**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN**

****

**BÀI TẬP LỚN**

**TRÍ TUỆ NHÂN TẠO**

**NGHIÊN CỨU PHÂN LOẠI EMAIL SPAM**

NGÀNH: KHOA HỌC MÁY TÍNH

CHUYÊN NGÀNH: TRÍ TUỆ NHÂN TẠO VÀ KHOA HỌC DỮ LIỆU

SINH VIÊN: **ĐỖ THU LIỄU**

LỚP: **12423TN**

GV GIẢNG DẠY: **PGS.TS.Nguyễn Văn Hậu**

**HƯNG YÊN – 2025**

**NHẬN XÉT**

Nhận xét của giáo viên hướng dẫn:

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

.................................................................................................................................

**GIÁO VIÊN HƯỚNG DẪN**

LỜI CAM ĐOAN

Tôi xin cam đoan bài tập lớn “Nghiên cứu phương pháp xây dựng email spam ” là kết quả thực hiện của bản thân tôi dưới sự hướng dẫn của cô Nguyễn Hoàng Điệp.

Những phần sử dụng tài liệu tham khảo trong bài tập lớn đã được nêu rõ trong phần tài liệu tham khảo. Các kết quả trình bày trong bài tập lớn và chương trình xây dựng được hoàn toàn là kết quả do bản thân tôi thực hiện.

Nếu vi phạm lời cam đoan này, tôi xin chịu hoàn toàn trách nhiệm trước khoa và nhà trường.

*Hưng Yên, ngày … tháng … năm 2025*

Sinh viên

Đỗ Thu Liễu

LỜI CẢM ƠN

Để có thể hoàn thành bài tập lớn này, lời đầu tiên tôi xin phép gửi lời cảm ơn tới bộ môn Khoa học máy tính, Khoa Công nghệ thông tin – Trường Đại học Sư phạm Kỹ thuật Hưng Yên đã tạo điều kiện thuận lợi cho tôi thực hiện bài tập lớn môn học này.

Đặc biệt tôi xin chân thành cảm ơn thầy Nguyễn Minh Tiến đã rất tận tình hướng dẫn, chỉ bảo tôi trong suốt thời gian thực hiện bài tập lớn vừa qua.

Tôi cũng xin chân thành cảm ơn tất cả các Thầy, các Cô trong Trường đã tận tình giảng dạy, trang bị cho tôi những kiến thức cần thiết, quý báu để giúp tôi thực hiện được bài tập lớn này.

Mặc dù tôi đã có cố gắng, nhưng với trình độ còn hạn chế, trong quá trình thực hiện đề tài không tránh khỏi những thiếu sót. Tôi hy vọng sẽ nhận được những ý kiến nhận xét, góp ý của các Thầy cô về những kết quả triển khai trong bài tập lớn.

Tôi xin trân trọng cảm ơn!

MỤC LỤC

Contents

[LỜI CAM ĐOAN 3](#_Toc218023268)

[LỜI CẢM ƠN 4](#_Toc218023269)

[MỤC LỤC 5](#_Toc218023270)

[DANH MỤC HÌNH ẢNH 6](#_Toc218023271)

[CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI 8](#_Toc218023272)

[1.1. Lý do chọn đề tài 8](#_Toc218023273)

[1.2. Mục tiêu của đề tài 9](#_Toc218023274)

[1.2.1. Mục tiêu tổng quát 9](#_Toc218023275)

[1.2.2. Mục tiêu cụ thể 10](#_Toc218023276)

[1.3. Giới hạn và phạm vi của đề tài 10](#_Toc218023277)

[1.3.1. Dữ liệu nghiên cứu 11](#_Toc218023278)

[1.3.2. Kỹ thuật phân tích 11](#_Toc218023279)

[1.3.3. Phạm vi ứng dụng 11](#_Toc218023280)

[1.3.4. Giới hạn nghiên cứu 12](#_Toc218023281)

[1.4. Nội dung thực hiện 12](#_Toc218023282)

[1.5. Phương pháp tiếp cận 13](#_Toc218023283)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 15](#_Toc218023284)

[2.1. Tổng quan về học máy (Machine Learning) 15](#_Toc218023285)

[2.1.1. Khái niệm học máy 15](#_Toc218023286)

[2.1.2. Đặc điểm nổi bật của học máy 16](#_Toc218023287)

[2.1.3. Phân loại các phương pháp học máy 17](#_Toc218023288)

[2.1.4. Ứng dụng của học máy trong thực tế. 18](#_Toc218023289)

[2.1.5. Vai trò của học máy trong đề tài 19](#_Toc218023290)

[2.1.6. Tổng quan về quá trình huấn luyện 20](#_Toc218023291)

[2.1.7. Chuẩn bị và tiền xử lý dữ liệu 20](#_Toc218023292)

[2.1.8. Quá trình huấn luyện mô hình 21](#_Toc218023293)

[2.2. Logistic Regression 21](#_Toc218023294)

[2.2.1. Giới thiệu thuật toán Logistic Regression 21](#_Toc218023295)

[2.2.2. Cấu trúc của mô hình Logistic Regression 22](#_Toc218023296)

[2.2.3. Quá trình xây dựng mô hình Logistic Regression 22](#_Toc218023297)

[2.2.4. Ưu điểm của Logistic Regression 23](#_Toc218023298)

[2.2.5. Nhược điểm của Logistic Regression 23](#_Toc218023299)

[2.3. Naïve Bayes 23](#_Toc218023300)

[2.3.1. Giới thiệu thuật toán Naïve Bayes 23](#_Toc218023301)

[2.3.2. Cấu trúc của mô hình Naïve Bayes 24](#_Toc218023302)

[2.3.3. Phương pháp mã hóa văn bản. 24](#_Toc218023303)

[2.3.4. Đánh giá mô hình học máy 26](#_Toc218023304)

[2.4. Tổng kết 27](#_Toc218023305)

[CHƯƠNG 3: GIẢI PHÁP 28](#_Toc218023306)

[3.1. Tìm hiểu dữ liệu bài toán 28](#_Toc218023307)

[3.2. Khám phá dữ liệu 28](#_Toc218023308)

[3.3. Khám phá dữ liệu 30](#_Toc218023309)

[3.3.1. Phân phối email ham và email spam 30](#_Toc218023310)

[3.3.2. Phân bố độ dài câu 31](#_Toc218023311)

[3.3.3. Trực quan các từ thường xuất hiện trong email 32](#_Toc218023312)

[3.4. Chuẩn bị dữ liệu 36](#_Toc218023313)

[3.4.1. Xử lý dữ liệu 36](#_Toc218023314)

[3.4.2. Trực quan dữ liệu sau khi xử lý 37](#_Toc218023315)

[3.4.3. Chuẩn hóa dữ liệu 39](#_Toc218023316)

[3.4.4. Chia dữ liệu huấn luyện và kiểm tra 42](#_Toc218023317)

[3.5. Xây dựng và đánh giá mô hình 43](#_Toc218023318)

[3.5.1. Hàm đánh giá mô hình 43](#_Toc218023319)

[3.5.2. Huấn luyện mô hình 43](#_Toc218023320)

[3.5.3. Huấn luyện với phương pháp TF-IDF . 44](#_Toc218023321)

[3.5.4. Huấn luyện với phương pháp BOW. 47](#_Toc218023322)

[3.6. So sánh các mô hình dựa trên bài toán 50](#_Toc218023323)

[3.6.1. Tiêu chí so sánh 50](#_Toc218023324)

[3.6.2. So sánh trực tiếp 51](#_Toc218023325)

[4.1. Cài đặt Streamlit 53](#_Toc218023326)

[KẾT LUẬN 54](#_Toc218023327)

DANH MỤC HÌNH ẢNH

[Hình 2. 1 .Tổng quan mối quan hệ giữa AI – Machine Learning 15](#_Toc218023438)

[Hình 3. 1: SMS Spam Collection Dataset. 29](#_Toc218023448)

[Hình 3. 2: Kích thước dataframe. 30](#_Toc218023449)

[Hình 3. 3: Thông tin tổng quan dữ liệu 30](#_Toc218023450)

[Hình 3. 4: Kiểu dữ liệu từng cột 30](#_Toc218023451)

[Hình 3. 5: Kiểm tra số giá trị bị thiếu 30](#_Toc218023452)

[Hình 3. 6: Số lượng từ trong từ điển. 31](#_Toc218023453)

[Hình 3. 7: Số bản ghi trùng lặp. 31](#_Toc218023454)

[Hình 3. 8 Biểu đồ phân phối email spam được và email ham. 31](#_Toc218023455)

[Hình 3. 9:Độ dài câu theo số kí tự. 32](#_Toc218023456)

[Hình 3. 11 Biểu đồ thể hiện phân phối độ dài câu trước khi xử lý dữ liệu. 32](#_Toc218023457)

[Hình 3. 12: Các từ xuất hiện nhiều trong email spam. 34](#_Toc218023458)

[Hình 3. 13: Các từ xuất hiện nhiều trong email ham. 35](#_Toc218023459)

[Hình 3. 14: Biểu đồ trực quan bằng Wordcloud. 36](#_Toc218023460)

[Hình 3. 15: Dữ liệu còn lại sau khi khóa bản ghi trùng lặp. 37](#_Toc218023461)

[Hình 3. 16: Xử lý tin nhắn email. 38](#_Toc218023462)

[Hình 3. 17: Dữ liệu sau khi xử lý. 38](#_Toc218023463)

[Hình 3. 18: Lượng từ còn lại trong từ điển. 39](#_Toc218023464)

[Hình 3. 19: Phân bố độ dài câu sau khi làm sạch dữ liệu 39](#_Toc218023465)

[Hình 3. 20: Trực quan từ cloud cho spam và ham sau khi làm sạch 40](#_Toc218023466)

[Hình 3. 21: Tính độ dài câu theo số ký tự làm sạch 40](#_Toc218023467)

[Hình 3. 22: Mã hóa dữ liệu bằng phương pháp Bag of Words. 40](#_Toc218023468)

[Hình 3. 23: Mã hóa dữ liệu bằng phương pháp TF-IDF. 41](#_Toc218023469)

[Hình 3. 24: Dùng Smote để xử lý mất cân bằng dữ liệu. 42](#_Toc218023470)

[Hình 3. 25: Dùng MaxAbsScaler để chuẩn hóa dữ liệu. 43](#_Toc218023471)

[Hình 3. 26: Kết quả của 2 mô hình chạy với TF-IDF. 46](#_Toc218023472)

[Hình 3. 27: Mã trận nhầm lẫn Logistic Regression với TF-IDF. 46](#_Toc218023473)

[Hình 3. 28: Mã trận nhầm lẫn Naïve Bays với TF-IDF. 47](#_Toc218023474)

[Hình 3. 29: Kết quả của 2 mô hình chạy với BOW. 48](#_Toc218023475)

[Hình 3. 30: Mã trận nhầm lẫn Naïve Bays với BOW 49](#_Toc218023476)

[Hình 3. 31: Mã trận nhầm lẫn Logistic Regression với BOW. 50](#_Toc218023477)

[Hình 4. 1 Giao diện Demo chính 54](#_Toc218023478)

[Hình 4. 2 Kết quả dự đoán 55](#_Toc218023479)

1. TỔNG QUAN VỀ ĐỀ TÀI
   1. Lý do chọn đề tài

Email là một trong những phương tiện giao tiếp điện tử phổ biến và quan trọng nhất trong đời sống hiện đại, được sử dụng rộng rãi trong học tập, công việc và các hoạt động kinh doanh. Bên cạnh những email hữu ích phục vụ trao đổi thông tin, quảng bá sản phẩm hay chăm sóc khách hàng, người dùng ngày càng phải đối mặt với số lượng lớn email rác (spam), gây phiền toái, lãng phí thời gian và tiềm ẩn nhiều rủi ro về bảo mật thông tin.

Cùng với sự phát triển mạnh mẽ của Internet và các nền tảng trực tuyến, lượng email được gửi và nhận mỗi ngày tăng lên với tốc độ rất nhanh. Các hệ thống email phải xử lý một khối lượng dữ liệu khổng lồ bao gồm nội dung văn bản, tiêu đề, liên kết và tệp đính kèm. Việc phân loại, lọc và quản lý email một cách chính xác và hiệu quả trở thành một yêu cầu cấp thiết nhằm bảo vệ người dùng khỏi các nội dung không mong muốn, lừa đảo hoặc chứa mã độc.

Tuy nhiên, trong nhiều trường hợp, việc phân loại email spam và email hợp lệ (ham) vẫn dựa nhiều vào các quy tắc thủ công hoặc sự can thiệp của con người, dẫn đến tốn kém thời gian, công sức và dễ xảy ra sai sót. Khi số lượng email ngày càng gia tăng và nội dung spam ngày càng tinh vi, các phương pháp truyền thống không còn đáp ứng tốt yêu cầu về tốc độ và độ chính xác. Do đó, việc ứng dụng các phương pháp học máy và xử lý ngôn ngữ tự nhiên để tự động hóa quá trình phân loại email spam trở nên cần thiết và có ý nghĩa thực tiễn cao.

Song song với giá trị ứng dụng, việc nghiên cứu và xây dựng hệ thống phân loại email spam còn mang lại giá trị học tập lớn cho sinh viên. Quá trình thực hiện đề tài giúp người học tiếp cận với tư duy khoa học, rèn luyện kỹ năng phân tích dữ liệu, xử lý văn bản, hiểu rõ quy trình xây dựng mô hình học máy và áp dụng công nghệ vào các bài toán thực tế.

Bên cạnh đó, đề tài còn góp phần nâng cao nhận thức về vai trò của công nghệ thông tin trong việc đảm bảo an toàn và hiệu quả cho các hệ thống giao tiếp điện tử. Một hệ thống phân loại email spam hoạt động hiệu quả không chỉ giúp giảm thiểu email rác mà còn nâng cao trải nghiệm người dùng, bảo vệ thông tin cá nhân và tăng độ tin cậy của các dịch vụ email.

Từ những lý do trên, tôi lựa chọn đề tài “Xây dựng hệ thống phân loại email spam” với mong muốn tìm hiểu và áp dụng các kỹ thuật học máy và xử lý ngôn ngữ tự nhiên vào một bài toán thực tiễn, góp phần nâng cao hiệu quả quản lý email, đồng thời giúp người học hiểu rõ hơn về sự ứng dụng của công nghệ trong đời sống số hiện nay.Thời trang là một trong những lĩnh vực gần gũi và gắn bó mật thiết với đời sống con người, thể hiện không chỉ nhu cầu mặc đẹp mà còn là cách con người thể hiện cá tính, phong cách và văn hóa. Cùng với sự phát triển nhanh chóng của xã hội hiện đại, ngành thời trang ngày càng trở thành một lĩnh vực có ảnh hưởng mạnh mẽ, không chỉ trong đời sống thường nhật mà còn trong kinh tế, quảng cáo và thương mại điện tử.

* 1. Mục tiêu của đề tài
     1. Mục tiêu tổng quát

Mục tiêu chính của đề tài là nghiên cứu và xây dựng mô hình học máy (Machine Learning) nhằm tự động phân loại email spam và email hợp lệ (ham) dựa trên nội dung văn bản của email. Đề tài hướng tới việc phát triển một hệ thống thông minh có khả năng nhận diện và phân loại chính xác các email không mong muốn thông qua việc huấn luyện mô hình trên tập dữ liệu email lớn đã được gán nhãn.

Cụ thể, hệ thống sẽ áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) kết hợp với các thuật toán học máy phổ biến như Naive Bayes, Logistic Regression, Support Vector Machine hoặc Decision Tree, cùng với các phương pháp biểu diễn văn bản như Bag of Words (BoW) và TF-IDF để trích xuất đặc trưng từ nội dung email. Thông qua quá trình huấn luyện và đánh giá, mô hình sẽ học được các đặc trưng giúp phân biệt email spam và email hợp lệ một cách hiệu quả.

Hệ thống phân loại email spam được xây dựng không chỉ giúp tự động hóa quá trình lọc email, giảm thiểu email rác và các nguy cơ lừa đảo, mà còn góp phần nâng cao trải nghiệm người dùng, tiết kiệm thời gian và tăng độ an toàn cho các hệ thống thư điện tử. Đồng thời, việc thực hiện đề tài còn giúp người học nắm vững quy trình xây dựng một mô hình học máy từ tiền xử lý dữ liệu, trích xuất đặc trưng, huấn luyện mô hình đến đánh giá và cải tiến hiệu quả phân loại trong các bài toán thực tế.

* + 1. Mục tiêu cụ thể

Tìm hiểu cơ sở lý thuyết về học máy (Machine Learning) và các phương pháp xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP), làm nền tảng cho bài toán phân loại email spam.

Khảo sát và tiền xử lý tập dữ liệu email, bao gồm các bước như: làm sạch văn bản (loại bỏ ký tự đặc biệt, số, dấu câu), chuyển chữ thường, loại bỏ từ dừng (stopwords), và chuẩn hóa dữ liệu văn bản; sau đó chia tập dữ liệu thành các phần huấn luyện – kiểm thử – đánh giá.

Áp dụng các phương pháp biểu diễn văn bản như Bag of Words (BoW) và TF-IDF để chuyển đổi nội dung email từ dạng văn bản sang dạng đặc trưng số phục vụ cho mô hình học máy.

Xây dựng và huấn luyện các mô hình phân loại học máy như Naive Bayes, Logistic Regression, Support Vector Machine (SVM) hoặc Decision Tree, đồng thời sử dụng các kỹ thuật như cân bằng dữ liệu (ví dụ: SMOTE) nhằm giảm ảnh hưởng của hiện tượng mất cân bằng lớp trong dữ liệu email spam.

Huấn luyện và đánh giá mô hình thông qua các chỉ số như Accuracy, Precision, Recall, F1-score, cùng với ma trận nhầm lẫn (Confusion Matrix) để phân tích chi tiết hiệu quả phân loại email spam và email hợp lệ (ham).

Đưa ra nhận xét, phân tích kết quả thực nghiệm và đề xuất định hướng phát triển trong tương lai, chẳng hạn như thử nghiệm các mô hình khác, tinh chỉnh tham số, kết hợp n-gram, hoặc triển khai hệ thống phân loại email spam vào các ứng dụng thực tế nhằm nâng cao hiệu quả lọc email.

* 1. Giới hạn và phạm vi của đề tài

Nghiên cứu này tập trung vào việc phân tích và phân loại email spam dựa trên tập dữ liệu email spam (ví dụ: Spam Dataset được cung cấp từ Kaggle hoặc nguồn dữ liệu công khai tương đương). Phạm vi cụ thể của nghiên cứu bao gồm:

* + 1. Dữ liệu nghiên cứu

Nghiên cứu này tập trung vào việc phân loại email dựa trên tập dữ liệu SMS Spam Collection được cung cấp bởi Kaggle.

Tập dữ liệu bao gồm 5572 mẫu dữ liệu, được chia thành

* 4385 dữ liệu huấn luyện
* 774 dữ liệu kiểm tra

Mỗi tin nhắn đại diện cho một loại email thuộc2 nhãn (classes) khác nhau, bao gồm: email spam hoặc email ham

Tập dữ liệu đã được chuẩn hóa và cân bằng, không chứa giá trị bị thiếu, thuận lợi cho việc huấn luyện mô hình học máy.

* + 1. Kỹ thuật phân tích

Nghiên cứu thực hiện các bước phân tích và xử lý dữ liệu hình ảnh, bao gồm:

* **Khảo sát đặc điểm dữ liệu (EDA)** nhằm hiểu rõ phân bố của các lớp **spam** và **ham,** tần suất từ khóa và mức độ mất cân bằng dữ liệu trong tập email.
* **Áp dụng các kỹ thuật tiền xử lý dữ liệu văn bản,** bao gồm: làm sạch nội dung email, chuyển chữ thường, loại bỏ ký tự đặc biệt, số và từ dừng (stopwords), nhằm chuẩn hóa dữ liệu đầu vào cho mô hình.
* **Trích xuất đặc trưng văn bản** bằng các phương pháp như **Bag of Words (BoW)** và **TF-IDF,** chuyển đổi nội dung email từ dạng văn bản sang dạng số để phục vụ quá trình huấn luyện mô hình.
* **Xây dựng và huấn luyện các mô hình học máy** (ví dụ: **Naive Bayes, Logistic Regression),** nhằm học các đặc trưng giúp phân biệt email spam và email hợp lệ.
* **Huấn luyện mô hình trên tập dữ liệu huấn luyện,** đánh giá hiệu quả trên tập kiểm thử thông qua các chỉ số như **Accuracy, Precision, Recall, F1-score,** cùng với **ma trận nhầm lẫn (Confusion Matrix)** để phân tích chi tiết kết quả phân loại.
  + 1. Phạm vi ứng dụng

Đề tài tập trung vào việc phân loại email spam và email hợp lệ (ham) trong phạm vi bài toán phân loại nhị phân dựa trên nội dung văn bản email.

Mục tiêu là xây dựng một hệ thống tự động nhận diện email spam từ dữ liệu đầu vào, có thể được ứng dụng trong:

• Hệ thống lọc email cho các dịch vụ thư điện tử nhằm giảm thiểu email rác và nội dung không mong muốn.

• Công cụ phát hiện sớm các email lừa đảo, quảng cáo độc hại hoặc chứa liên kết nguy hiểm.

• Hệ thống hỗ trợ quản lý và bảo mật thông tin, giúp nâng cao trải nghiệm người dùng và độ tin cậy của các nền tảng giao tiếp điện tử.

* + 1. Giới hạn nghiên cứu

Đề tài chỉ sử dụng tập dữ liệu email spam có sẵn trên Kaggle, không bao gồm việc thu thập hoặc mở rộng dữ liệu email thực tế từ các nguồn bên ngoài.

Dữ liệu nghiên cứu chỉ bao gồm nội dung văn bản của email, chưa khai thác các thông tin nâng cao khác như tiêu đề chi tiết, siêu dữ liệu (metadata), tệp đính kèm hoặc liên kết bên ngoài.

Nghiên cứu tập trung vào bài toán phân loại nhị phân email spam và email hợp lệ (ham), chưa đi sâu vào việc phân loại các loại spam cụ thể như quảng cáo, lừa đảo (phishing) hay mã độc.

Kết quả của mô hình phụ thuộc lớn vào chất lượng, độ đa dạng và mức độ mất cân bằng của tập dữ liệu email, cũng như phương pháp tiền xử lý và trích xuất đặc trưng được sử dụng trong nghiên cứu.

* 1. Nội dung thực hiện

Đề tài gồm 4 chương:

Chương 1: Tổng quan về đề tài

Chương 2: Cơ sở lý thuyết

Chương 3: Giải pháp

Chương 4: Xây dựng Streamlit

* 1. Phương pháp tiếp cận

Đề tài “Xây dựng hệ thống phân loại email spam” sử dụng phương pháp tiếp cận dựa trên học máy (Machine Learning) kết hợp với xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) nhằm tự động nhận diện và phân loại email spam và email hợp lệ (ham) một cách chính xác và hiệu quả. Phương pháp tiếp cận của đề tài được chia thành ba giai đoạn chính: chuẩn bị và tiền xử lý dữ liệu, xây dựng và huấn luyện mô hình học máy, và đánh giá – tối ưu hóa mô hình.

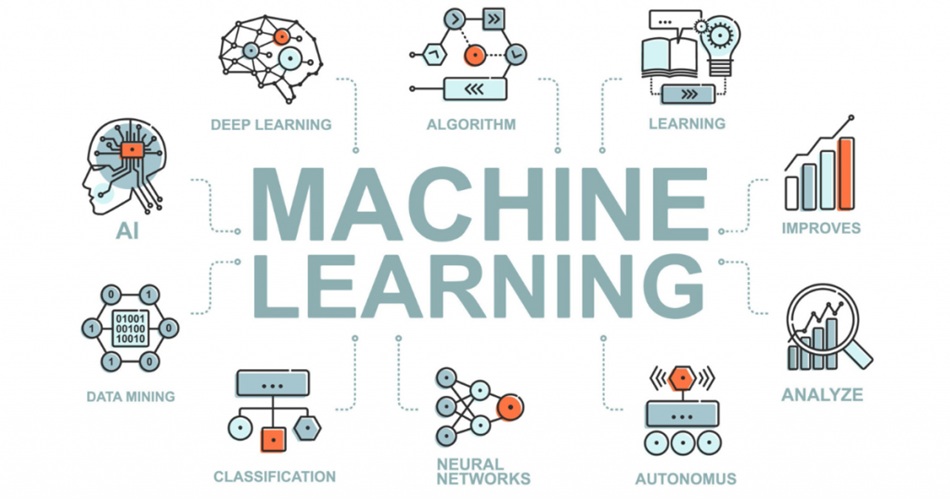
Trong giai đoạn đầu, tập dữ liệu email spam được sử dụng làm nguồn dữ liệu chính (ví dụ: spam.csv từ Kaggle). Tập dữ liệu bao gồm các email dưới dạng văn bản cùng nhãn phân loại tương ứng (spam/ham). Dữ liệu được tiền xử lý thông qua các bước như: loại bỏ ký tự đặc biệt, số và dấu câu; chuyển chữ thường; loại bỏ từ dừng (stopwords); và chuẩn hóa văn bản. Sau đó, dữ liệu được chia thành các tập huấn luyện, kiểm thử và xác thực. Nội dung email được biểu diễn dưới dạng đặc trưng số bằng các phương pháp như Bag of Words (BoW) hoặc TF-IDF nhằm phục vụ cho quá trình huấn luyện mô hình.

Ở giai đoạn tiếp theo, đề tài áp dụng các thuật toán học máy để xây dựng mô hình phân loại email spam, tiêu biểu như Naive Bayes, Logistic Regression hoặc Support Vector Machine (SVM). Các mô hình này học mối quan hệ giữa các đặc trưng văn bản và nhãn phân loại để phân biệt email spam và email hợp lệ. Trong trường hợp dữ liệu mất cân bằng, các kỹ thuật như cân bằng dữ liệu (SMOTE) hoặc điều chỉnh trọng số lớp được áp dụng nhằm cải thiện hiệu quả phân loại.

Cuối cùng, mô hình được đánh giá và tối ưu hóa dựa trên các chỉ số như Accuracy, Precision, Recall và F1-score. Kết quả thực nghiệm được phân tích thông qua ma trận nhầm lẫn (Confusion Matrix) và các báo cáo phân loại nhằm đánh giá chi tiết hiệu suất mô hình trong việc phát hiện email spam. Việc tinh chỉnh tham số và lựa chọn mô hình phù hợp giúp nâng cao độ chính xác và khả năng tổng quát hóa của hệ thống.

Phương pháp tiếp cận này giúp xây dựng một hệ thống phân loại email spam tự động, hiệu quả và dễ triển khai, góp phần giảm thiểu email rác, nâng cao trải nghiệm người dùng và tạo nền tảng cho việc ứng dụng vào các hệ thống lọc email và bảo mật thông tin trong thực tế.

1. CƠ SỞ LÝ THUYẾT
   1. Tổng quan về học máy (Machine Learning)



Hình 2. 1 . Tổng quan mối quan hệ giữa AI – Machine Learning

* + 1. Khái niệm học máy

Trong lĩnh vực trí tuệ nhân tạo (Artificial Intelligence – AI), học máy (Machine Learning) là một nhánh quan trọng, tập trung vào việc xây dựng các mô hình và thuật toán cho phép máy tính tự học từ dữ liệu và đưa ra dự đoán hoặc quyết định mà không cần được lập trình một cách tường minh cho từng trường hợp cụ thể.

Khác với các phương pháp lập trình truyền thống dựa trên các quy tắc cố định, học máy cho phép hệ thống tự động phát hiện các mẫu (patterns) và mối quan hệ trong dữ liệu thông qua quá trình huấn luyện trên tập dữ liệu có sẵn. Các mô hình học máy sử dụng các đặc trưng (features) được trích xuất từ dữ liệu đầu vào để học cách ánh xạ từ dữ liệu sang kết quả mong muốn.

Về bản chất, học máy mô phỏng cách con người học hỏi từ kinh nghiệm. Khi được cung cấp dữ liệu huấn luyện, mô hình sẽ dần cải thiện hiệu suất thông qua việc điều chỉnh các tham số nhằm giảm sai số dự đoán. Chất lượng của mô hình phụ thuộc vào dữ liệu, phương pháp trích xuất đặc trưng và thuật toán học máy được lựa chọn.

Nhờ khả năng xử lý và phân tích dữ liệu hiệu quả, học máy đã và đang được ứng dụng rộng rãi trong nhiều lĩnh vực như phân loại văn bản, lọc email spam, nhận dạng mẫu, dự đoán xu hướng, và nhiều bài toán thực tế khác, đóng vai trò quan trọng trong sự phát triển của các hệ thống thông minh hiện nay.

* + 1. Đặc điểm nổi bật của học máy

Học máy (Machine Learning) có nhiều đặc điểm nổi bật so với các phương pháp lập trình truyền thống, thể hiện ở những khía cạnh cơ bản sau:

Khả năng học từ dữ liệu: Thay vì được lập trình sẵn bằng các quy tắc cứng nhắc, mô hình học máy có thể tự học các mẫu và mối quan hệ trong dữ liệu thông qua quá trình huấn luyện. Trong bài toán phân loại email spam, mô hình học máy học được sự khác biệt giữa email spam và email hợp lệ dựa trên nội dung văn bản và tần suất xuất hiện của các từ khóa.

Phụ thuộc vào đặc trưng đầu vào: Khác với học sâu, học máy truyền thống yêu cầu xác định và trích xuất đặc trưng thủ công từ dữ liệu đầu vào. Ví dụ, trong phân loại email spam, các đặc trưng thường là tần suất từ, n-gram hoặc trọng số TF-IDF, đóng vai trò quan trọng trong hiệu quả của mô hình.

Khả năng mô hình hóa mối quan hệ tuyến tính và phi tuyến: Nhiều thuật toán học máy như Logistic Regression, Decision Tree, SVM có khả năng mô hình hóa các mối quan hệ tuyến tính và phi tuyến trong dữ liệu, giúp phân biệt các lớp dữ liệu một cách hiệu quả.

Hiệu quả với tập dữ liệu vừa và nhỏ: So với học sâu, các mô hình học máy hoạt động hiệu quả với tập dữ liệu có quy mô vừa, không đòi hỏi tài nguyên phần cứng quá lớn và có thời gian huấn luyện nhanh hơn.

Dễ triển khai và giải thích: Các mô hình học máy truyền thống thường dễ triển khai, dễ huấn luyện và dễ giải thích kết quả hơn so với học sâu. Điều này đặc biệt quan trọng trong các bài toán như lọc email spam, nơi cần hiểu rõ vì sao một email bị đánh dấu là spam.

Ứng dụng rộng rãi trong xử lý văn bản: Học máy đã chứng minh hiệu quả cao trong các bài toán phân loại văn bản, phân tích cảm xúc, lọc thư rác, và nhiều ứng dụng thực tế khác trong lĩnh vực xử lý ngôn ngữ tự nhiên.

* + 1. Phân loại các phương pháp học máy

Các thuật toán học máy được chia thành ba nhóm chính dựa trên cách hệ thống học từ dữ liệu:

* Học có giám sát (Supervised Learning)
* Dữ liệu huấn luyện có sẵn nhãn (label), tức là biết trước kết quả mong muốn.
* Mục tiêu của mô hình là học mối quan hệ giữa đầu vào (input) và đầu ra (output).
* Sau khi huấn luyện, mô hình có thể dự đoán nhãn cho dữ liệu mới.
* Ví dụ: Dự đoán có gian lận thẻ không (0 – không, 1 – có).
* Một số thuật toán phổ biển:
* Hồi quy tuyến tính (Linear Regression).
* Cây quyết định (Decision Tree).
* K – láng giềng gần nhất (KNN).
* Máy vector hỗ trợ (SVM).
* Rừng ngẫu nhiên (Random Forest).
* Học không giám sát
* Dữ liệu không có nhãn, mô hình phải tự khám phá cấu trúc ẩn trong dữ liệu.
* Thường được dùng để phân cụm (clustering) hoặc giảm chiều dữ liệu (dimensionality reduction).
* Ví dụ: Nhóm khách hàng có hành vi mua hàng tương tự nhau, hoặc phân loại các đặc điểm bệnh nhân theo nhóm triệu chứng.
* Một số thuật toán tiêu biểu:
* K-means Clustering
* Hierarchical Clustering
* Principal Component Analysis (PCA)
* Học tăng cường
* Mô hình tương tác với môi trường, học thông qua thử – sai (trial and error), nhận phần thưởng (reward) hoặc hình phạt (penalty) dựa trên hành động.
* Mục tiêu: tìm ra chiến lược tối ưu để đạt phần thưởng cao nhất.
* Ứng dụng: robot tự hành, hệ thống gợi ý thông minh, và các trò chơi tự động.
* Thuật toán tiêu biểu: Q-Learning, Deep Q-Network (DQN), Policy Gradient.
  + 1. Ứng dụng của học máy trong thực tế.

Học máy (Machine Learning) hiện nay đã trở thành nền tảng của nhiều công nghệ thông minh trong đời sống hiện đại. Nhờ khả năng học từ dữ liệu và tự cải thiện hiệu suất theo thời gian, các mô hình học máy được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau.

Trong lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP), học máy được sử dụng để phân loại và phân tích văn bản. Các ứng dụng tiêu biểu bao gồm lọc email spam, phân tích cảm xúc, phân loại tin tức, gợi ý nội dung và hỗ trợ tìm kiếm thông tin. Thông qua việc học từ nội dung văn bản và tần suất từ khóa, các hệ thống có thể tự động phân biệt giữa thông tin hữu ích và nội dung không mong muốn.

Trong y học, học máy hỗ trợ phân tích dữ liệu bệnh nhân, dự đoán nguy cơ mắc bệnh và hỗ trợ chẩn đoán dựa trên các chỉ số lâm sàng. Các mô hình học máy giúp bác sĩ đưa ra quyết định nhanh hơn, chính xác hơn, đặc biệt trong việc phát hiện sớm các bệnh lý nguy hiểm.

Trong tài chính – ngân hàng, học máy được ứng dụng trong phát hiện gian lận giao dịch, đánh giá rủi ro tín dụng và dự đoán xu hướng thị trường. Bằng cách phân tích hành vi giao dịch, hệ thống có thể phát hiện các giao dịch bất thường và giảm thiểu rủi ro cho tổ chức tài chính.

Trong thương mại điện tử, học máy đóng vai trò quan trọng trong hệ thống gợi ý sản phẩm, phân tích hành vi người dùng và cá nhân hóa trải nghiệm mua sắm. Các thuật toán học máy giúp đề xuất sản phẩm phù hợp với sở thích của khách hàng, tối ưu hóa quảng cáo và nâng cao hiệu quả kinh doanh.

Ngoài ra, trong giao thông và logistics, học máy được sử dụng để dự đoán lưu lượng, tối ưu hóa tuyến đường và quản lý vận hành. Các hệ thống này giúp tiết kiệm chi phí, thời gian và nâng cao hiệu quả hoạt động.

Từ những ứng dụng trên có thể thấy, học máy không chỉ là một công cụ kỹ thuật mà còn là cầu nối giữa dữ liệu và tri thức, giữa công nghệ và đời sống thực tiễn. Việc ứng dụng học máy đã và đang góp phần nâng cao hiệu quả làm việc, chất lượng dịch vụ và tạo ra nhiều giá trị thiết thực cho xã hội hiện đại.

* + 1. Vai trò của học máy trong đề tài

Trong khuôn khổ của đề tài **“Xây dựng hệ thống phân loại email spam”, học máy (Machine Learning)** giữ vai trò trung tâm trong toàn bộ quá trình nghiên cứu và phát triển hệ thống. Bài toán đặt ra là **phân loại tự động các email thành hai nhóm chính: email spam và email hợp lệ (ham)** — một nhiệm vụ khó có thể thực hiện hiệu quả bằng các phương pháp thủ công do số lượng email lớn và nội dung ngày càng đa dạng, phức tạp. Việc áp dụng học máy cho phép hệ thống **tự động học từ dữ liệu email** và đưa ra quyết định phân loại mà không cần sự can thiệp trực tiếp của con người.

Thông qua các thuật toán học máy, hệ thống có khả năng **phân tích và khai thác các đặc trưng quan trọng trong nội dung email** như tần suất từ khóa, cấu trúc văn bản và các mẫu ngôn ngữ thường xuất hiện trong email spam. Các mô hình học máy như **Naive Bayes, Logistic Regression** giúp học được mối quan hệ giữa các đặc trưng này và nhãn phân loại, từ đó đưa ra kết quả phân loại chính xác ngay cả khi nội dung spam được thay đổi hoặc che giấu tinh vi.

Bên cạnh ý nghĩa về mặt kỹ thuật, việc ứng dụng học máy trong đề tài còn mang lại **giá trị học tập và thực tiễn rõ rệt.** Quá trình thực hiện giúp người học hiểu rõ **quy trình xây dựng một hệ thống học máy hoàn chỉnh**, từ thu thập và tiền xử lý dữ liệu văn bản, trích xuất đặc trưng, huấn luyện mô hình, đánh giá hiệu quả cho đến triển khai ứng dụng. Qua đó, đề tài góp phần nâng cao tư duy phân tích, kỹ năng xử lý dữ liệu và khả năng ứng dụng công nghệ vào các bài toán thực tế.

Về mặt ứng dụng, hệ thống phân loại email spam **giúp tự động hóa quá trình lọc email,** giảm thiểu email rác, hạn chế nguy cơ lừa đảo và nâng cao trải nghiệm người dùng trong các hệ thống thư điện tử. Điều này cho thấy học máy không chỉ là một công cụ nghiên cứu lý thuyết, mà còn là **giải pháp hiệu quả cho các vấn đề thực tiễn**, đặc biệt trong lĩnh vực an toàn thông tin và truyền thông số hiện nay.

* + 1. Tổng quan về quá trình huấn luyện

Trong học máy (Machine Learning), quá trình huấn luyện mô hình (Model Training) là giai đoạn quan trọng nhằm giúp mô hình học được mối quan hệ giữa dữ liệu đầu vào và đầu ra mong muốn. Mục tiêu của quá trình huấn luyện là tối ưu các tham số của mô hình (ví dụ: trọng số, hệ số, tham số mô hình) sao cho sai số giữa giá trị dự đoán và giá trị thực tế được giảm xuống mức thấp nhất.

Quá trình huấn luyện thường được thực hiện thông qua nhiều vòng lặp (iterations hoặc epochs). Trong mỗi vòng lặp, mô hình sử dụng dữ liệu huấn luyện để tạo ra dự đoán, sau đó so sánh với giá trị thực nhằm tính toán hàm mất mát (loss function). Dựa trên sai số này, các tham số của mô hình được cập nhật theo thuật toán tối ưu (như Gradient Descent) để cải thiện hiệu suất dự đoán.

Nhờ quá trình lặp lại liên tục, mô hình học máy dần nâng cao khả năng tổng quát hóa và đưa ra dự đoán chính xác hơn trên dữ liệu mới, tương tự như cách con người học hỏi thông qua việc rút kinh nghiệm từ các sai sót trong quá khứ.

* + 1. Chuẩn bị và tiền xử lý dữ liệu

Trước khi đưa dữ liệu vào huấn luyện, cần thực hiện tiền xử lý (data preprocessing) nhằm đảm bảo dữ liệu ở định dạng và phạm vi phù hợp với mô hình.  
Các bước phổ biến bao gồm:

* Chuẩn hóa (Normalization): đưa giá trị của dữ liệu về cùng một thang đo, thường trong khoảng [0,1] hoặc [-1,1], giúp mô hình hội tụ nhanh hơn.
* Mã hóa dữ liệu phân loại (Encoding): biến đổi các nhãn phân loại thành dạng số hoặc vector nhị phân (one-hot encoding).
* Chia tập dữ liệu: tách dữ liệu thành ba phần chính:
  + *Tập huấn luyện (Training set)* – dùng để điều chỉnh trọng số mô hình.
  + *Tập kiểm thử (Test set)* – dùng để đánh giá khách quan mô hình sau khi huấn luyện xong.

Tiền xử lý dữ liệu là một bước thiết yếu vì chất lượng dữ liệu đầu vào có ảnh hưởng trực tiếp đến hiệu quả học của mô hình.

* + 1. Quá trình huấn luyện mô hình

Quá trình huấn luyện một mô hình học máy thường bao gồm nhiều bước liên tiếp, được mô tả tổng quát như sau:

Chuẩn bị và khởi tạo mô hình: Lựa chọn thuật toán học máy phù hợp (ví dụ: Naive Bayes, Logistic Regression) và khởi tạo các tham số ban đầu của mô hình theo cấu hình mặc định hoặc do người dùng thiết lập.

Huấn luyện mô hình (Training): Dữ liệu huấn luyện, sau khi đã được tiền xử lý và trích xuất đặc trưng (ví dụ: TF-IDF, BOW được chuẩn hóa qua MaxAbsScaler)), được đưa vào mô hình. Mô hình sẽ học mối quan hệ giữa các đặc trưng đầu vào và nhãn đầu ra (spam/ham) thông qua việc điều chỉnh các tham số nội bộ.

Tính toán hàm mất mát hoặc hàm đánh giá: Trong quá trình huấn luyện, mô hình sử dụng các hàm mục tiêu (objective function) hoặc tiêu chí tối ưu để đo lường mức độ sai lệch giữa kết quả dự đoán và nhãn thực tế, từ đó cải thiện khả năng phân loại.

Cập nhật tham số mô hình: Dựa trên kết quả đánh giá, các tham số của mô hình được điều chỉnh nhằm giảm sai số dự đoán. Quá trình này diễn ra tự động bên trong thuật toán học máy đã chọn.

Đánh giá mô hình: Sau khi huấn luyện, mô hình được kiểm tra trên tập dữ liệu kiểm thử để đánh giá khả năng tổng quát hóa thông qua các chỉ số như Accuracy, Precision, Recall và F1-score, cùng với ma trận nhầm lẫn (Confusion Matrix).

Lặp và tối ưu hóa: Quá trình huấn luyện – đánh giá có thể được lặp lại nhiều lần với các cấu hình khác nhau (thay đổi thuật toán, tham số, phương pháp trích xuất đặc trưng) nhằm lựa chọn mô hình có hiệu suất tốt nhất.

Trong suốt quá trình này, các kỹ thuật như chia tập dữ liệu hợp lý, xử lý mất cân bằng dữ liệu, và tinh chỉnh tham số (hyperparameter tuning) đóng vai trò quan trọng trong việc nâng cao độ chính xác và độ ổn định của mô hình học máy.

* 1. Logistic Regression
     1. Giới thiệu thuật toán Logistic Regression

Logistic Regression là một thuật toán học máy thuộc nhóm học có giám sát (supervised learning), chủ yếu được sử dụng cho bài toán phân loại nhị phân (dự đoán giữa hai lớp, ví dụ: 0 hoặc 1, như bệnh/có bệnh hoặc spam/không spam). Mặc dù có tên là "regression", Logistic Regression thực chất là một mô hình phân loại, vì nó dự đoán xác suất một điểm dữ liệu thuộc vào một lớp cụ thể. Thuật toán sử dụng hàm sigmoid để ánh xạ các giá trị thực thành xác suất trong khoảng [0, 1].

* + 1. Cấu trúc của mô hình Logistic Regression
* Hàm sigmoid: Hàm sigmoid chuyển đổi đầu ra tuyến tính thành xác suất.

Trong đó, với:

w: Vector trọng số (hệ số hồi quy).

x: Vector đặc trưng đầu vào.

b: Hằng số (bias).

* Xác suất đầu ra: Xác suất một điểm dữ liệu x thuộc lớp C1 (lớp dương) được tính bằng:
* Xác suất thuộc lớp C0 (lớp âm) là:
* Ngưỡng phân loại: Thông thường, ngưỡng 0.5 được sử dụng để quyết định lớp:
  + 1. Quá trình xây dựng mô hình Logistic Regression
* Chuẩn hóa dữ liệu: Các đặc trưng cần được chuẩn hóa (ví dụ: chuẩn hóa Min-Max hoặc Z-score) để đảm bảo tất cả đặc trưng có cùng thang đo, giúp tối ưu hóa hiệu quả.
* Tối ưu hàm mất mát: Sử dụng hàm mất mát log-loss (hay còn gọi là Binary Cross-Entropy):

Trong đó:

:Nhãn thực tế của mẫu thứ i.

Xác suất dự đoán.

N: Số lượng mẫu.

* Hàm mất mát được tối ưu bằng Gradient Descent hoặc các biến thể (như SGD, Adam) để tìm **w** và ***b*** tối ưu.
* Dự đoán: Dựa trên xác suất đầu ra, gán lớp bằng ngưỡng (thường là 0.5, nhưng có thể điều chỉnh tùy bài toán).
  + 1. Ưu điểm của Logistic Regression
* Đơn giản và hiệu quả: Thuật toán dễ triển khai, tính toán nhanh, phù hợp cho các bài toán phân loại nhị phân cơ bản.
* Kết quả dễ giải thích: Hệ số w cho biết mức độ quan trọng của từng đặc trưng.
* Tổng quát hóa tốt: Hoạt động hiệu quả với dữ liệu có tính tuyến tính.
  + 1. Nhược điểm của Logistic Regression
* Đơn giản và hiệu quả: Thuật toán dễ triển khai, tính toán nhanh, phù hợp cho các bài toán phân loại nhị phân cơ bản.
* Kết quả dễ giải thích: Hệ số w cho biết mức độ quan trọng của từng đặc trưng.
  1. Naïve Bayes
     1. Giới thiệu thuật toán Naïve Bayes

Thuật toán Naïve Bayes dựa trên Định lý Bayes trong xác suất, giả định rằng các đặc trưng (features) độc lập với nhau. Mặc dù giả định này đơn giản, nhưng Naïve Bayes hoạt động rất hiệu quả trong nhiều bài toán phân loại, đặc biệt là xử lý văn bản và phân tích cảm xúc.

* + 1. Cấu trúc của mô hình Naïve Bayes
* Định lý Bayes:

P(C∣X)=(P(X∣C)×P(C))/(P(X))

* Trong đó:
  + P(C∣X): Xác suất thuộc lớp C khi biết X
  + P(X∣C): Xác suất quan sát thấy X khi biết C
  + P(C): Xác suất tiên nghiệm của lớp C
  + Giả định độc lập: Các biến X1, X2, …, Xn độc lập với nhau.
* Quá trình xây dựng mô hình Naïve Bayes
  + Bước 1: Tính xác suất tiên nghiệm P(C)của từng lớp.
  + Bước 2: Tính xác suất có điều kiện P(X\_i∣C)cho từng đặc trưng.
  + Bước 3: Áp dụng định lý Bayes để tính P(C∣X).
  + Bước 4: Chọn lớp có xác suất lớn nhất làm kết quả dự đoán.
* Ưu điểm của Naïve Bayes
  + Đơn giản, nhanh và dễ triển khai.
  + Hiệu quả với dữ liệu lớn và nhiều đặc trưng.
  + Hoạt động tốt với dữ liệu rời rạc hoặc văn bản.
  + Không yêu cầu nhiều tham số.
* Nhược điểm của Naïve Bayes
  + Giả định độc lập tuyệt đối giữa các đặc trưng là không thực tế.
  + Dự đoán xác suất có thể sai lệch nếu các biến phụ thuộc mạnh.
* Hiệu suất kém với dữ liệu có mối tương quan cao giữa các thuộc tính.
  + 1. Phương pháp mã hóa văn bản.
       1. TF-IDF
* Định nghĩa:

TF-IDF là phương pháp đánh giá tầm quan trọng của một từ trong một tài liệu so với toàn bộ tập tài liệu.

* Cấu thành:

**TF** (Term Frequency – Tần suất từ)

* Đo lường tần suất xuất hiện của từ trong một tài liệu.
* Từ xuất hiện nhiều → TF cao.

**IDF** (Inverse Document Frequency – Tần suất ngược tài liệu)

* Đo lường tầm quan trọng của từ trong toàn bộ tập dữ liệu.
* Từ xuất hiện nhiều trong nhiều tài liệu → IDF thấp (ít đặc trưng).
* Từ hiếm → IDF cao (đặc trưng hơn).

**TF-IDF**

TF-IDF=TF×IDF

* Từ quan trọng trong tài liệu nhưng hiếm trong tập dữ liệu → TF-IDF cao.
* Ứng dụng:
* Nhấn mạnh từ đặc trưng trong văn bản, giảm ảnh hưởng của từ phổ biến.
* Sử dụng rộng rãi trong phân loại văn bản, tìm kiếm thông tin, spam email.
  + - 1. BOW
* Định nghĩa:
* Bag of Words là phương pháp biểu diễn văn bản dưới dạng tập từ (words) không quan tâm đến thứ tự, mỗi từ trở thành một đặc trưng (feature) trong vector.
* Cách hoạt động:
* Mỗi văn bản được chuyển thành vector số, giá trị vector là tần suất từ xuất hiện trong văn bản.
* Ví dụ: văn bản "win a prize" với từ điển ["win", "a", "prize", "money"] → vector [1, 1, 1, 0].
* Ưu điểm:
* Dễ triển khai, trực quan, phù hợp với các mô hình học máy như Naive Bayes hay Logistic Regression.
* Giữ thông tin về tần suất từ, giúp mô hình học phân biệt nội dung văn bản.
* Hạn chế:
* Không giữ thứ tự từ và ngữ cảnh, vector có thể rất sparse (nhiều giá trị 0) khi từ điển lớn.
  + 1. Đánh giá mô hình học máy

Sau khi huấn luyện xong, mô hình cần được đánh giá một cách toàn diện để kiểm tra mức độ chính xác và khả năng tổng quát hóa trên dữ liệu mới.  
Các tiêu chí đánh giá phổ biến trong học sâu bao gồm:

* Độ chính xác (Accuracy): Là tỷ lệ giữa số lượng dự đoán đúng và tổng số mẫu. Đây là chỉ số đơn giản nhưng trực quan, thường được dùng cho bài toán phân loại.

A black and white text

AI-generated content may be incorrect.

* Độ chính xác theo lớp (Precision):Cho biết trong số những mẫu được mô hình dự đoán là dương tính, có bao nhiêu mẫu thực sự đúng.

A black and white math equation

AI-generated content may be incorrect.

* Độ nhạy (Recall): Phản ánh khả năng của mô hình trong việc phát hiện đúng các mẫu thuộc về lớp thực sự.

A math equation with black text

AI-generated content may be incorrect.

* F1-score: Là trung bình điều hòa của Precision và Recall, giúp đánh giá cân bằng giữa hai yếu tố này.

A close-up of words

AI-generated content may be incorrect.

* Ma trận nhầm lẫn (Confusion Matrix): Là bảng thể hiện trực quan mối quan hệ giữa nhãn thực tế và nhãn dự đoán, giúp phát hiện các lớp mà mô hình hay nhầm lẫn.

Các chỉ số này thường được tính toán dựa trên tập kiểm thử (test set) – tập dữ liệu độc lập chưa từng được mô hình “nhìn thấy” trong quá trình học.

* 1. Tổng kết

Chương 2 đã hệ thống hóa các cơ sở lý thuyết liên quan đến học sâu và mạng nơ-ron nhân tạo, làm rõ nguyên lý hoạt động, cấu trúc và vai trò của các thành phần như hàm kích hoạt, hàm mất mát, bộ tối ưu hóa và các kỹ thuật huấn luyện. Bên cạnh đó, chương cũng trình bày chi tiết về mạng nơ-ron tích chập (CNN) – một kiến trúc tiêu biểu trong xử lý hình ảnh, cùng các kỹ thuật hỗ trợ như BatchNormalization, MaxPooling và Dropout, giúp nâng cao hiệu quả và độ ổn định của mô hình.

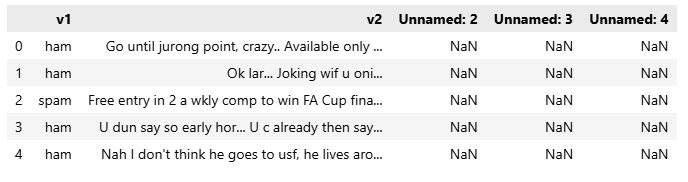
Những nội dung lý thuyết được trình bày trong chương này đóng vai trò là nền tảng quan trọng cho các bước thực nghiệm ở phần sau. Trên cơ sở đó, Chương 3 sẽ tập trung vào việc trình bày phương pháp thực hiện, bao gồm quy trình xử lý dữ liệu, thiết kế mô hình, huấn luyện và đánh giá hiệu quả nhằm hoàn thiện hệ thống phân loại hình ảnh.

CHƯƠNG 3: GIẢI PHÁP

3.1. Tìm hiểu dữ liệu bài toán

Bộ Dataset: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

**Dataset:** SMS Spam Collection Dataset là một tập dữ liệu phân loại văn bản, được sử dụng rộng rãi trong bài toán phân loại tin nhắn/email spam. Mục tiêu của tập dữ liệu là dự đoán một tin nhắn thuộc loại spam hay không spam (ham) dựa trên nội dung văn bản của tin nhắn.

Tập dữ liệu bao gồm 5.574 mẫu tin nhắn SMS, trong đó mỗi mẫu được gán một nhãn phân loại tương ứng: 

Hình 3. 1: SMS Spam Collection Dataset.

Mô tả dữ liệu:

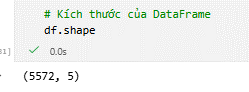
* spam: tin nhắn rác, quảng cáo hoặc lừa đảo
* ham: tin nhắn hợp lệ, không phải spam

Mỗi bản ghi trong tập dữ liệu gồm các thuộc tính chính:

* Label: nhãn phân loại (spam hoặc ham)
* Message: nội dung văn bản của tin nhắn

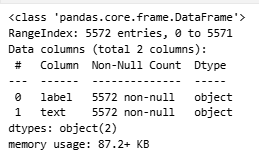
3.2. Khám phá dữ liệu

* df.shape: Xem kích thước dataframe.



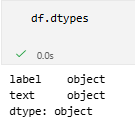
Hình 3. 2: Kích thước dataframe.

* df.info(): Xem thông tin tổng quan như số cột, kiểu dữ liệu, và số giá trị không rỗng.



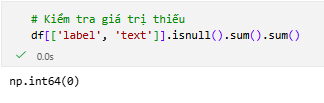
Hình 3. 3: Thông tin tổng quan dữ liệu

* df.dtypes: Xem kiểu dữ liệu của từng cột.



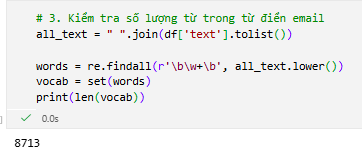
Hình 3. 4: Kiểu dữ liệu từng cột

* df.isnull().sum(): Kiểm tra số giá trị bị thiếu trong mỗi cột.



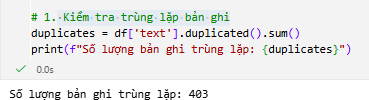
Hình 3. 5: Kiểm tra số giá trị bị thiếu

* Lượng từ trong từ điển



Hình 3. 6: Số lượng từ trong từ điển.

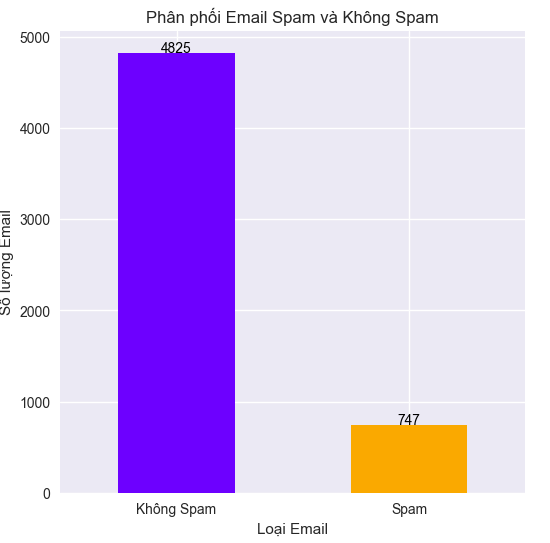
* Kiểm tra trùng lặp bản ghi



Hình 3. 7: Số bản ghi trùng lặp.

3.3. Khám phá dữ liệu

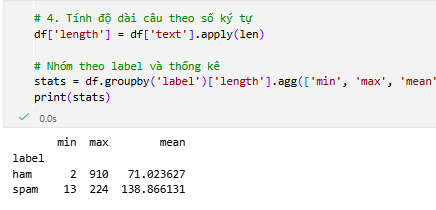
3.3.1. Phân phối email ham và email spam



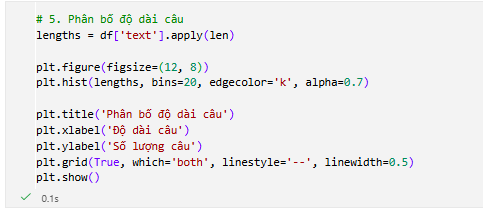
Hình 3. 8 Biểu đồ phân phối email spam được và email ham.

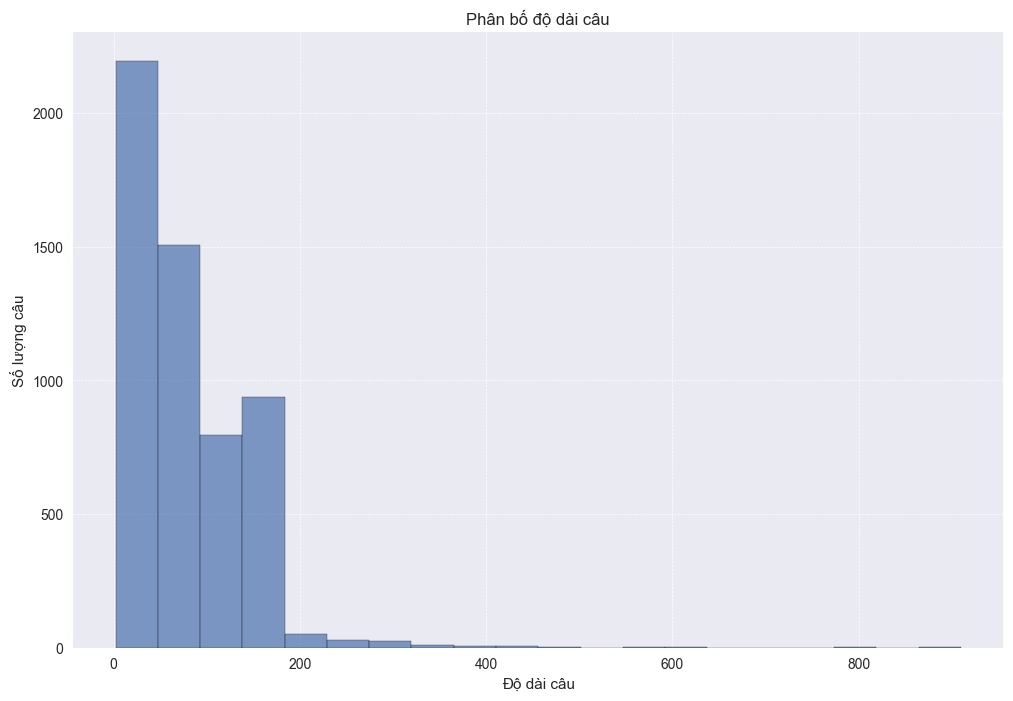
* Ý nghĩa biểu đồ:
* Biểu đồ cột giúp so sánh số lượng email ham được và spam được trong tập dữ liệu.
* Ứng dụng: Xem xét phân phối dữ liệu để nhận định liệu dữ liệu có cân bằng hay không (balanced vs. imbalanced dataset).

3.3.2. Phân bố độ dài câu



Hình 3. 9:Độ dài câu theo số kí tự.

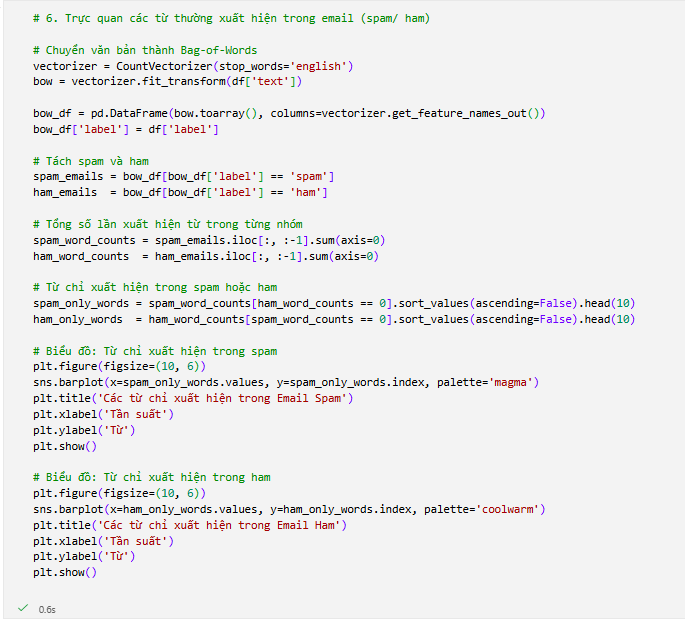


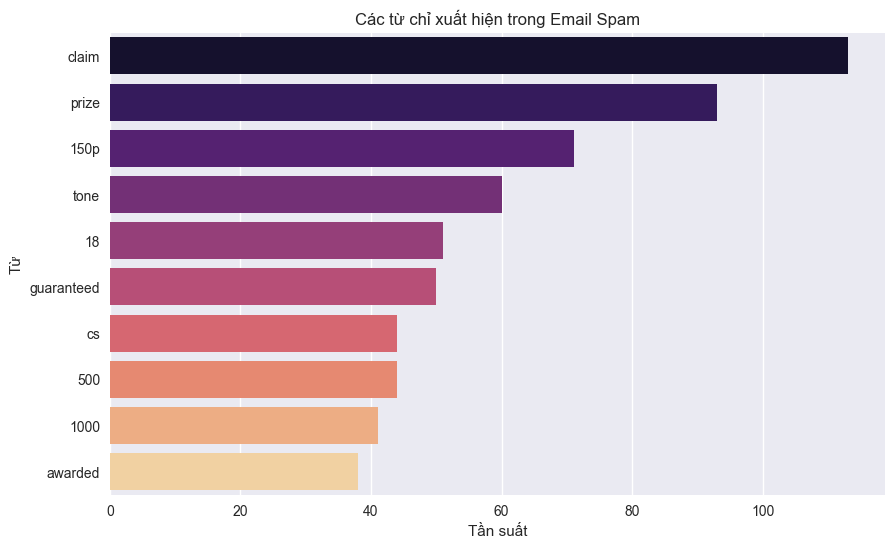


Hình 3. 11 Biểu đồ thể hiện phân phối độ dài câu trước khi xử lý dữ liệu.

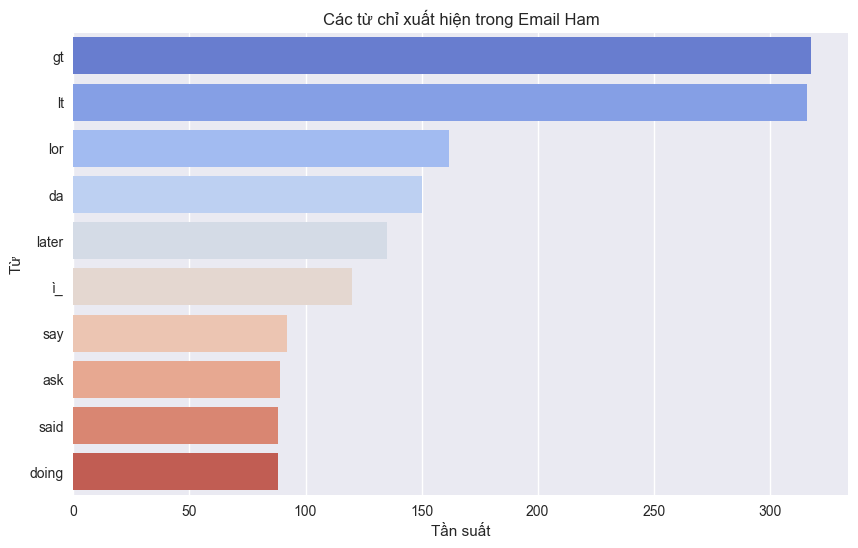
* Ý nghĩa biểu đồ:
* Biểu đồ thể hiện sự phân phối độ dài câu (hoặc số lượng từ) trong tập dữ liệu email, giúp chúng ta hiểu rõ đặc điểm văn bản của các email spam và email hợp lệ (ham).
* Thông qua biểu đồ, có thể quan sát được sự khác biệt về độ dài nội dung giữa các loại email, ví dụ: email spam thường có xu hướng ngắn gọn hoặc lặp lại nhiều từ khóa quảng cáo, trong khi email ham thường có nội dung dài và đa dạng hơn.
* Ứng dụng:
* Biểu đồ giúp phát hiện đặc trưng phân biệt giữa email spam và email hợp lệ, từ đó hỗ trợ quá trình khai phá đặc trưng (feature engineering).
* Dựa vào phân phối độ dài email, ta có thể:
* Bổ sung đặc trưng độ dài email (số ký tự hoặc số từ) vào mô hình học máy để cải thiện độ chính xác.
* Xác định ngưỡng độ dài bất thường, hỗ trợ nhận diện các email spam có nội dung quá ngắn hoặc quá dài.
* Biểu đồ cũng giúp đánh giá chất lượng dữ liệu, phát hiện các email trống, email quá ngắn hoặc nhiễu cần được xử lý trong bước tiền xử lý dữ liệu.
* Ngoài ra, việc hiểu rõ phân phối độ dài email giúp lựa chọn phương pháp biểu diễn văn bản phù hợp (BoW, TF-IDF), từ đó nâng cao hiệu quả của mô hình phân loại.

3.3.3. Trực quan các từ thường xuất hiện trong email

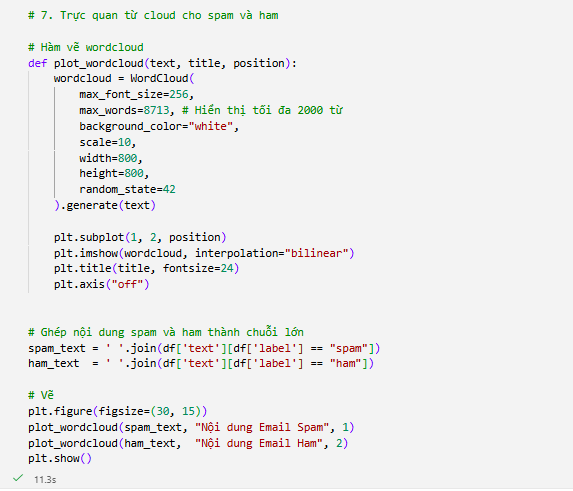




Hình 3. 12: Các từ xuất hiện nhiều trong email spam.



Hình 3. 13: Các từ xuất hiện nhiều trong email ham.





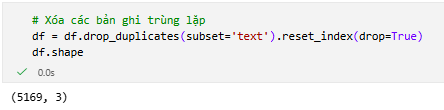
Hình 3. 14: Biểu đồ trực quan bằng Wordcloud.

* **Ý nghĩa:**
* Biểu đồ trực quan hóa các từ xuất hiện với tần suất cao trong email spam và email hợp lệ (ham), giúp làm rõ sự khác biệt về mặt ngữ nghĩa và nội dung giữa hai loại email.
* Thông qua biểu đồ, có thể dễ dàng nhận thấy rằng:
* Email spam thường chứa các từ mang tính quảng cáo, khuyến mãi, lôi kéo hành động (ví dụ: free, win, won, prize, money, mobile).
* Email ham thường chứa các từ mang tính trao đổi thông tin, công việc hoặc cá nhân, phản ánh nội dung giao tiếp bình thường.
* **Ứng dụng:**
* Biểu đồ giúp xác định các từ khóa đặc trưng cho từng lớp, từ đó hỗ trợ quá trình xây dựng và lựa chọn đặc trưng trong mô hình học máy.
* Các từ xuất hiện nhiều trong spam có thể được sử dụng để:
* Nâng cao hiệu quả của các phương pháp biểu diễn văn bản như Bag of Words (BoW) và TF-IDF.
* Tăng khả năng phân biệt giữa email spam và ham khi huấn luyện mô hình phân loại.
* Biểu đồ cũng giúp kiểm tra và đánh giá chất lượng tiền xử lý dữ liệu, chẳng hạn như loại bỏ từ dừng (stopwords) hoặc chuẩn hóa từ.
* Ngoài ra, việc so sánh các từ phổ biến giữa hai lớp giúp giải thích được kết quả của mô hình, tăng tính minh bạch và khả năng diễn giải (interpretability).

3.4. Chuẩn bị dữ liệu

3.4.1. Xử lý dữ liệu

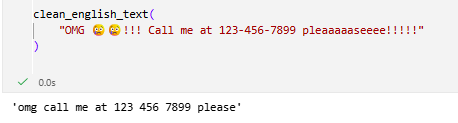
* Xóa các bản ghi trùng lặp



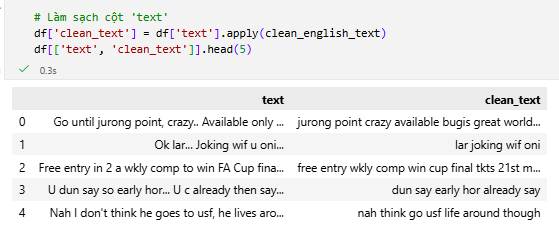
Hình 3. 15: Dữ liệu còn lại sau khi khóa bản ghi trùng lặp.

* Xử lý dữ liệu tin nhắn email.
* Kiểm tra dữ liệu đầu vào
* Chuyển về chữ thường (Lowercase)
* Loại bỏ ký tự đặc biệt & emoji.
* Giảm ký tự lặp.
* Chuẩn hóa khoảng trắng.
* Tách từ (Tokenization).
* Trả về chuỗi hoàn chỉnh





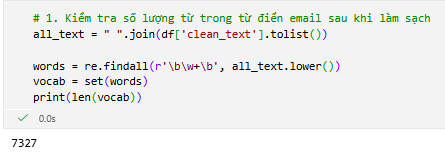
Hình 3. 16: Xử lý tin nhắn email.



Hình 3. 17: Dữ liệu sau khi xử lý.

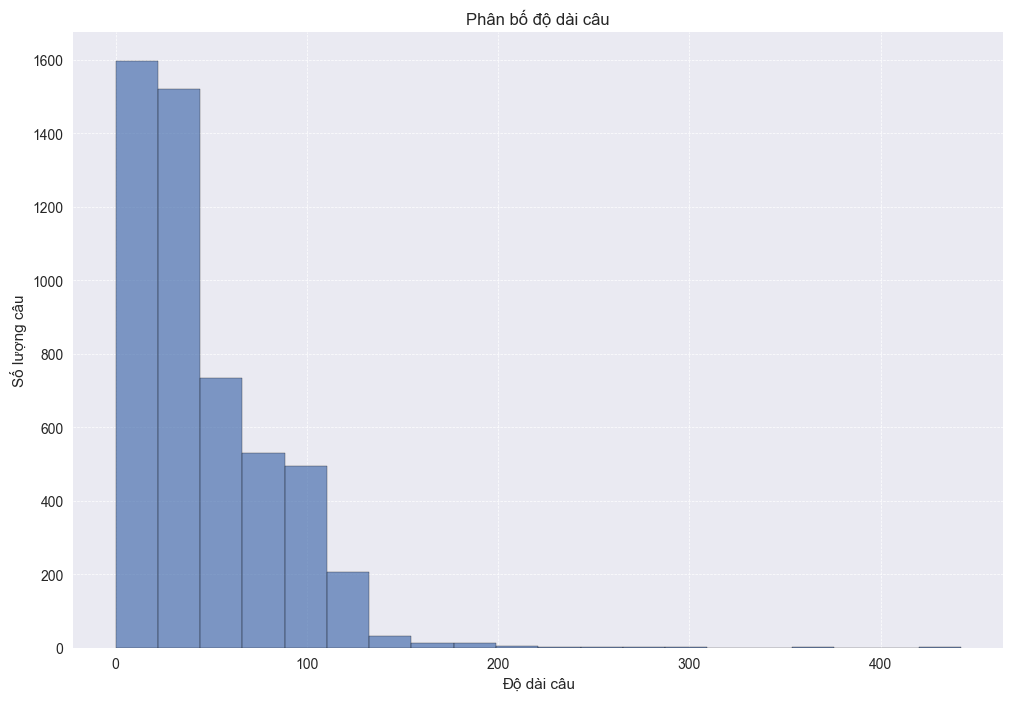
3.4.2. Trực quan dữ liệu sau khi xử lý

* Lượng từ điển còn lại



Hình 3. 18: Lượng từ còn lại trong từ điển.

* Phân bố độ dài câu sau khi làm sạch dữ liệu



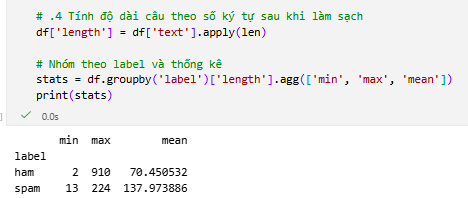
Hình 3. 19: Phân bố độ dài câu sau khi làm sạch dữ liệu

* Nhận xét: Độ dài câu giảm đáng kẻ so với ban đầu.
* Trực quan từ cloud cho spam và ham sau khi làm sạch



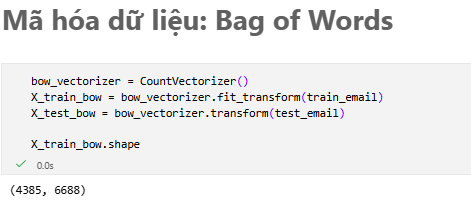
Hình 3. 20: Trực quan từ cloud cho spam và ham sau khi làm sạch

* Nhận xét: Các từ khóa giúp phân loại email được nổi bật rõ ràng
* Tính độ dài câu theo số ký tự làm sạch



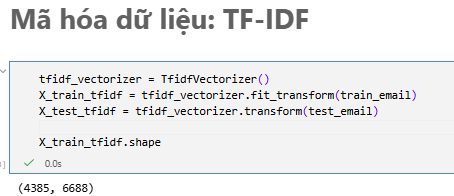
Hình 3. 21: Tính độ dài câu theo số ký tự làm sạch

3.4.3. Chuẩn hóa dữ liệu



Hình 3. 22: Mã hóa dữ liệu bằng phương pháp Bag of Words.

* Tác dụng
* Biến văn bản thành **dạng số** dựa trên **tần suất xuất hiện của từ.**
* Mỗi từ trong từ điển là một đặc trưng (feature), giá trị là số lần từ đó xuất hiện trong email.
* Ý nghĩa trong phân loại email spam
* Giúp mô hình học được các **từ khóa thường gặp trong spam** (ví dụ: free, win, offer).
* Phù hợp với các mô hình đơn giản như **Naive Bayes** và **Logistic Regression.**
* Hạn chế
* Không phản ánh mức độ quan trọng của từ.
* Các từ phổ biến xuất hiện nhiều trong cả spam và ham có thể gây nhiễu.



Hình 3. 23: Mã hóa dữ liệu bằng phương pháp TF-IDF.

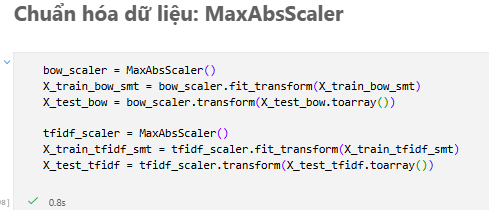
* Tác dụng
* Là phiên bản **nâng cao của Bag of Words.**
* Giảm trọng số của các từ xuất hiện quá thường xuyên, tăng trọng số cho các từ **mang tính phân biệt cao.**
* Ý nghĩa trong phân loại email spam
* Giúp mô hình **tập trung vào các từ khóa đặc trưng cho spam,** thay vì các từ thông dụng.
* Thường cho kết quả **chính xác hơn BoW,** đặc biệt với Logistic Regression hoặc SVM.



Hình 3. 24: Dùng Smote để xử lý mất cân bằng dữ liệu.

Tác dụng

* Xử lý **mất cân bằng dữ liệu** bằng cách **tạo thêm dữ liệu giả cho lớp thiểu số** (ví dụ: spam hoặc ham).
* Không sao chép dữ liệu cũ mà tạo mẫu mới dựa trên các điểm lân cận.
* Ý nghĩa trong phân loại email spam
* Giúp mô hình **học tốt hơn lớp ít dữ liệu,** tránh thiên lệch về lớp chiếm đa số.
* Cải thiện **Recall và F1-score** cho lớp thiểu số.

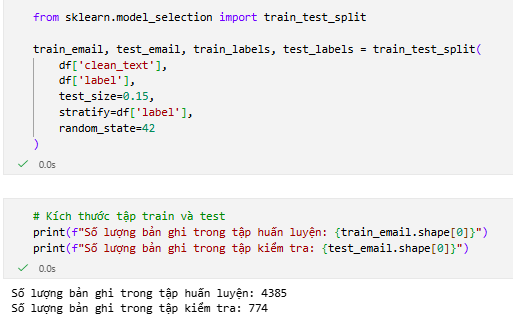


Hình 3. 25: Dùng MaxAbsScaler để chuẩn hóa dữ liệu.

* Tác dụng
* Chuẩn hóa dữ liệu về khoảng **[-1, 1]** dựa trên giá trị tuyệt đối lớn nhất.
* **Không làm mất tính thưa (sparse)** của dữ liệu BoW/TF-IDF.
* Ý nghĩa trong phân loại email spam
* Giúp các đặc trưng có **cùng thang đo**, hỗ trợ các mô hình nhạy với giá trị đầu vào như:
  + Logistic Regression
  + SVM
* Tăng tốc độ hội tụ và ổn định quá trình huấn luyện.

3.4.4. Chia dữ liệu huấn luyện và kiểm tra

* Chia dữ liệu thành:
  + 85% để huấn luyện.
  + 15% để kiểm tra.



3.5. Xây dựng và đánh giá mô hình

3.5.1. Hàm đánh giá mô hình

Chức năng chính của hàm:

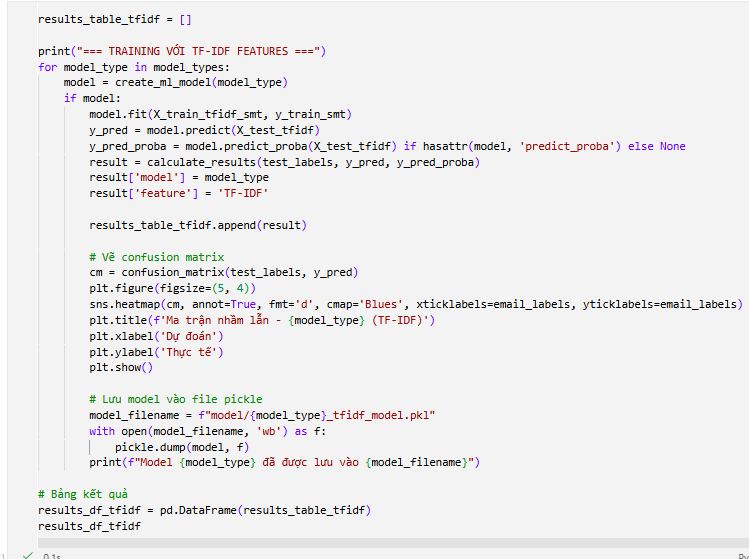
* Dự đoán nhãn mục tiêu dựa trên tập kiểm tra (X\_test).
* Hiển thị các chỉ số đánh giá hiệu suất (precision, recall, F1-score, ...).
* Tạo và trực quan hóa ma trận nhầm lẫn (confusion matrix).

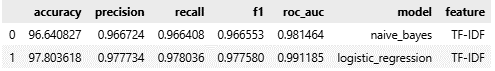
3.5.2. Huấn luyện mô hình



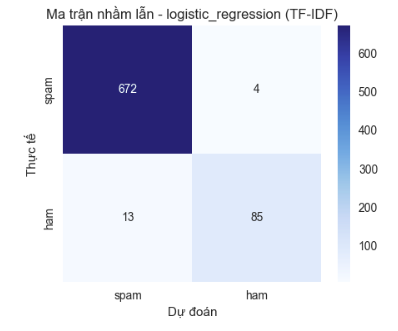


3.5.3. Huấn luyện với phương pháp TF-IDF .





Hình 3. 26: Kết quả của 2 mô hình chạy với TF-IDF.



Hình 3. 27: Mã trận nhầm lẫn Logistic Regression với TF-IDF.

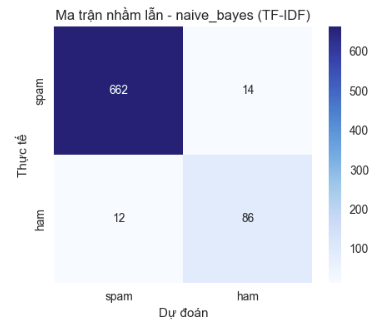
**Với mô hình logistic regression**

* False Negatives (Spam bị phân loại nhầm thành Ham): rất thấp (4 mẫu)
* False Positives (Ham bị phân loại nhầm thành Spam): thấp (13 mẫu)
* Sai sót phân loại là **không đáng kể,** đặc biệt quan trọng trong bài toán lọc spam vì hạn chế việc bỏ sót email spam.

**Phân tích:**

* Accuracy ≈ 97.8%: Mô hình đạt độ chính xác tổng thể rất cao, chứng tỏ khả năng dự đoán đúng trên toàn bộ tập dữ liệu là tốt.
* Precision ≈ 97.7%: Khi mô hình dự đoán là spam, khả năng dự đoán đúng là rất cao (ít spam giả).,
* Recall ≈ 97.8%: Mô hình nhận diện được gần như toàn bộ các email spam thực sự (chỉ bỏ sót 4 trường hợp).
* Cả precision và recall đều cao, cho thấy **mô hình hoạt động ổn định và đáng tin cậy.**
* F1-score ≈ 97.75%: F1-score cao chứng tỏ mô hình đạt sự cân bằng rất tốt giữa precision và recall, không thiên lệch về một phía.

**Ma trận nhầm lẫn:**



Hình 3. 28: Mã trận nhầm lẫn Naïve Bays với TF-IDF.

* **Với mô hình naïve bayes**
* False Negatives (Spam bị phân loại nhầm thành Ham): rất thấp (14 mẫu)
* False Positives (Ham bị phân loại nhầm thành Spam): thấp (12 mẫu)

**Độ chính xác:**

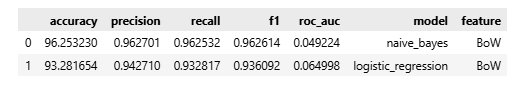
* Accuracy ≈ 96.6%: Mô hình dự đoán đúng trên hầu hết các email, đạt độ chính xác tổng thể cao.
* Precision ≈ 96.7%: Khi mô hình dự đoán là spam, khả năng dự đoán đúng cao, spam giả ít.
* Recall ≈ 96.6%: Mô hình nhận diện được hầu hết các email spam thực sự, chỉ bỏ sót 14 trường hợp.
* F1-score ≈ 96.7%: Chỉ số F1 cao, cân bằng tốt giữa precision và recall, mô hình ổn định.
* ROC AUC ≈ 98.15%: Mô hình phân biệt spam và ham rất hiệu quả.

**Kết luận:**

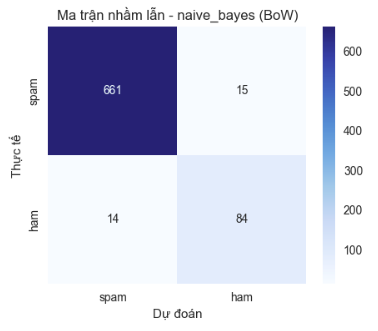
* Mô hình Logistic Regression kết hợp TF-IDF cho kết quả rất tốt:
* Độ chính xác cao
* Khả năng phân biệt giữa spam và ham rõ ràng
* Sai số thấp, cân bằng tốt giữa các chỉ số đánh giá

3.5.4. Huấn luyện với phương pháp BOW.





Hình 3. 29: Kết quả của 2 mô hình chạy với BOW.



Hình 3. 30: Mã trận nhầm lẫn Naïve Bays với BOW

**Đánh giá mô hình Naive Bayes (BoW)**

**Ma trận nhầm lẫn**

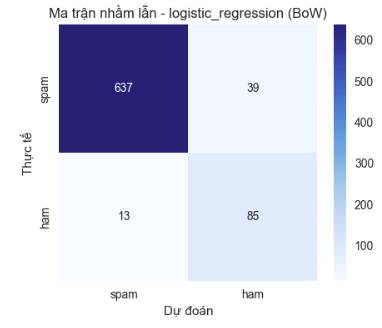
* **True Positives (TP)** cho lớp 1 là 661 và **False Positives (FP)** là 14, cho thấy mô hình phân loại rất chính xác lớp 1.
* **True Negatives (TN)** cho lớp 0 là 84 trong khi **False Negatives (FN)** là 14, cho thấy mô hình phân loại lớp 0 cũng rất tốt.
* Mô hình Naive Bayes với BoW phân loại khá tốt, tuy nhiên số lượng **spam bị bỏ sót (FN= 15)** vẫn còn tồn tại.

**Độ chính xác:**

* Accuracy ≈ 96.25%: Độ chính xác tổng thể cao, cho thấy mô hình hoạt động ổn định trên tập dữ liệu.
* Precision ≈ 96.27%: Khi dự đoán email là spam, mô hình có độ tin cậy cao.
* Recall ≈ 96.25%: Mô hình nhận diện phần lớn email spam, nhưng vẫn bỏ sót một số trường hợp.
* Precision và recall khá cân bằng, phù hợp cho bài toán phân loại spam cơ bản.
* F1-score ≈ 96.26%: F1-score cao cho thấy mô hình có sự cân bằng tốt giữa precision và recall.

**Nhận xét chung**

* Lỗi **False Negatives (15)** cho thấy một số email spam vẫn bị phân loại nhầm thành ham.
* Tuy nhiên, tổng số lỗi phân loại không nhiều và mô hình vẫn đạt hiệu quả tốt.



Hình 3. 31: Mã trận nhầm lẫn Logistic Regression với BOW.

**Đánh giá mô hình Logistic Regression (BoW)**

**Ma trận nhầm lẫn:**

* True Positives (TP) cho lớp 1 là 637 và False Positives (FP) là 13, cho thấy mô hình phân loại rất chính xác lớp 1.
* True Negatives (TN) cho lớp 0 là 85 trong khi False Negatives (FN) là 39, cho thấy mô hình phân loại lớp 0 cũng rất tốt.

**Độ chính xác:**

* Accuracy ≈ 93.28%
* Độ chính xác thấp hơn đáng kể so với Naive Bayes khi cùng sử dụng BoW.
* Precision ≈ 94.27%khi dự đoán email là spam, mô hình có độ tin cậy cao.
* Recall ≈ 93.28%mô hình nhận diện phần lớn email spam, nhưng vẫn bỏ sót một số trường hợp.
* Recall thấp hơn cho thấy mô hình chưa nhận diện tốt các email spam, đặc biệt là các trường hợp khó phân biệt.
* F1-score ≈ 93.61%
* F1-score thấp hơn cho thấy sự cân bằng giữa precision và recall chưa tối ưu.

**Nhận xét chung**

* False Negatives cao (39) là điểm yếu lớn trong bài toán lọc spam vì có thể gây bỏ sót email spam nguy hiểm.
* Logistic Regression với BoW không tận dụng tốt đặc trưng so với TF-IDF.

**Kết luận:**

Mô hình Logistic Regression và mô hình Naïve Bays này có hiệu suất rất cao, đạt độ chính xác 93% và 96%. Sự phân biệt giữa các lớp 0 và 1 được thực hiện rất tốt với ít sai sót.

3.6. So sánh các mô hình dựa trên bài toán

****3.6.1. Tiêu chí so sánh****

Dựa trên bài toán phân loại nấm ăn được (Edible - 1) và không ăn được (Inedible - 0), các mô hình Logistic Regression, Decision Tree, và KNN được so sánh cụ thể qua các tiêu chí sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mô hình** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Ghi chú** |
| Naïve Bayes | 96.25% | 96.27% | 96.25% | 96.26% | BOW |
| Logistic Regression | 93.28% | 94.27% | 93.28% | 93.60% |
| Naïve Bayes | 96.64% | 96.67% | 96.64% | 95.65% | TF-IDF |
| Logistic Regression | 97.80% | 97.77% | 97.75% | 97.75% |

3.6.2. So sánh trực tiếp

a. Accuracy (Độ chính xác tổng thể)

* Naïve Bayes (BOW) đạt Accuracy 96.25%, mô hình dự đoán đúng phần lớn email, hiệu suất khá tốt.
* Logistic Regression (BOW) đạt Accuracy 93.28%, thấp hơn Naïve Bayes, phản ánh khả năng dự đoán tổng thể kém hơn một chút.
* Naïve Bayes (TF-IDF) đạt Accuracy 96.64%, cải thiện nhẹ so với BOW.
* Logistic Regression (TF-IDF) đạt Accuracy cao nhất 97.80%, cho thấy mô hình này với TF-IDF dự đoán chính xác hầu hết các email.

b. Precision (Độ chính xác khi dự đoán spam)

* Naïve Bayes (BOW) đạt Precision 96.27%, khi dự đoán spam mô hình rất chính xác, ít báo động giả.
* Logistic Regression (BOW) Precision 94.27%, thấp hơn Naïve Bayes, vẫn chấp nhận được.
* Naïve Bayes (TF-IDF) Precision 96.67%, tăng nhẹ so với BOW.
* Logistic Regression (TF-IDF) Precision 97.77%, rất cao, cho thấy hầu hết email dự đoán là spam đều đúng.

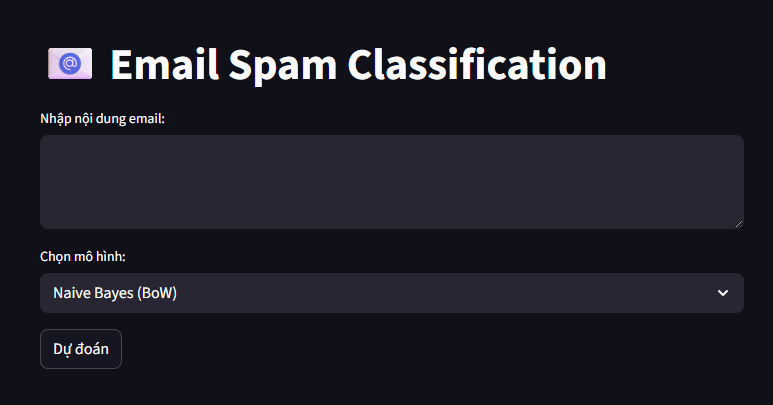
c. Recall (Độ nhạy, khả năng phát hiện spam thực sự)

* Naïve Bayes (BOW) Recall 96.25%, hầu hết spam thực sự được nhận diện chính xác.
* Logistic Regression (BOW) Recall 93.28%, bỏ sót spam nhiều hơn Naïve Bayes.
* Naïve Bayes (TF-IDF) Recall 96.64%, cải thiện nhẹ so với BOW.
* Logistic Regression (TF-IDF) Recall 97.75%, mô hình nhận diện gần như toàn bộ spam thực sự.

d. F1-Score (Cân bằng Precision và Recall)

* Naïve Bayes (BOW) F1-Score 96.26%, cân bằng tốt giữa Precision và Recall.
* Logistic Regression (BOW) F1-Score 93.60%, cân bằng kém hơn nhưng vẫn chấp nhận được.
* Naïve Bayes (TF-IDF) F1-Score 95.65%, mô hình ổn định, cân bằng tốt.
* Logistic Regression (TF-IDF) F1-Score 97.75%, cân bằng xuất sắc, cho thấy mô hình TF-IDF hoạt động tốt nhất.
* **Tóm lại:**
* Mô hình Logistic Regression với TF-IDF đạt hiệu suất cao nhất, cân bằng tốt giữa precision và recall.
* Naïve Bayes cũng hoạt động ổn định, đặc biệt với TF-IDF.
* Sử dụng TF-IDF cải thiện hiệu suất so với BOW, đặc biệt với Logistic Regression.CHƯƠNG 4: XÂY DỰNG Streamlit

4.1. Cài đặt Streamlit



Hình 4. 1 Giao diện Demo chính

* **Các thành phần nhập liệu**:
  + Nội dung tin nhắn email: Người dùng nhập nội dung tin nhắn
  + Chọn mô hình: Naïve Bays (Bow), Naïve Bayes (TF-IDF), Logistic Regression (Bow), Logistic Regression(TF-IDF).

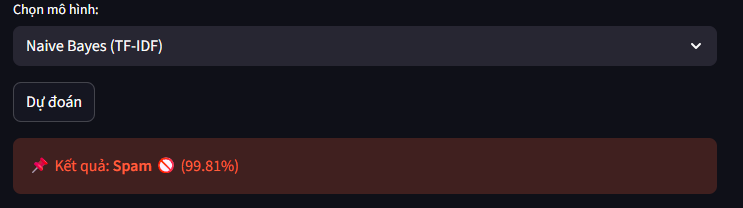
• Khi người dùng nhấn nút "Dự đoán", dữ liệu email được nhập từ trường văn bản sẽ được thu thập và gửi đến backend của ứng dụng Streamlit.

• Tiến hành tiền xử lý và vector hóa email dựa trên mô hình được chọn (Naive Bayes hoặc Logistic Regression), sau đó thực hiện dự đoán.

Hiển thị kết quả:

• Kết quả dự đoán sẽ được hiển thị ngay dưới form nhập liệu, với thông tin rõ ràng về email:

* Email được dự đoán là SPAM hay HAM.
* Xác suất dự đoán, ví dụ: "Email SPAM (Xác suất: 92%)" hoặc "Email HAM (Xác suất: 85%)".



Hình 4. 2 Kết quả dự đoán

KẾT LUẬN

**Kết quả đạt được:**

Trong bài toán phân loại email thành spam và ham, nghiên cứu đã triển khai và đánh giá hai mô hình học máy phổ biến là Naive Bayes và Logistic Regression. Kết quả cho thấy:

* Naive Bayes thể hiện khả năng xử lý tốt với dữ liệu văn bản, đặc biệt khi các từ xuất hiện độc lập với nhau, mô hình này có tốc độ huấn luyện nhanh và hiệu quả cao với tập dữ liệu nhỏ.
* Logistic Regression đạt hiệu suất phân loại ổn định, có khả năng tổng quát hóa tốt và linh hoạt với các tập dữ liệu lớn hơn, đặc biệt khi kết hợp với các kỹ thuật tiền xử lý như Bag of Words hoặc TF-IDF.

**Hạn chế của đề tài:**

Cả hai mô hình đều đạt hiệu quả đáng kể trong việc phân loại email, giúp phát hiện spam và bảo vệ người dùng trước các thư không mong muốn. Tuy nhiên, hiệu suất của các mô hình vẫn phụ thuộc vào **chất lượng tiền xử lý dữ liệu**, chẳng hạn như loại bỏ từ dừng, chuẩn hóa văn bản và vector hóa đặc trưng.

**Hướng phát triển trong tương lai:**

Trong tương lai, nghiên cứu có thể được mở rộng bằng việc thử nghiệm các mô hình học sâu (Deep Learning) như mạng nơ-ron tích chập (CNN) hoặc mô hình biến áp (Transformer), nhằm nâng cao độ chính xác và khả năng phát hiện các mẫu spam phức tạp hơn.

TÀI LIỆU THAM KHẢO

[1] Matplotlib: [Giới thiệu về Matplotlib - Nguyen Van Hoang (viblo.asia)](https://viblo.asia/p/gioi-thieu-ve-matplotlib-mot-thu-vien-rat-huu-ich-cua-python-dung-de-ve-do-thi-yMnKMN6gZ7P)

[2] Pandas: [Giới thiệu về Pandas - Nguyen Van Hoang (viblo.asia)](https://viblo.asia/p/gioi-thieu-ve-pandas-mot-thu-vien-pho-bien-cua-python-cho-viec-phan-tich-du-lieu-aWj53Nnel6m)

[3] Numpy: [Giới thiệu về Numpy - Nguyen Van Hoang (viblo.asia)](https://viblo.asia/p/gioi-thieu-ve-numpy-mot-thu-vien-chu-yeu-phuc-vu-cho-khoa-hoc-may-tinh-cua-python-maGK7kz9Kj2)

[4] [Gradient Boosting - Tất tần tật về thuật toán mạnh mẽ nhất trong Machine](https://viblo.asia/p/gradient-boosting-tat-tan-tat-ve-thuat-toan-manh-me-nhat-trong-machine-learning-YWOZrN7vZQ0)

[Learning - Bui Tien Tung (viblo.asia)](https://viblo.asia/p/gradient-boosting-tat-tan-tat-ve-thuat-toan-manh-me-nhat-trong-machine-learning-YWOZrN7vZQ0)

[5] N. V. Hậu, P. M. Chuẩn, và N. V. Quyết, “Giáo trình Học máy cơ bản”, Nhà

      xuất bản Khoa học và Kỹ thuật, 2022.