TRƯỜNG ĐẠI HỌC SÀI GÒN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG NGHIÊN CỨU KHOA HỌC
ÁP DỤNG MÁY HỌC VÀO PHÂN LOẠI THƯ RÁC ĐIỆN TỬ

HỌC PHẦN: PHƯƠNG PHÁP NGHIÊN CỨU KHOA HỌC

THUỘC NHÓM NGÀNH: CÔNG NGHỆ THÔNG TIN

GIẢNG VIÊN: ĐỖ NHƯ TÀI

THÀNH VIÊN NHÓM 77:

3123410024 - NGÔ THƯỢNG BẢO

3123410242 - NGUYỄN TRỌNG NHÂN

3123410174 - NGUYỄN PHẠM TUẤN KHÔI

3123410387 - NGUYỄN HỮU TRI

MỤC LỤC

1. Lý do chọn đề tài 2

2. Tổng quan vấn đề cần nghiên cứu 3

2.1. Tình hình nghiên cứu hiện tại 3

2.2. Hướng tiếp cận.....4

3. Mục đích và nhiệm vụ nghiên cứu.....5

4. Đối tượng và phạm vi nghiên cứu.....5

5. Phương pháp nghiên cứu.....6

6. Giả thuyết khoa học 8

7. Dự kiến kế hoạch nghiên cứu 8

8. Dự kiến nội dung của luận văn 8

9. Tài liệu tham khảo.....10

ĐỀ CƯƠNG NGHIÊN CỨU KHOA HỌC

1. Lý do chọn đề tài

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, sự phổ biến của email cũng kéo theo một vấn đề nhức nhối: thư rác (spam).

Theo thống kê từ các tổ chức bảo mật lớn trên thế giới, hơn 50% lưu lượng email toàn cầu mỗi ngày là thư rác. Các email này không chỉ làm gián đoạn hoạt động giao tiếp chính thống, mà còn có thể chứa liên kết độc hại, mã độc, hoặc nội dung lừa đảo tinh vi nhằm chiếm đoạt thông tin người dùng. Những hệ lụy này ảnh hưởng nghiêm trọng đến năng suất lao động, an toàn thông tin cá nhân và uy tín tổ chức.

Mặc dù nhiều hệ thống lọc spam đã được triển khai từ sớm, nhưng đa phần dựa trên phương pháp thủ công hoặc quy tắc tĩnh (rule-based), dễ bị qua mặt khi các kỹ thuật tạo thư spam ngày càng tinh vi. Trong bối cảnh đó, Trí tuệ nhân tạo (AI), đặc biệt là các kỹ thuật học máy (Machine Learning) và học sâu (Deep Learning), mở ra một hướng tiếp cận hiện đại và hiệu quả hơn cho bài toán này.

Vì vậy, đề tài "Áp dụng máy học vào phân loại thư rác điện tử" được lựa chọn với mong muốn:

- Giải quyết một vấn đề thực tiễn và cấp thiết của xã hội số hiện đại.
- Ứng dụng các thành tựu công nghệ tiên tiến trong lĩnh vực trí tuệ nhân tạo.
- Tạo nền tảng cho việc phát triển các hệ thống lọc email thông minh, linh hoạt và thích nghi tốt với các dạng spam mới.

- Góp phần nâng cao nhận thức và năng lực ứng phó với các hình thức lừa đảo qua email trong cộng đồng người dùng.

2. Tổng quan vấn đề cần nghiên cứu

2.1. Tình hình nghiên cứu hiện tại

a) Phương pháp truyền thống:

- Bộ lọc từ khóa và luật định sẵn: Dựa trên danh sách các từ khóa đặc trưng như "miễn phí", "trúng thưởng", "giảm giá",... và các luật như kiểm tra tiêu đề, người gửi, hay số lượng liên kết. Tuy nhiên, phương pháp này có độ chính xác thấp và dễ bị né tránh.

- Phương pháp Bayes[1]: Sử dụng xác suất thống kê để tính toán khả năng một email là spam dựa trên tần suất từ vựng. Dù phổ biến và hiệu quả ở mức cơ bản, nhưng Bayes yêu cầu lượng lớn dữ liệu và dễ bị lỗi khi gặp từ mới hoặc nội dung không phổ biến.

- Hệ thống điểm tin cậy (Reputation System): Dựa vào uy tín của địa chỉ email người gửi. Tuy nhiên, với sự xuất hiện của các địa chỉ email giả mạo hoặc bị tấn công, phương pháp này cũng mất dần hiệu quả.

b) Phương pháp hiện đại sử dụng AI:

- Machine Learning: Các mô hình như Naive Bayes[1], SVM[2], Random Forest[3],... đã được áp dụng và cho kết quả tốt hơn các phương pháp truyền thống. Chúng học từ dữ liệu và tự rút ra quy luật phân loại.

- Deep Learning: Các mạng nơ-ron (CNN, RNN, LSTM)[4]e có khả năng trích xuất đặc trưng sâu hơn từ nội dung email, kể cả các cấu trúc phức tạp hoặc ngữ cảnh ngôn ngữ.

- Xử lý ngôn ngữ tự nhiên (NLP): Với sự phát triển của các thư viện như NLTK, spaCy, BERT,... hệ thống có thể hiểu được ý nghĩa, ngữ cảnh của câu văn, từ đó phân biệt thư spam một cách tinh vi hơn.

- Học tăng cường (Reinforcement Learning): Dù chưa phổ biến rộng rãi trong spam detection, nhưng đây là hướng nghiên cứu tiềm năng, giúp mô hình học từ phản hồi thực tế và liên tục cải tiến.

2.2. Hướng tiếp cận

Trong khuôn khổ nghiên cứu này, đề tài tập trung vào ba hướng tiếp cận chính:

a) Phân tích và đánh giá:

- Nghiên cứu đặc điểm nhận dạng của các loại thư spam phổ biến: từ cấu trúc, nội dung, đến cách sử dụng ngôn ngữ.

- Đánh giá các phương pháp hiện tại: So sánh ưu nhược điểm giữa rule-based, ML truyền thống và Deep Learning.

- Thực hiện khảo sát hiệu quả của các mô hình AI hiện hành trong môi trường giả lập.

b) Phát triển giải pháp nhận diện:

- Chuẩn bị bộ dữ liệu đa dạng và đại diện cho cả hai lớp "spam" và "ham".

- Huấn luyện các mô hình học máy khác nhau như SVM, Random Forest, LSTM, BERT,... và so sánh hiệu quả.

- Tối ưu hóa tham số và tích hợp cơ chế tự học để mô hình cải tiến theo thời gian.

c) Đánh giá và cải tiến:

- Kiểm thử mô hình trên dữ liệu thực tế hoặc dữ liệu mới chưa từng xuất hiện trong huấn luyện.

- Phân tích các trường hợp mô hình phân loại sai (false positives và false negatives).

- Đề xuất hướng cải tiến hệ thống theo hướng học liên tục (online learning) hoặc kết hợp nhiều mô hình (ensemble learning).

3. Mục đích và nhiệm vụ nghiên cứu

Mục đích nghiên cứu: Đề tài hướng đến phát triển một hệ thống 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.

Nhiệm vụ nghiên cứu:

- 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 và hành vi người dùng 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).

- Phát hiện thời gian thực và khả năng thích ứng: Thiết kế hệ thống có khả năng phát hiện gian lận theo thời gian thực, đảm bảo khả năng cập nhật và thích ứng với các chiến thuật spam mới liên tục phát triển.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

- Dữ liệu email: Sử dụng các bộ dữ liệu công khai như Enron Spam Dataset[1] 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.

- 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.

- Đặc trưng đầu vào: Siêu dữ liệu (người gửi, tiêu đề, thời gian gửi) và hành vi người dùng (mẫu phản hồi, tỷ lệ nhấp chuột).

Phạm vi nghiên cứu:

- Phân tích ngôn ngữ, văn bản, bố cục và nội dung: 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,...Kết hợp xây dựng tập dữ liệu từ các trường hợp đã xác minh là gian lận để phân tích các chủ đề thường gặp (như giả mạo ngân hàng, thông báo trúng thưởng, đầu tư sinh lời cao).

- Nghiên cứu thủ đoạn và xu hướng lừa đảo: Thu thập và phân tích các thủ đoạn gian lận như giả mạo danh tính, giả danh tổ chức uy tín, sử dụng liên kết độc hại, lừa đảo chiếm đoạt tài sản qua ví điện tử hoặc mã OTP. Hành vi người dùng: Nghiên cứu các hành vi người dùng dễ bị lợi dụng (ví dụ: nhấp vào đường link lạ, chia sẻ thông tin cá nhân), từ đó đưa vào mô hình học máy các yếu tố hành vi.

- Mở rộng đối tượng phân tích theo thời gian: Đa nền tảng, đa lĩnh vực: Mở rộng phân tích gian lận không chỉ trong tài chính mà còn ở các nền tảng xã hội, sàn thương mại điện tử, ứng dụng nhắn tin.

5. Phương pháp nghiên cứu

Đề tài áp dụng phương pháp nghiên cứu kết hợp giữa phương pháp lý thuyết, phương pháp thực nghiệm và phương pháp chuyên gia để đảm bảo tính khoa học và hoành chỉnh trong suốt quá trình thực hiện nghiên cứu

Phương pháp lý thuyết:

- Tìm hiểu các nghiên cứu liên quan đến phân loại văn bản nói chung và phân loại thư rác nói riêng thông qua tài liệu, tạp chí hay bài báo khoa học
- Phân tích và tổng hợp lý thuyết về các chủ đề như NLP (Xử lý ngôn ngữ tự nhiên), Text Vectorization, Tokenization cũng như cách thức áp dụng cơ chế chú ý vào mô hình máy học, kèm theo đó là các thuật toán machine learning cơ bản đã từng được áp dụng trước đây như Random Forest hay Naïve Bayes
- Lựa chọn nguồn tài liệu uy tín như arXiv, IEEE hay Google Scholar để tham khảo và nghiên cứu

Phương pháp thực nghiệm:

- Thiết kế và triển khai kịch bản thực nghiệm các thuật toán và tích hợp cơ chế chú ý vào máy học
- Cài đặt mô hình bằng ngôn ngữ python, sử dụng các thư viện học sâu như TensorFlow hay PyTorch
- Tiền xử lý dữ liệu và sử dụng tập dữ liệu Enron-Spam cho bài toán phân loại thư rác điện tử, đảm bảo đáp ứng đủ về mặt dữ liệu cho mô hình
- Thực hiện huấn luyện và kiểm thử mô hình, đánh giá kết quả thông qua các chỉ số như Accuracy, Precision, Recall và F1-score
- So sánh, phân tích độ hiệu quả giữa các thuật toán, mô hình có sử dụng cơ chế chú ý và không, cũng như xét về mặt phức tạp cũng từng hướng tiếp cận, từ đó rút ra kết luận

Phương pháp chuyên gia:

- Trao đổi thường xuyên với giảng viên hướng dẫn, nhận phản hồi để đảm bảo hướng đi của nghiên cứu
- Dựa trên ý kiến giảng viên, điều chỉnh cách tiếp cận, tối ưu hóa quá trình thực hiện và đảm bảo tiến độ thực hiện nghien cứu

Phương pháp nghiên cứu này kết hợp sư chặt chẽ giữa nghiên cứu và thực nghiệm, cùng với sự hỗ trợ từ chuyên gia, đảm bảo tính khoa học và khả thi của nghiên cứu

6. Giả thuyết khoa học

Nghiên cứu giải quyết được vấn đề phân loại thư rác điện tử dựa trên các đặc tính như số lượng kí tự, tần suất từ, địa chỉ người gửi, từ khóa,… Cũng như cải thiện hiệu suất phân loại văn bản so với các thuật toán/mô hình được sử dụng trước đây trước đây.

7. Dự kiến kế hoạch nghiên cứu

STT	Nội dung công việc	Thời gian dự kiến thực hiện
1	Nghiên cứu, chọn đề tài	1 tuần
2	Định hướng cho đề tài	1 tuần
3	Phân tích, khảo sát dữ liệu	2 tuần
4	Xây dựng mô hình máy học	2 tuần
5	Nghiên cứu, viết báo cáo	2 tuần
6	Chỉnh sửa, hoàn thành báo cáo	2 tuần

Bảng kế hoạch nghiên cứu dự kiến

8. Dự kiến nội dung của luận văn

Chương 1: Giới thiệu

- Lý do chọn đề tài.
- Mục tiêu nghiên cứu.
- Đối tượng và phạm vi nghiên cứu.
- Phương pháp nghiên cứu.
- Những đóng góp mới của đề tài.
- Cấu trúc luận văn.

Chương 2: Cơ sở lý thuyết

Khái niệm về thư điện tử, sự khác nhau giữa thư điện tử rác và thư điện tử bình thường

Tổng quan về xử lý ngôn ngữ tự nhiên, tokenization và text vectorization

Giới thiệu về các thuật toán phân loại:

- Naïve Bayes
- Support Vector Machine
- K-nearest neighbors
- Random Forest
- Neural Network (ANN và CNN)

Tổng quan về bài toán phân loại văn bản và tập dữ liệu Enron-Spam

Chương 3: Phương pháp nghiên cứu và mô hình đề xuất

Quy trình nghiên cứu và các bước thực hiện

Xem xét khả năng ứng dụng cơ chế chú ý để giải quyết bài toán nhận dạng thư rác điện tử

Thiết kế kịch bản thực nghiệm:

- Chuẩn bị dữ liệu: Tiền xử lý, chia tập huấn luyện và kiểm thử
- Cài đặt mô hình: sử dụng python với các thư viện học sâu

Chương 4: Thực nghiệm và kết quả

Thực hiện các thí nghiệm trên tập dữ liệu Enron-Spam

- Huấn luyện và kiểm thử mô hình có sử dụng cơ chế chú ý và không
- Đánh giá hiệu quả giữa các thuật toán được sử dụng trong mô hình, giữa mô hình có sử dụng cơ chế chú ý và không thông qua các chỉ số như Accuracy, Precision, Recall, F1-score

Phân tích từng loại thuật toán được sử dùng trong mô hình

- So sánh độ hiệu quả giữa các thuật toán và mô hình có và không sử dụng cơ chế chú ý

Biểu diễn các kết quả bằng biểu đồ và các hình ảnh trực quan

Chương 5: Kết luận và hướng phát triển

- Tóm tắt các kết quả đạt được.
- Đánh giá
- Hạn chế của nghiên cứu
- Đề xuất hướng nghiên cứu và ứng dụng trong tương lai.

9. Tài liệu tham khảo

[1] Metsis, Vangelis & Androutsopoulos, Ion & Paliouras, Georgios. (2006). Spam Filtering with Naive Bayes - Which Naive Bayes?. In CEAS.

[2] Mammone, Alessia, Marco Turchi, and Nello Cristianini. "Support vector machines." *Wiley Interdisciplinary Reviews: Computational Statistics* 1.3 (2009): 283-289.

[3] Rigatti, Steven J. "Random forest." *Journal of Insurance Medicine* 47.1 (2017): 31-39.

[4] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436-444.