Bài toán phân loại email spam

Trình bày: Đỗ Thu Liễu.

Giảng viên hướng dẫn: Nguyễn Hoàng Điệp.

Ngày thuyết trình: 26/12/2025

Nội dung

1. **Tổng quan về chủ đề.**
2. **Phân tích dữ liệu.**
3. **Tiền xử lý dữ liệu.**
4. **Mô hình đào tạo Machine Learning.**
5. **Đánh giá mô hình.**
6. **Tổng kết.**

Tìm hiểu bài toán

- Phân loại email (email classification) là một lĩnh vực quan trọng trong xử lý ngôn ngữ tự nhiên (NLP).
- Mục tiêu: Xây dựng mô hình phân biệt chính xác đâu là thư rác (spam) và đâu là thư thường (ham).

1. Tổng quan về chủ đề

- **Ví dụ 1 (Spam – quảng cáo):** “Congratulations! You have won a FREE gift voucher worth $500.Click the link below to claim now:http://xxxxxxx.win-now.com(Do not miss this limited-time offer!)”

- **Ví dụ 2 (Spam – thông báo giả mạo):** “Your bank account has been temporarily suspended.Please verify your identity immediately at:http://secure-banking-xxxx.comFailure to do so may result in permanent lock. 500.Click the link below to claim now:http://xxxxxxx.win-now.com(Do not miss this limited-time offer!)””

1.Tổng quan về chủ đề

Dataset

- Nguồn dataset: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
- Dataset: SMS Spam Collection (5572 mẫu).
- Gồm 2 cột chính: v1, v2.
- Hai nhãn: “ham” (bình thường) và “spam”.

v1 class		v2 sms			
ham spam	87% 13%	5169 unique values		[null] bt not his girlfrnd..... Other (47)	99% 0% 1%	[null] MK17 92H. 450Pp... Other (10)	100% 0% 0%
ham		Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got a...			
ham		Ok lar... Joking wif u oni...			
spam		Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entr...			

Hình 1: Dataset.

2. Phân tích dữ liệu

Ý tưởng để giải quyết vấn đề

- Tập dữ liệu:
  - **EDA** - Phân tích và khai phá dữ liệu
- Xử lý:
  - Làm sạch dữ liệu.
  - Biểu diễn văn bản:
    - **BoW**
    - **TF-IDF**
  - Kỹ thuật chuẩn hóa dữ liệu:
    - **MaxAbsScaler.**
  - Xử lý bài toán mất cân bằng: **SMOTE.**

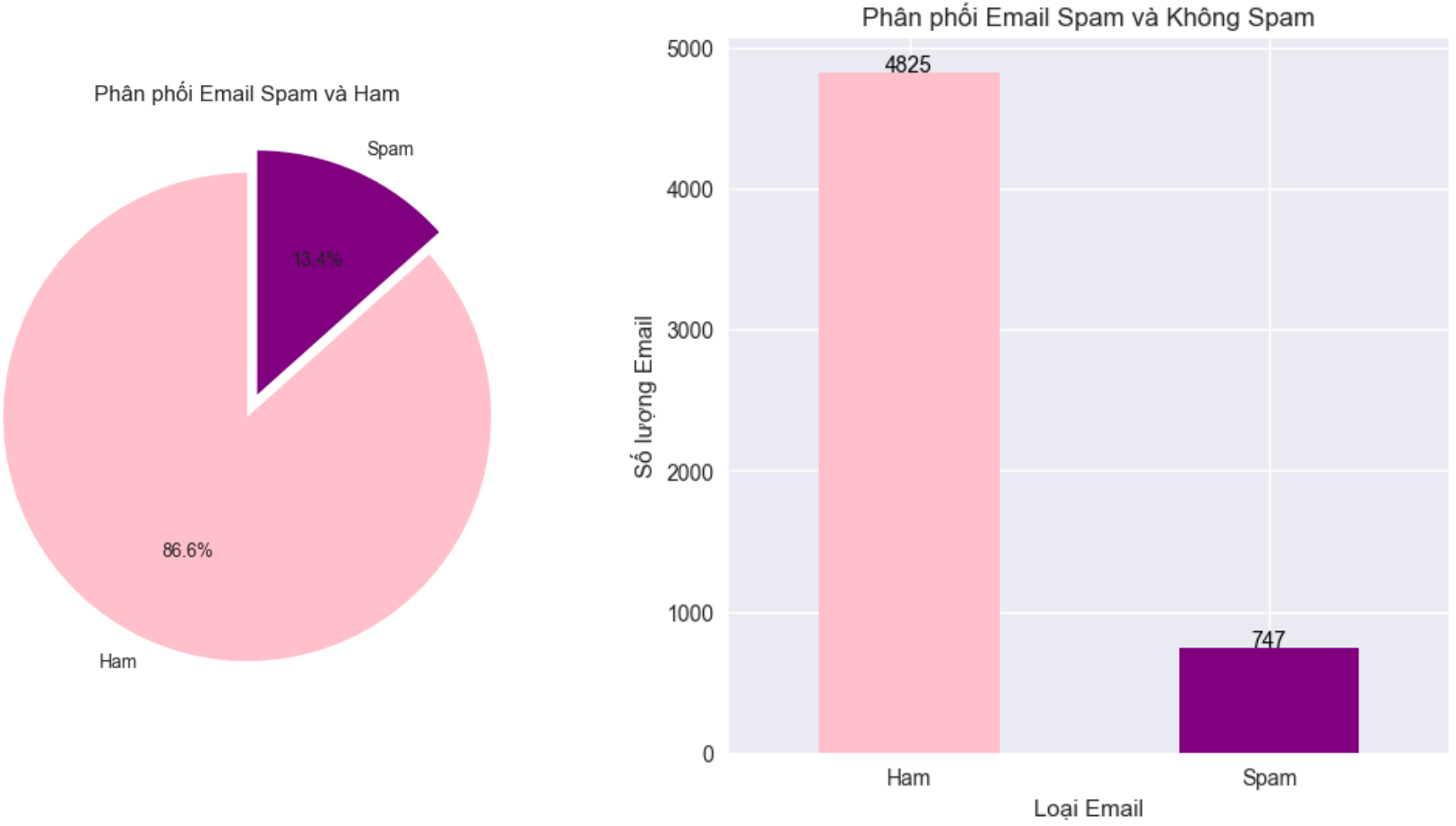
2. Phân tích dữ liệu

Ý tưởng để giải quyết vấn đề

- Mô hình học máy:
  - Naïve bayes.
  - Logistic regression.
- Đánh giá:
  - Ma trận nhầm lẫn.
  - Chỉ số đánh giá.

2. Phân tích dữ liệu

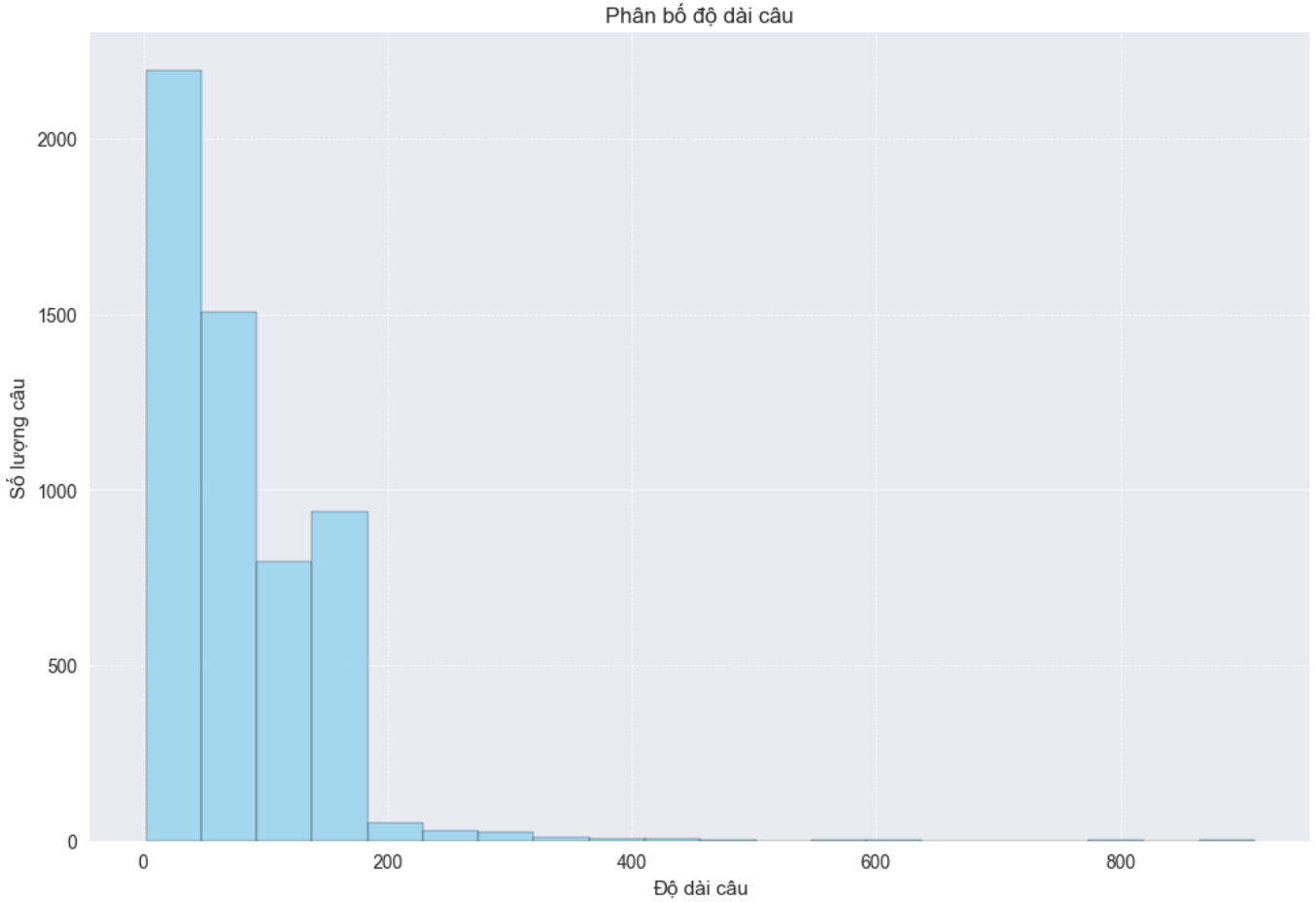
Tập dữ liệu mất cân bằng (spam chiếm ~13,4%).



Hình 2: Biểu đồ thống kê số lượng từng nhãn.

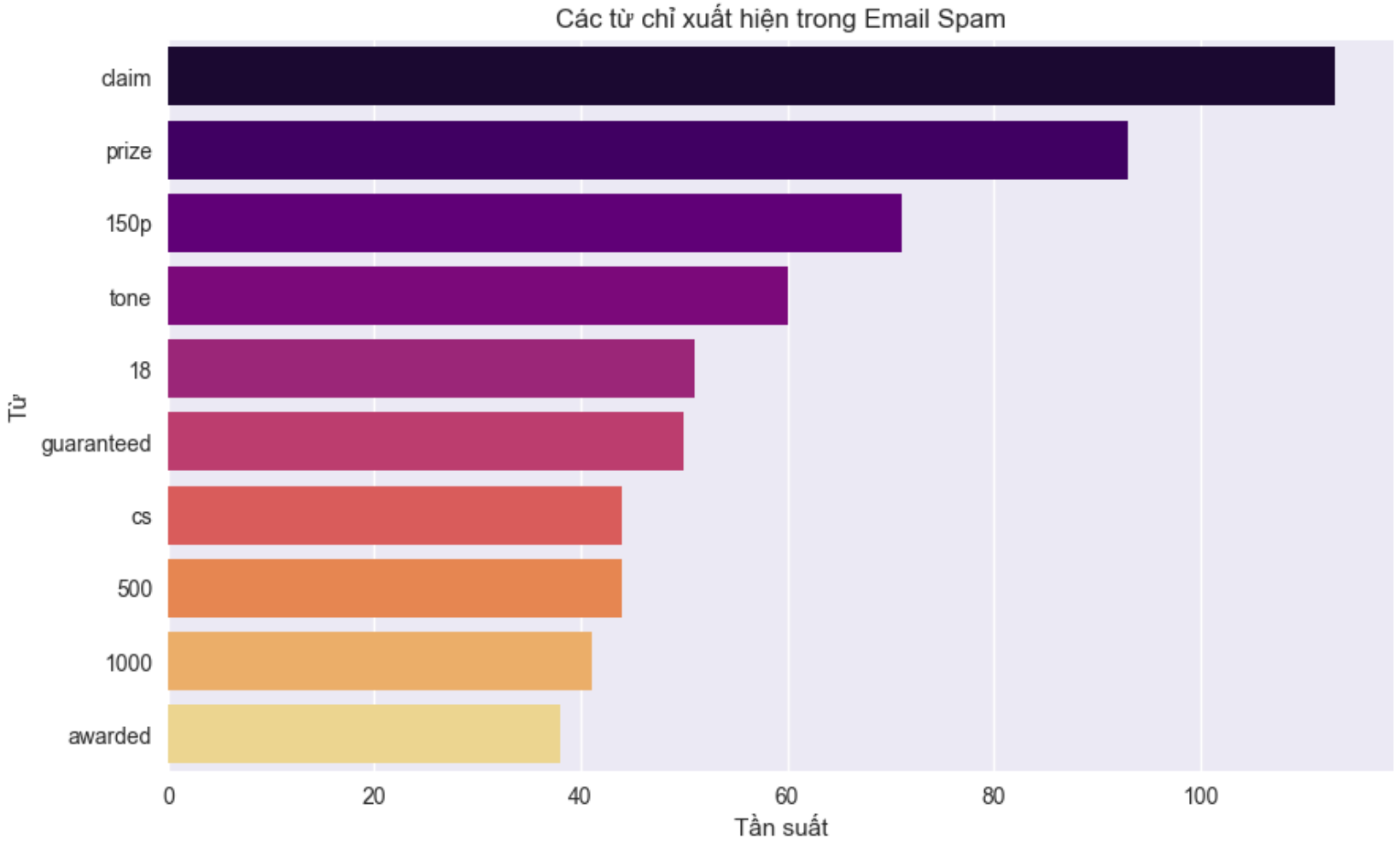
2. Phân tích dữ liệu

- Email spam thường ngắn hơn, chứa nhiều từ nhấn mạnh.
- Email ham có xu hướng dài hơn, mang ngữ cảnh đầy đủ và ít ký tự đặc biệt.



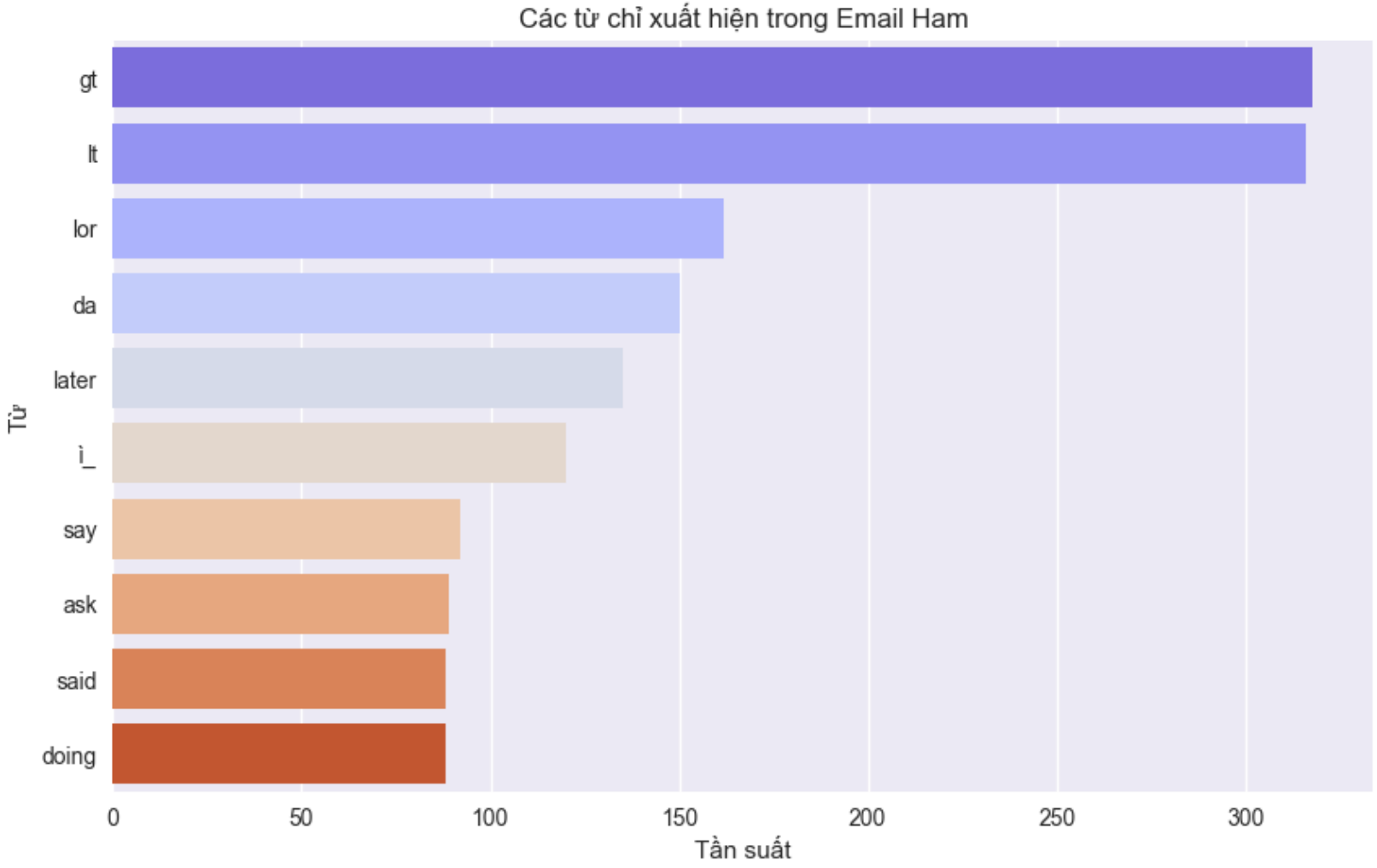
Hình 3: Biểu đồ thể hiện phân phối độ dài tin nhắn.

2. Phân tích dữ liệu



Hình 4: Biểu đồ thể hiện tần suất các từ xuất hiện nhiều nhất trong tin nhắn email spam.

2. Phân tích dữ liệu



Hình 5: Biểu đồ thể hiện tần suất các từ xuất hiện nhiều nhất trong tin nhắn email ham.

Số lượng từ trong từ điển là : 8713.

Hình 6: Biểu đồ thể hiện các từ phổ biến nhất trong nhóm email "ham" và "spam".

3. Tiền xử lý dữ liệu

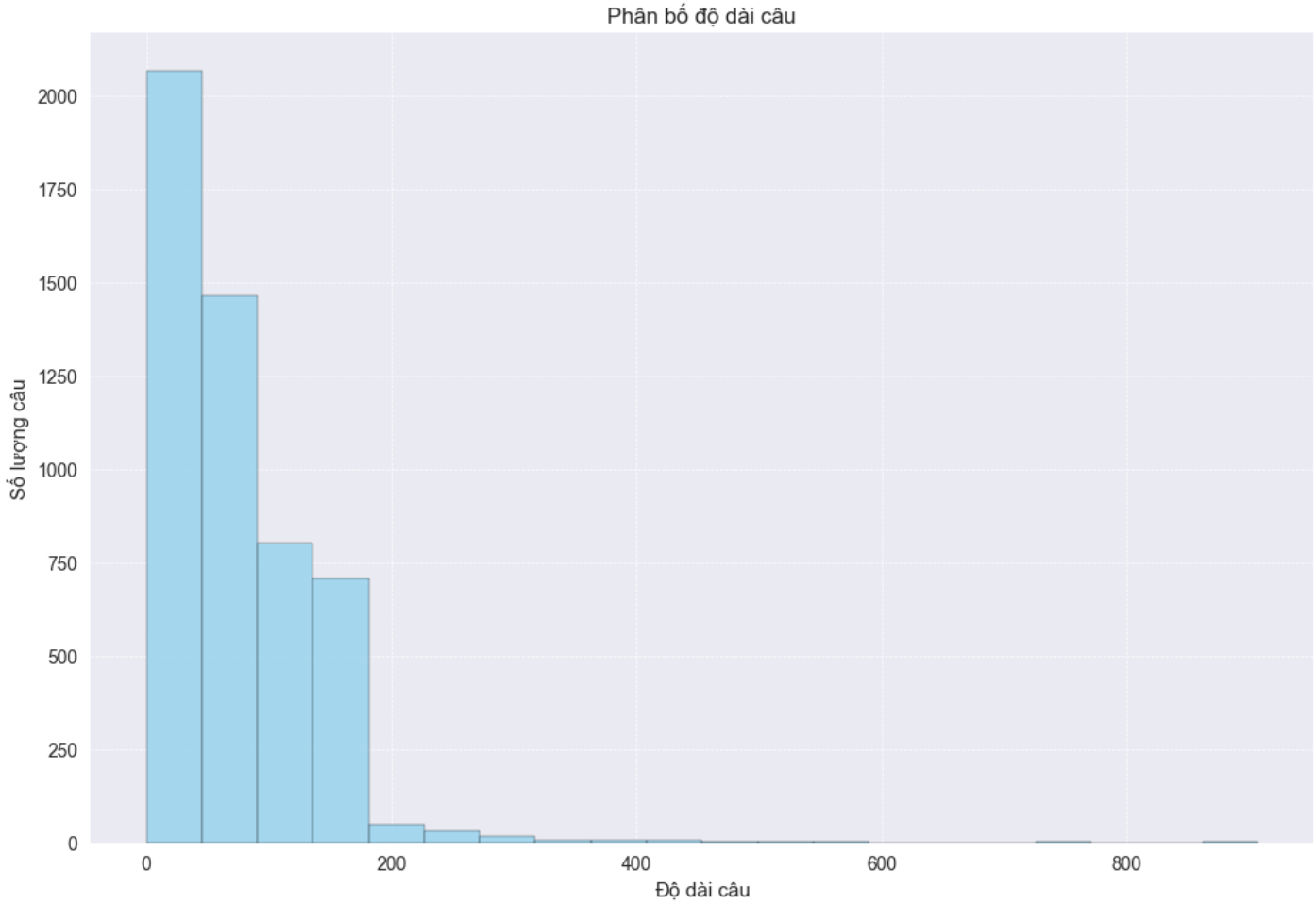
❑ Làm sạch văn bản:

- Kiểm tra dữ liệu đầu vào
- Xử lý dữ liệu trùng lặp.
- 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 .

Trước làm sạch	Sau làm sạch
"OMG 😭😭!!! Call me at 123-456-7899 pleaaaaaseeee!!!!!»	"omg call me at 123 456 7899 please"

3. Tiền xử lý dữ liệu

Sau khi làm sạch, độ dài tin nhắn giảm trung bình 12–20%.



Hình 7: Biểu đồ thể hiện phân phối độ dài tin nhắn sau khi làm sạch dữ liệu.

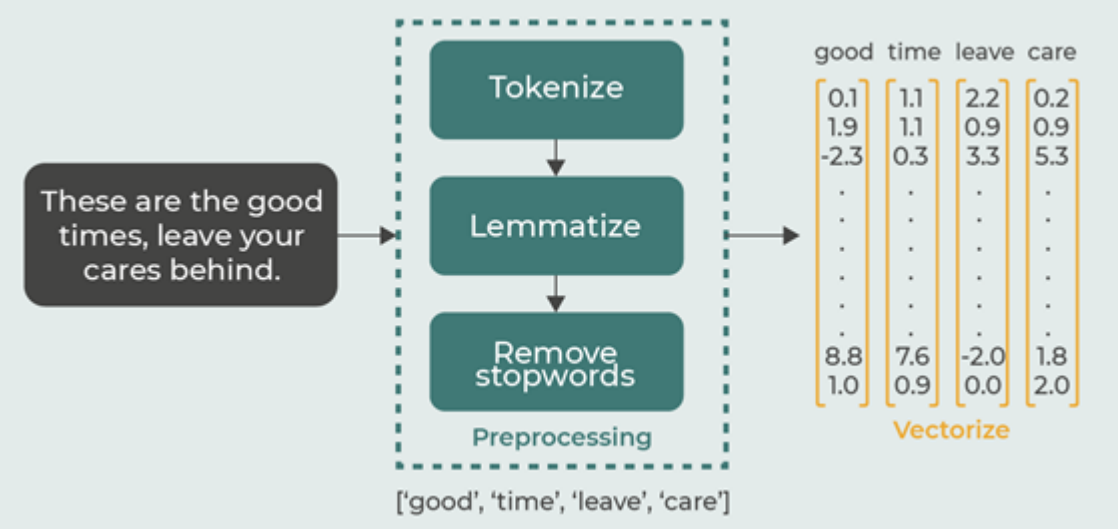
3. Tiền xử lý dữ liệu

Số lượng từ trong từ điển là : 8591.

Hình 8: Biểu đồ thể hiện các từ phổ biến nhất trong nhóm email "ham" và "spam" sau khi làm sạch dữ liệu.

Biểu diễn văn bản

- Dữ liệu văn bản cần được biểu diễn dưới dạng số để các mô hình xử lý.
- Các phương pháp phổ biến:
  - Túi từ (BoW).
  - Tần suất thuật ngữ - Tần suất tài liệu nghịch đảo (TF-IDF).



Hình 9: Hình ảnh quy trình xử lý văn bản.

Bag-of-Words (BoW)

- **Bag-of-Words (BoW):** Biểu diễn mỗi tài liệu dưới dạng một vectơ có độ dài bằng số từ trong từ vựng. Giá trị của mỗi phần tử trong vectơ là tần suất xuất hiện của từ tương ứng trong tài liệu.
- **Ví dụ:**
  - “I love machine learning”
  - “I love learning ML”
  - Bộ từ vựng sẽ là : ["I", "love", "machine", "learning“, ”ML”]
  - Các vectơ BoW cho các câu này sẽ là:
    - "I love machine learning": [1, 1, 1, 1, 0]
    - "I love learning ML": [1, 1, 0, 1, 1]

3. Tiền xử lý dữ liệu

❖ **TF-IDF** (Term Frequency – Inverse Document Frequency) : Đánh giá **tầm quan trọng của một từ** trong một tài liệu hoặc toàn bộ tập tài liệu.

➢ **Tần suất Thuật ngữ (TF):**

- TF đo lường tần suất xuất hiện của một từ trong một tài liệu cụ thể.

➢ **Tần suất Tài liệu Nghịch đảo (IDF):**

- IDF đo lường tầm quan trọng của một từ trong toàn bộ tập hợp tài liệu. Nếu một từ xuất hiện trong nhiều tài liệu, giá trị IDF của nó sẽ thấp hơn.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

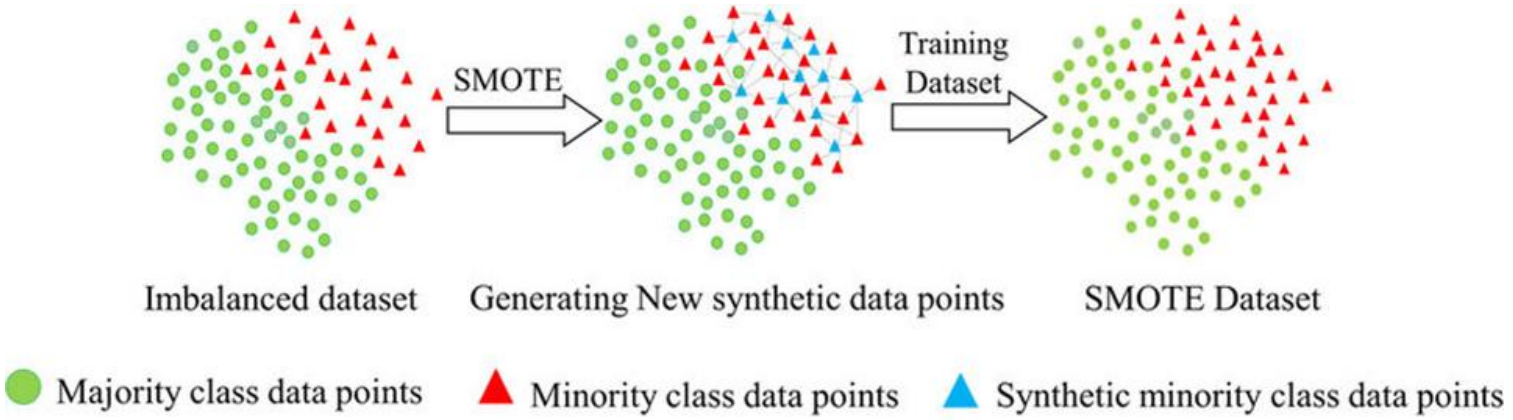
$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

3. Tiền xử lý dữ liệu

❖ **Smote:** phương pháp xử lý mất cân bằng dữ liệu. Nó hoạt động bằng cách tạo ra các điểm dữ liệu nhân tạo mới cho lớp Spam dựa trên những điểm lân cận, giúp cân bằng lại tập dữ liệu

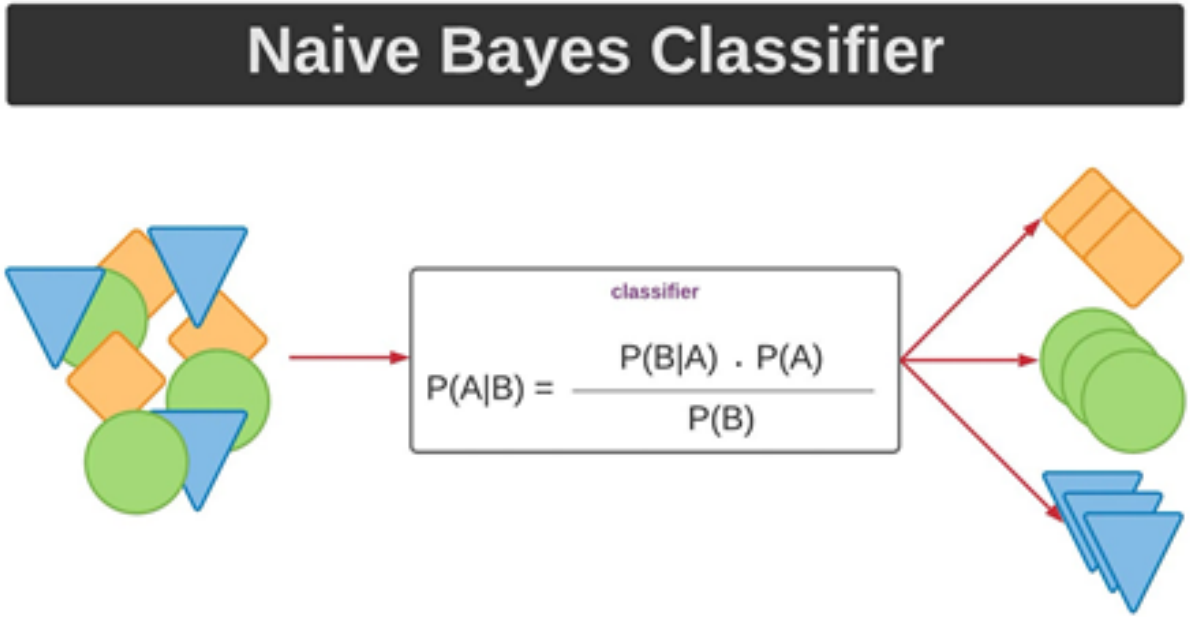


MaxAbsScaler

- **Mục đích:** Chuẩn hóa dữ liệu sao cho giá trị nằm trong [-1, 1] hoặc [0, 1] (với dữ liệu dương).
- **Cách hoạt động:** Chia mỗi giá trị của feature cho giá trị tuyệt đối lớn nhất của feature đó.

$$X_{\text{scaled}} = \frac{X}{\max(|X|)}$$

Naïve Bayes: là một thuật toán phân loại dựa trên Định lý Bayes, giả định rằng các đặc trưng (features) là độc lập với nhau.



Naïve Bayes

❑ **Ưu điểm:**

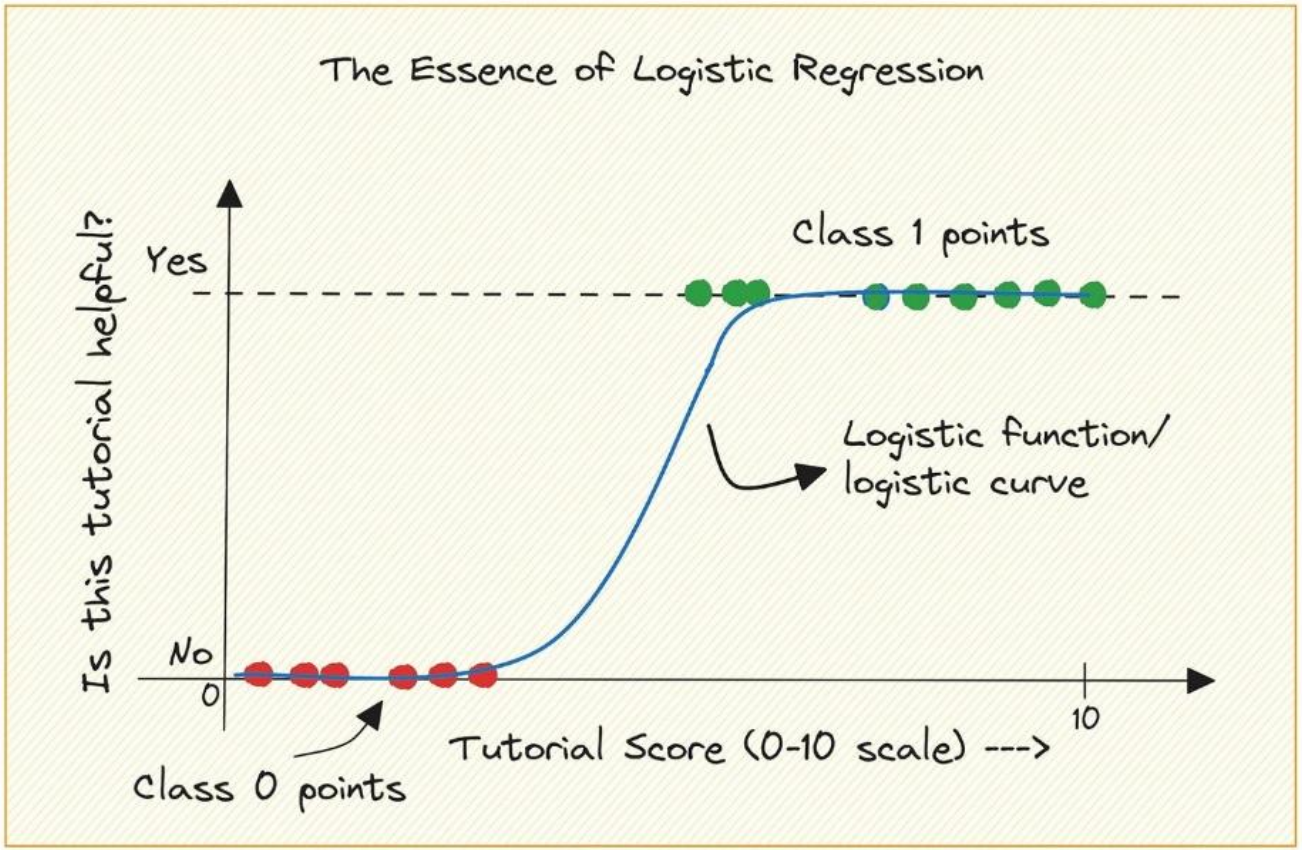
- Nhanh, nhẹ, dễ huấn luyện
- Phù hợp với dữ liệu văn bản nhiều chiều
- Hoạt động tốt với ít dữ liệu.
- Ít overfitting.

❑ **Nhược điểm:**

- Giả định độc lập giữa các từ (không thực tế).
- Dễ bị ảnh hưởng bởi dữ liệu mất cân bằng.
- Không nắm bắt được ngữ nghĩa của từ.

4. Mô hình đào tạo Machine Learning

Logistic regression: là một mô hình hồi quy tuyến tính dùng để phân loại nhãn nhị phân, dự đoán xác suất của một sự kiện xảy ra.



Logistic regression

- **Ưu điểm:**
  - Đơn giản, dễ triển khai
  - Tốt với dữ liệu phân tách tuyến tính.
  - Có xác suất dự báo, giúp đánh giá mức độ chắc chắn.
- **Nhược điểm:**
  - Không phù hợp dữ liệu quá phức tạp.
  - Rất nhạy cảm với dữ liệu mất cân bằng.

5. Đánh giá mô hình

Text Representation	Model	Accuracy	Precision	Recall	F1
BOW	**Naïve Bayes**	**97.54**	**97.50**	**97.54**	**97.50**
	Logistic Regression	94.57	94.79	94.57	94.66
TF-IDF	Naïve Bayes	97.15	97.09	97.15	97.09
	Logistic Regression	90.56	91.48	90.56	88.01

- Kết luận: Sự kết hợp giữa biểu diễn BOW và mô hình Naive Bayes là kết quả tối ưu nhất cho tập dữ liệu này.

6. Tổng kết

1. Hiệu quả của quy trình xử lý dữ liệu

- Làm sạch văn bản giúp giảm nhiễu, loại bỏ ký tự thừa và tăng chất lượng đặc trưng.
- TF-IDF biểu diễn văn bản tốt hơn BoW nhờ làm nổi bật các từ khóa quan trọng.
- SMOTE giải quyết mất cân bằng nhãn, cải thiện khả năng phát hiện spam.
- Chuẩn hóa bằng MaxAbsScaler giúp mô hình ổn định hơn.

6. Tổng kết

2. Đánh giá mô hình Machine Learning

- Naive Bayes: nhanh, hiệu quả với dữ liệu lớn, nhưng phụ thuộc giả định độc lập giữa các từ.
- Logistic Regressaion: ổn định, dự báo xác suất tốt, phù hợp với TF-IDF và dữ liệu dạng văn bản.

3. Mô hình tối ưu cho bài toán

- Sự kết hợp BOW + Naive Bayes cho kết quả tốt nhất về độ chính xác, độ ổn định .

DEMO

THANKS FOR LISTENING!