spam-phishing-report-2023/112015/

[2] Email Spam Classification Dataset CSV. (2020, March 10). Kaggle. https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv

[3] K. Agarwal and T . Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690.

[4] Harisinghaney, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. "Text and image-based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." In Optimization, Reliabilty, and Information T echnology (ICROIT), 2014 International Conference on, pp.153 -155. IEEE, 2014

[5] 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.