E-mail Categorization using KNN, Gaussian Naive Bayes, and Logistic Regression

Báo cáo môn Machine Learning

Nguyễn Duy Vũ - 22000132

Vũ Đức Duy - 22000078

Đoàn Phạm Ngọc Linh - 22000102

Giảng viên hướng dẫn: Thầy Cao Văn Chung

Trường Đại học Khoa học Tự nhiên – ĐHQGHN
Khoa Toán – Cơ – Tin học
Ngành Toán Tin
Học phần: Machine Learning

Hà Nội, tháng 4 năm 2025

Contents

1 Tóm tắt

Báo cáo này tập trung vào việc xây dựng hệ thống phân loại email tự động dựa trên ba thuật toán học máy: **K-Nearest Neighbors (KNN)**, **Gaussian Naive Bayes** và **Logistic Regression**. Mục tiêu là phân loại email vào các nhóm chủ đề như công việc, quảng cáo, spam và cá nhân, dựa trên các đặc trưng văn bản như tần suất từ khóa, độ dài và thông tin metadata (người gửi, tiêu đề).

Dữ liệu sử dụng là tập email công khai. Quá trình tiền xử lý bao gồm tách từ, loại bỏ từ dừng, chuẩn hóa . Dữ liệu được chia thành hai phần: huấn luyện (80%) và kiểm thử (20%), đảm bảo cân bằng giữa các lớp.

Kết quả thực nghiệm:

Mô hình	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.9720	0.9320	0.9730	0.9521
K-Nearest Neighbors	0.9063	0.7639	0.9730	0.8559
Gaussian Naive Bayes	0.6986	0.4675	0.3885	0.4244

Dựa trên kết quả thực nghiệm, **Logistic Regression** cho hiệu suất vượt trội với độ chính xác và F1 Score cao nhất, cho thấy mô hình này phù hợp nhất cho bài toán phân loại email. **K-Nearest Neighbor** có khả năng nhận diện (Recall) tốt nhưng độ chính xác tổng thể và F1 Score thấp hơn, trong khi **Gaussian Naive Bayes** cho kết quả thấp nhất trên mọi chỉ số. Như vậy, Logistic Regression là lựa chọn tối ưu cho bài toán này.

2 Giới thiệu đề tài

2.1 Mục tiêu và bài toán phân loại email

Phân loại email là một bài toán quan trọng trong lĩnh vực học máy và xử lý ngôn ngữ tự nhiên. Nó đóng vai trò then chốt trong việc xây dựng các hệ thống tự động hỗ trợ người dùng quản lý hộp thư một cách hiệu quả. Với mục tiêu phân loại email thành các nhóm chủ đề (chẳng hạn như công việc, quảng cáo, spam, cá nhân), bài toán này không chỉ liên quan đến lý thuyết phân loại mà còn có nhiều ứng dụng thực tế trong môi trường làm việc hiện đại.

Mục tiêu chính của bài toán phân loại email sử dụng mô hình học máy là xác định nhóm chủ đề của một email dựa trên các đặc trưng văn bản, bao gồm nội dung, tiêu đề, người gửi, và các thông tin metadata khác. Với sự phát triển của các kỹ thuật học máy và xử lý ngôn ngữ tự nhiên, việc xây dựng các hệ thống phân loại email tự động trở nên ngày càng khả thi và hiệu quả. Những hệ thống này có thể giúp người dùng tiết kiệm thời gian, phát hiện email quan trọng hoặc lọc bỏ thư rác một cách tự động.

2.2 Ứng dụng thực tế

Phân loại email có thể được ứng dụng trong nhiều lĩnh vực, đặc biệt là trong các hệ thống quản lý thông tin và truyền thông. Dưới đây là một số ứng dụng tiêu biểu:

- **Lọc thư rác (spam filter)**: Hệ thống phân loại có thể phát hiện và đưa các email rác vào thư mục riêng, giúp người dùng không bị làm phiền và tránh nguy cơ lừa đảo.
- **Tự động gắn nhãn và sắp xếp email**: Các email có thể được tự động phân loại thành các nhóm như công việc, cá nhân, quảng cáo,... giúp tối ưu việc tìm kiếm và phản hồi email.
- **Trợ lý ảo và chatbot**: Các trợ lý thông minh có thể sử dụng hệ thống phân loại email để ưu tiên các thư quan trọng hoặc đưa ra phản hồi tự động phù hợp với nội dung email.

- **Phân tích hành vi người dùng và hỗ trợ chăm sóc khách hàng**: Trong các tổ chức lớn, việc phân tích nội dung email đến có thể giúp hiểu rõ hơn về nhu cầu người dùng và cải thiện dịch vụ chăm sóc khách hàng.

2.3 Các phương pháp được lựa chọn

Trong nghiên cứu này, chúng tôi lựa chọn ba phương pháp học máy phổ biến để giải quyết bài toán phân loại email. Mỗi phương pháp đều có những đặc điểm riêng biệt và phù hợp với từng loại dữ liệu hoặc mục tiêu phân tích.

- **K-Nearest Neighbors (KNN)**: KNN là một thuật toán phân loại dựa trên việc so sánh điểm dữ liệu cần phân loại với các điểm lân cận gần nhất trong tập huấn luyện. Trong bài toán phân loại văn bản, KNN hoạt động hiệu quả khi sử dụng các đặc trưng như vector TF-IDF và khoảng cách cosine.

- **Gaussian Naive Bayes (GNB)**: Naive Bayes là một phương pháp phân loại xác suất đơn giản nhưng hiệu quả, giả định tính độc lập giữa các đặc trưng. Mặc dù giả định này không hoàn toàn đúng với dữ liệu văn bản, Gaussian Naive Bayes vẫn cho kết quả tốt khi dữ liệu có phân phối tương đối chuẩn và số chiều lớn.

- **Logistic Regression (LR)**: Logistic Regression là một mô hình phân loại tuyến tính mạnh mẽ, đặc biệt hiệu quả trong bài toán phân loại văn bản có đặc trưng thưa (sparse features). Nhờ khả năng xử lý tốt dữ liệu nhiều chiều và dễ diễn giải, Logistic Regression thường được sử dụng như một mô hình baseline cho nhiều bài toán phân loại.

Các phương pháp trên được lựa chọn vì tính phổ biến, dễ triển khai và hiệu quả đã được kiểm chứng trong các bài toán phân loại văn bản. Bằng cách áp dụng cả ba phương pháp, chúng tôi hướng đến việc so sánh và đánh giá hiệu suất phân loại dựa trên các tiêu chí như độ chính xác (accuracy), độ nhạy (recall), độ đặc hiệu (specificity) và diện tích dưới đường cong ROC (AUC).

Việc lựa chọn ba phương pháp KNN, Gaussian Naive Bayes và Logistic Regression không chỉ dựa trên tính phổ biến mà còn vì mỗi phương pháp đại diện cho một hướng tiếp cận khác nhau trong học máy:

- **KNN** đại diện cho nhóm thuật toán phi tham số, không xây dựng mô hình huấn luyện rõ ràng mà dựa trên khoảng cách trong không gian đặc trưng để phân loại. Điều này giúp mô hình đơn giản, dễ hiểu và phù hợp với dữ liệu có độ tương đồng cao.

- **Gaussian Naive Bayes** là một mô hình xác suất dựa trên định lý Bayes và giả định tính độc lập giữa các đặc trưng. Phương pháp này đặc biệt hiệu quả với dữ liệu có phân phối gần chuẩn và mang lại tốc độ huấn luyện rất nhanh, thích hợp với dữ liệu có nhiều chiều như văn bản.

- **Logistic Regression** là một mô hình tuyến tính có khả năng học trọng số của các đặc trưng để tối ưu hóa phân loại. Phương pháp này thường hoạt động tốt trên dữ liệu thưa, có tính tuyến tính, và là một trong những baseline đáng tin cậy trong phân loại văn bản.

Bằng việc kết hợp ba phương pháp thuộc ba hướng tiếp cận khác nhau – dựa trên khoảng cách (KNN), xác suất (Naive Bayes), và mô hình tuyến tính (Logistic Regression) – nghiên cứu nhằm mục tiêu đưa ra đánh giá khách quan, toàn diện hơn về hiệu quả của từng mô hình đối với bài toán phân loại email. Ngoài ra, cả ba phương pháp đều có ưu điểm là dễ cài đặt, thời gian huấn luyện ngắn, và phù hợp với tập dữ liệu có quy mô vừa, giúp đảm bảo tính khả thi trong phạm vi nghiên cứu.

3 Tổng quan về các phương pháp

3.1 K-Nearest Neighbors (KNN)

3.1.1 Giới thiệu

K-Nearest Neighbors (KNN) là một phương pháp phân loại đơn giản nhưng mạnh mẽ. Phương pháp này không yêu cầu quá trình huấn luyện phức tạp mà thay vào đó, sử dụng toàn bộ dữ liệu huấn luyện để đưa ra quyết định phân loại dựa trên những điểm dữ liệu gần nhất.

3.1.2 Nguyên lý hoạt động

KNN hoạt động bằng cách tìm kiếm k điểm dữ liệu gần nhất với điểm cần phân loại trong không gian đặc trưng. Dự đoán của mô hình sẽ dựa trên đa số nhãn của các điểm gần nhất.

Công thức tính khoảng cách phổ biến nhất là khoảng cách Euclidean:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Trong đó x và y là hai điểm trong không gian đặc trưng với các giá trị thuộc n-chiều.

3.1.3 Ưu điểm và nhược điểm

- **Ưu điểm:**
  - Đơn giản, dễ hiểu và triển khai.
  - Không yêu cầu quá trình huấn luyện.
  - Hoạt động tốt với các dữ liệu không có quá nhiều nhiễu.
- **Nhược điểm:**
  - Chi phí tính toán cao khi dữ liệu lớn.
  - Kém hiệu quả với các dữ liệu có không gian đặc trưng cao.

– Kết quả dự đoán phụ thuộc vào giá trị của k, nếu chọn k không phù hợp sẽ dẫn đến kết quả không chính xác.

3.2 Gaussian Naive Bayes (GNB)

3.2.1 Giới thiệu

Gaussian Naive Bayes (GNB) là một phương pháp phân loại dựa trên lý thuyết Bayes, trong đó giả định rằng các đặc trưng trong dữ liệu độc lập với nhau và có phân phối Gaussian (phân phối chuẩn)

3.2.2 Nguyên lý hoạt động

Phương pháp Naive Bayes sử dụng định lý Bayes để tính xác suất của mỗi lớp dựa trên các đặc trưng của dữ liệu:

$$p(k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid k)p(k)}{p(\mathbf{x})} \propto p(\mathbf{x} \mid k)p(k)$$

Giả thiết Naive cho rằng các đặc trưng độc lập với nhau:

$$p(\mathbf{x} \mid k) = \prod_{i=1}^{d} p(x_i \mid k)$$

Với Gaussian Naive Bayes, mỗi đặc trưng x_i tuân theo phân phối chuẩn trong mỗi lớp k:

$$p(x_i \mid k) = \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} \exp\left(-\frac{(x_i - \mu_{ki})^2}{2\sigma_{ki}^2}\right)$$

Tham số μ_{ki}, σ_{ki}^2 được ước lượng từ dữ liệu huấn luyện[**?**].

3.2.3 Ưu điểm và nhược điểm

- **Ưu điểm:**
  - Đơn giản, nhanh chóng và hiệu quả, đặc biệt với bộ dữ liệu lớn.
  - Hoạt động tốt khi các đặc trưng độc lập và có phân phối chuẩn.

- **Nhược điểm:**
  - Giả định độc lập giữa các đặc trưng thường không thực tế.
  - Kém hiệu quả khi các đặc trưng có mối quan hệ mạnh mẽ với nhau.

3.3 Logistic Regression (LR)

3.3.1 Giới thiệu

Logistic Regression (LR) là một phương pháp phân loại phổ biến, mặc dù tên gọi là ”hồi quy”, nhưng nó thực chất là một thuật toán phân loại. LR được sử dụng để dự đoán xác suất của một lớp dựa trên các đặc trưng đầu vào

3.3.2 Nguyên lý hoạt động

Logistic Regression sử dụng hàm sigmoid để chuyển đổi kết quả của mô hình hồi quy tuyến tính thành xác suất:

$$h_\theta(x) = \frac{1}{1 + \exp(-\theta^T x)}$$

Xác suất phân lớp:

$$P(y = 1|x; \theta) = h_\theta(x), \quad P(y = 0|x; \theta) = 1 - h_\theta(x)$$

Tham số θ được ước lượng bằng phương pháp cực đại hóa log-likelihood:

$$\ell(\theta) = \sum_{n=1}^{N} [y_n \log h_\theta(x_n) + (1 - y_n) \log(1 - h_\theta(x_n))]$$

Có thể sử dụng các thuật toán tối ưu như Gradient Descent hoặc Newton-Raphson để tìm

3.3.3 Ưu điểm và nhược điểm

- **Ưu điểm:**
  - Dễ hiểu, dễ triển khai.

  - Hiệu quả với dữ liệu có quan hệ tuyến tính giữa đặc trưng và lớp.
  - Có thể mở rộng cho bài toán đa lớp.

- **Nhược điểm:**
  - Không phù hợp với dữ liệu có quan hệ phi tuyến tính.
  - Nhạy cảm với đa cộng tuyến và ngoại lệ.

4 Dữ liệu và thực nghiệm

4.1 Dữ liệu

Bộ dữ liệu này bao gồm thông tin về phân loại email tự động, với mục tiêu phân loại email theo các nhóm chủ đề khác nhau như công việc, quảng cáo, spam, và cá nhân. Bộ dữ liệu được thu thập từ các email công khai và chứa nhiều đặc trưng văn bản, bao gồm tần suất từ khóa, độ dài email, và các thông tin metadata như người gửi và tiêu đề. Bộ dữ liệu này rất hữu ích cho các nhà nghiên cứu và các chuyên gia trong lĩnh vực học máy và xử lý ngôn ngữ tự nhiên, đặc biệt là trong việc phát triển các mô hình phân loại văn bản.

4.1.1 Cấu trúc dữ liệu

- **Số lượng bản ghi:** 5172 email.
- **Số lượng cột:** 3008.
- **Loại dữ liệu:**
  - Dữ liệu số thực: Tần suất xuất hiện của các từ phổ biến trong email.
  - Dữ liệu số nguyên: Các chỉ số thống kê khác như độ dài văn bản, số lượng từ.
  - Dữ liệu dạng chuỗi: Tên email, tiêu đề,...

4.1.2 Một số cột đáng chú ý

- **Email No. / Email Name:** Mã định danh và tên email, dùng để tra cứu nhưng không được sử dụng trong quá trình huấn luyện mô hình.
- **Most Common Word 1 - 9:** Các cột thể hiện tần suất xuất hiện của 9 từ phổ biến nhất trong email, bao gồm các từ như “the”, “to”, “ect”, “and”, “for”, “of”, “a”, “you”, “hou”. Đây là những đặc trưng quan trọng phản ánh nội dung và ngữ cảnh email.

- **Tần suất từ:** Giá trị trung bình (mean) và độ lệch chuẩn (std) của mỗi từ cho thấy sự phân tán và phân bố khác nhau giữa các email. Ví dụ: từ ”you” có giá trị trung bình là 55.5 và độ lệch chuẩn 87.6, cho thấy tần suất sử dụng rất khác nhau giữa các email.
- **Label:** Nhãn mục tiêu được gán cho từng email để phục vụ cho bài toán phân loại.

4.1.3 Thông tin bổ sung

Tỷ lệ giữa các lớp trong bộ dữ liệu này khá cân bằng, với tỷ lệ spam và không spam gần như đều nhau. Điều này giúp mô hình phân loại có thể học được các đặc trưng chung mà không bị thiên lệch quá mức về một lớp nào đó.

4.2 Tiền xử lý dữ liệu

Bộ dữ liệu này đã được tiền xử lý sẵn với các đặc trưng được trích xuất từ nội dung email. Tuy nhiên, để cải thiện hiệu suất của các mô hình phân loại, chúng tôi đã thực hiện các bước tiền xử lý bổ sung như sau:

- **Loại bỏ các cột không cần thiết:**
  - `Email No.`: Loại bỏ vì đây chỉ là số thứ tự, không có ý nghĩa phân tích.
- **Chuẩn hóa dữ liệu:**
  - Thực hiện chuẩn hóa dữ liệu về phân phối chuẩn với kỳ vọng là 0 và phương sai là 1 sử dụng `StandardScaler`.
  - Áp dụng chuẩn hóa cho tất cả các đặc trưng số để đảm bảo các đặc trưng có cùng thang đo.
- **Giảm chiều dữ liệu:**
  - Áp dụng `PCA (Principal Component Analysis)` để giảm số chiều của dữ liệu từ 3008 xuống còn 774 thành phần chính.

  - Các thành phần chính này giữ lại hầu hết thông tin quan trọng trong dữ liệu gốc, đồng thời giảm đáng kể độ phức tạp của mô hình.

- **Chia dữ liệu:**
  - Chia dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%) để đánh giá hiệu suất của mô hình.
  - Sử dụng phương pháp phân tầng để đảm bảo tỷ lệ email spam và không phải spam được giữ nguyên trong cả hai tập.

Quá trình tiền xử lý này giúp cải thiện hiệu suất của các mô hình phân loại bằng cách loại bỏ thông tin không cần thiết, chuẩn hóa các đặc trưng và giảm độ phức tạp của dữ liệu.

4.3 Trực quan hóa dữ liệu

4.3.1 Trực quan hóa dữ liệu và phân tích thành phần chính

Trong phần này, chúng ta tiến hành phân tích hiệu quả của phương pháp PCA trong việc giảm chiều dữ liệu email để phân loại spam. Đầu tiên, chúng ta phân tích lượng thông tin được giữ lại ở từng thành phần chính thông qua phương sai giải thích.

Hình **??** thể hiện đường cong phương sai tích lũy của dữ liệu email.

- **Phân bố dữ liệu:** Biểu đồ cho thấy cách các email được phân bố trong không gian hai chiều của hai thành phần chính đầu tiên. Điểm màu xanh đại diện cho email không phải spam, điểm màu đỏ đại diện cho email spam.
- **Mức độ phân tách:** Mức độ phân tách giữa các điểm màu xanh và đỏ cho thấy khả năng phân biệt giữa email spam và không spam của hai thành phần chính đầu tiên. Nếu các điểm phân tách rõ ràng, điều này chứng tỏ PCA đã giữ lại được thông tin phân biệt quan trọng.
- **Tâm của các nhóm (Centroids):** Các dấu sao đánh dấu tâm của nhóm spam và không spam. Khoảng cách giữa hai tâm này cho thấy mức độ khác biệt trung bình giữa hai loại email.

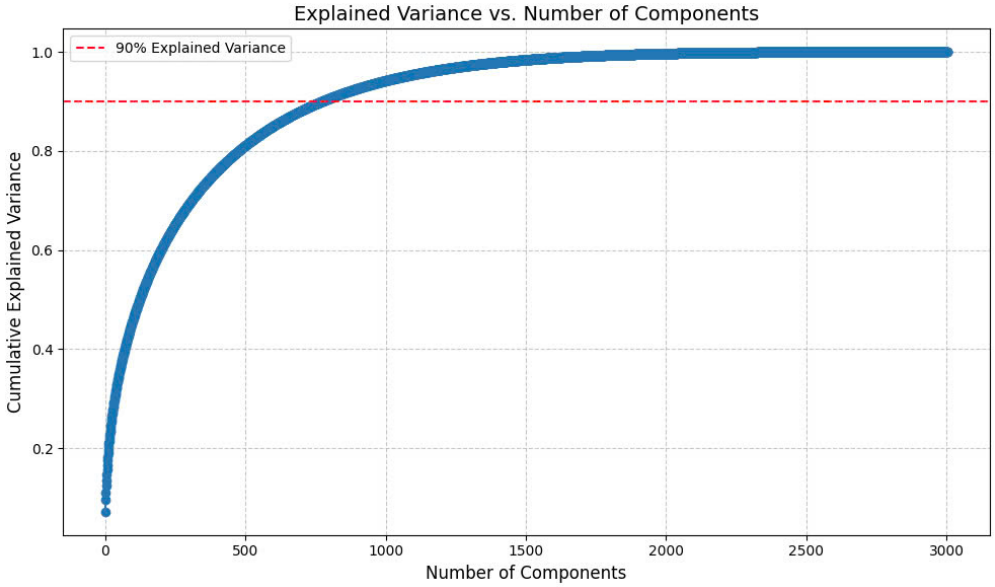


Figure 1: Biểu đồ phương sai tích lũy theo số lượng thành phần chính

- **Đường nối giữa các tâm:** Đường kẻ nối giữa hai tâm có thể được xem như một đường phân chia đơn giản giữa hai lớp. Nếu các điểm phân bố rõ ràng về hai phía của đường này, điều đó cho thấy khả năng phân loại tốt ngay cả với mô hình đơn giản.

Hình 2 trình bày phân bố của dữ liệu email trong không gian hai thành phần chính đầu tiên. Các điểm màu xanh đại diện cho email không phải spam, các điểm màu đỏ đại diện cho email spam. Hai điểm đánh dấu tâm của nhóm spam và không spam, với đường thẳng nối giữa hai tâm này.

- **Tốc độ tăng của đường cong:** Đường cong càng dốc ở đầu, càng chứng tỏ một số ít thành phần chính đầu tiên đã nắm giữ phần lớn thông tin trong dữ liệu..
- **Điểm cắt với ngưỡng 90% :** Điểm mà đường cong cắt đường ngang màu đỏ cho biết cần bao nhiêu thành phần chính để giữ lại 90% thông tin trong dữ liệu gốc.
- **Hiệu quả của việc giảm chiều :** Nếu chỉ cần một số nhỏ thành phần chính để đạt được 90% phương sai, điều này chứng tỏ việc giảm chiều dữ liệu rất hiệu quả và không làm mất nhiều thông tin quan trọng.

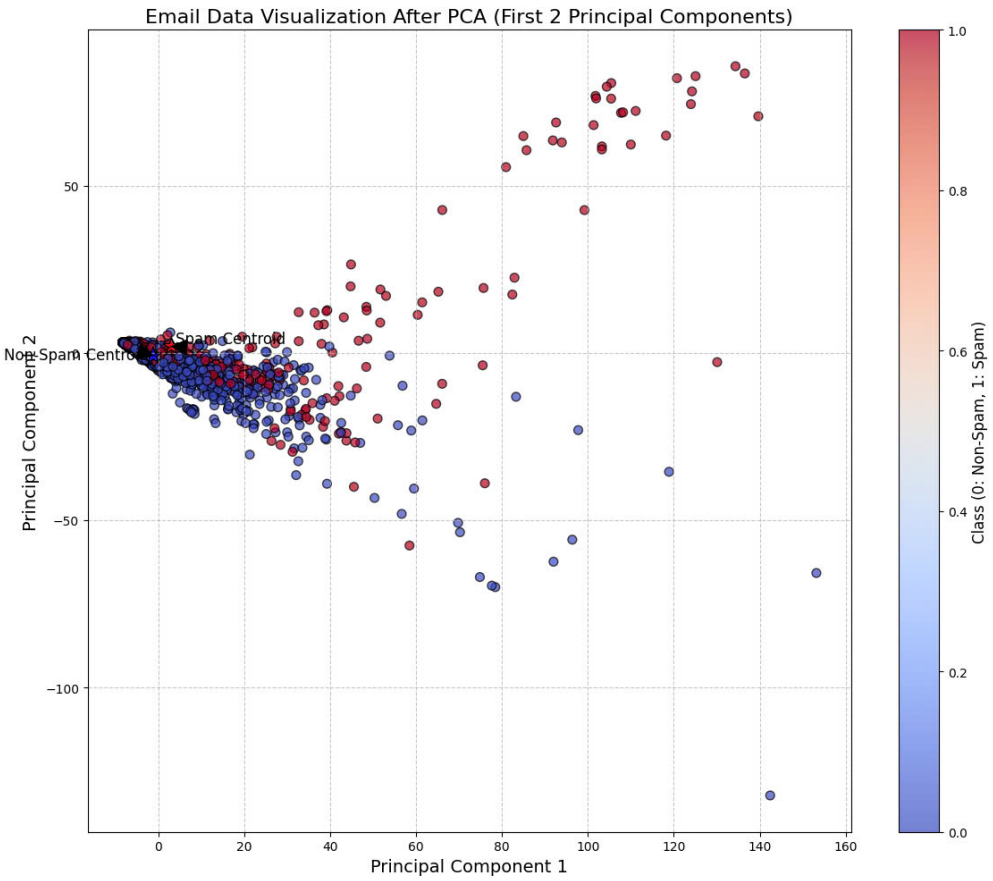


Figure 2: Biểu đồ phân tán của dữ liệu email trên hai thành phần chính đầu tiên

4.4 Thực nghiệm

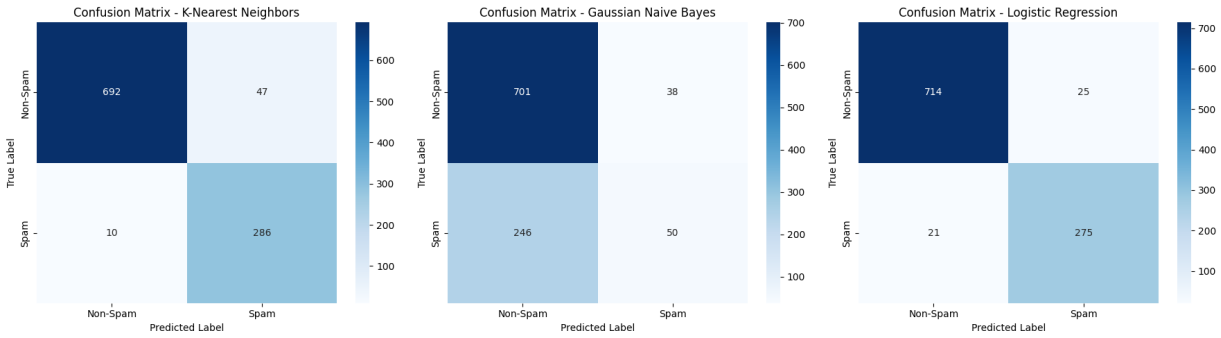
4.4.1 Dữ liệu gốc:

Bảng kết quả trên dữ liệu gốc

Kết quả các mô hình trên dữ liệu gốc

Mô hình	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9556	0.9167	0.9291	0.9228
K-Nearest Neighbors	0.9449	0.8589	0.9662	0.9094
Gaussian Naive Bayes	0.7256	0.5682	0.1689	0.2604

Ma trận nhầm lẫn của từng mô hình:



Trên dữ liệu gốc, ba mô hình được so sánh gồm Logistic Regression, K-Nearest Neighbors (KNN) và Gaussian Naive Bayes. Kết quả như sau:

- **Logistic Regression** đạt độ chính xác (accuracy) 0.9556, F1 Score 0.9228, cho thấy mô hình này hoạt động rất hiệu quả trên dữ liệu gốc. Ma trận nhầm lẫn cho thấy số lượng dự đoán sai (False Positives và False Negatives) thấp, đồng nghĩa với việc mô hình phân biệt tốt giữa email spam và không spam. Điều này phù hợp với đặc điểm của Logistic Regression, vốn xử lý tốt các đặc trưng thưa và nhiều chiều, cũng như tận dụng tốt mối quan hệ tuyến tính giữa các đặc trưng và nhãn phân loại.

- **K-Nearest Neighbors** có accuracy 0.9449 và F1 Score 0.9094, cũng thể hiện hiệu suất tốt. KNN tận dụng được cấu trúc dữ liệu gốc, đặc biệt khi các đặc trưng chưa bị biến đổi nhiều, giúp mô hình xác định được các điểm lân cận chính xác hơn. Tuy nhiên, KNN có thể bị ảnh hưởng bởi các đặc trưng không đồng nhất về thang đo hoặc nhiễu trong dữ liệu.

- **Gaussian Naive Bayes** cho kết quả thấp nhất (accuracy 0.7256, F1 Score 0.2604). Mô hình này giả định các đặc trưng độc lập có điều kiện, điều không đúng với dữ liệu văn bản thực tế, nơi các từ thường có mối liên hệ ngữ nghĩa. Do đó, số lượng dự đoán sai cao (False Negatives lớn), dẫn đến Recall và F1 Score thấp.

Nhìn chung, trên dữ liệu gốc, Logistic Regression và KNN đều cho hiệu quả tốt, trong khi Gaussian Naive Bayes bị hạn chế do giả định đơn giản hóa về mối quan hệ giữa các đặc trưng.

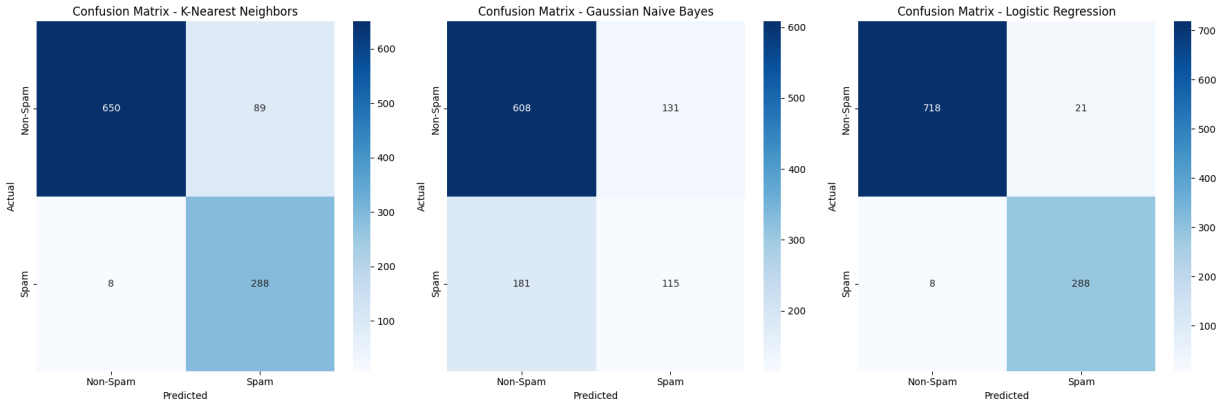
4.4.2 Dữ liệu đã chuẩn hóa và giảm chiều:

Bảng kết quả trên dữ liệu đã chuẩn hóa và giảm chiều

Kết quả các mô hình trên dữ liệu đã chuẩn hóa và giảm chiều

Mô hình	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9720	0.9320	0.9730	0.9521
K-Nearest Neighbors	0.9063	0.7639	0.9730	0.8559
Gaussian Naive Bayes	0.6986	0.4675	0.3885	0.4244

Ma trận nhầm lẫn của từng mô hình:



Phân tích kết quả trên dữ liệu đã chuẩn hóa và giảm chiều

Sau khi áp dụng các bước chuẩn hóa (đưa dữ liệu về cùng thang đo) và giảm chiều (PCA), kết quả các mô hình thay đổi rõ rệt:

- **Logistic Regression** tiếp tục vượt trội với accuracy tăng lên 0.9720 và F1 Score đạt 0.9521. Việc chuẩn hóa giúp các đặc trưng có cùng thang đo, còn PCA loại bỏ nhiễu và giữ lại thông tin quan trọng nhất, giúp Logistic Regression học hiệu quả hơn và tổng quát hóa tốt hơn trên dữ liệu kiểm thử.
- **K-Nearest Neighbors** lại giảm hiệu suất (accuracy còn 0.9063, F1 Score 0.8559). Nguyên nhân là do KNN phụ thuộc mạnh vào khoảng cách giữa các điểm dữ liệu trong không gian đặc trưng. Khi giảm chiều bằng PCA, cấu trúc khoảng cách có thể bị biến đổi, làm giảm khả năng phân biệt của KNN, đặc biệt nếu các thành phần chính không giữ được toàn bộ thông tin cần thiết cho phân loại.
- **Gaussian Naive Bayes** có cải thiện về F1 Score (lên 0.4244) nhưng accuracy giảm nhẹ (còn 0.6986). Chuẩn hóa và giảm chiều giúp giảm bớt sự phụ thuộc giữa các đặc trưng, phần nào phù hợp hơn với giả định của Naive Bayes, nhưng hiệu quả tổng thể vẫn thấp hơn hai mô hình còn lại do bản chất đơn giản hóa của thuật toán này.

Kết luận

- Logistic Regression là mô hình hưởng lợi rõ rệt nhất từ việc chuẩn hóa và giảm chiều, cho kết quả tốt nhất ở cả hai giai đoạn.
- KNN phù hợp hơn với dữ liệu gốc, nhưng hiệu suất giảm khi đặc trưng bị biến đổi qua PCA.
- Gaussian Naive Bayes có cải thiện nhẹ sau chuẩn hóa và giảm chiều, nhưng vẫn kém hiệu quả so với hai mô hình còn lại.

Việc lựa chọn phương pháp tiền xử lý phù hợp với từng mô hình là rất quan trọng để đạt hiệu suất tối ưu trong bài toán phân loại email.

5 Kết luận và Phương hướng phát triển

5.1 Kết luận

Trong báo cáo này, chúng tôi đã trình bày quy trình xây dựng hệ thống phân loại email tự động dựa trên ba mô hình học máy: **K-Nearest Neighbors (KNN)**, **Gaussian Naive Bayes (GNB)** và **Logistic Regression (LR)**. Qua quá trình phân tích, tiền xử lý, huấn luyện và đánh giá mô hình, chúng tôi nhận thấy rằng:

- **Logistic Regression** mang lại hiệu quả cao nhất với độ chính xác 97.2% và F1 Score vượt trội, phù hợp với dữ liệu có đặc trưng thưa.
- **KNN** hoạt động ổn định với khả năng nhận diện (Recall) tốt nhưng hiệu quả tổng thể thấp hơn LR do nhạy cảm với lựa chọn tham số k và chi phí tính toán cao.
- **Gaussian Naive Bayes** là mô hình đơn giản, tốc độ nhanh nhưng giả định độc lập giữa các đặc trưng không phù hợp với dữ liệu văn bản, dẫn đến độ chính xác thấp hơn.

Ngoài ra, bộ dữ liệu sử dụng trong bài toán có cấu trúc rõ ràng, tỷ lệ các lớp gần cân bằng, và không có dữ liệu thiếu. Điều này hỗ trợ tốt cho quá trình huấn luyện và đánh giá mô hình một cách khách quan.

5.2 Phương hướng phát triển

Trong tương lai, có thể mở rộng và cải tiến nghiên cứu theo các hướng sau:

- **Sử dụng đặc trưng nâng cao:** Áp dụng các kỹ thuật biểu diễn văn bản như TF-IDF, Word2Vec, FastText hoặc BERT thay vì chỉ dùng tần suất từ để mô hình hiểu rõ hơn ngữ cảnh.
- **Thử nghiệm mô hình khác:** Triển khai thêm các mô hình hiện đại như SVM, Random Forest, hoặc các mạng nơ-ron như LSTM, BERT, để so sánh hiệu suất và khả năng tổng quát hóa.

- **Tối ưu siêu tham số:** Sử dụng các kỹ thuật như Grid Search, Random Search hoặc Bayesian Optimization để tìm tập siêu tham số tốt nhất cho từng mô hình.
- **Mở rộng tập dữ liệu:** Thu thập thêm email thuộc nhiều chủ đề và phong cách khác nhau để tăng tính đa dạng, từ đó giúp mô hình học được nhiều đặc trưng hơn.
- **Triển khai ứng dụng thực tế:** Tích hợp mô hình vào hệ thống hỗ trợ quản lý email, thư mục thông minh hoặc trợ lý ảo để tự động lọc, sắp xếp và gợi ý phản hồi cho người dùng.

Những hướng phát triển trên không chỉ nâng cao hiệu suất mô hình mà còn giúp ứng dụng kết quả nghiên cứu vào thực tiễn một cách hiệu quả và bền vững.

6 Tài liệu tham khảo

1. Machine Learning cơ bản - Vũ Hữu Tiệp.
2. Mining of Massive Datasets - Jure Leskovec, Anand Rajaraman, Jeff.
3. Bài giảng Hệ gợi ý - Socit - Đại học Bách khoa Hà Nội.
4. Machine Learning TextBook - Andreas Lindholm, Niklas Wahlstrom, Fredrik Lindsten, Thomas B.Schon.