Spam Email Classification

Giảng viên:

PGS.TS Nguyễn Đình Thúc

ThS. Vũ Quốc Hoàng

ThS. Nguyễn Ngọc Toàn

ThS. Nguyễn Thị Hường

Sinh viên:

21120107 - Nguyễn Minh Nhật

21120056 - Nguyễn Đặng Tường Duy

NỘI DUNG

01 Đánh giá kết quả

02 Phân tích dữ liệu

03 Phương pháp đề xuất

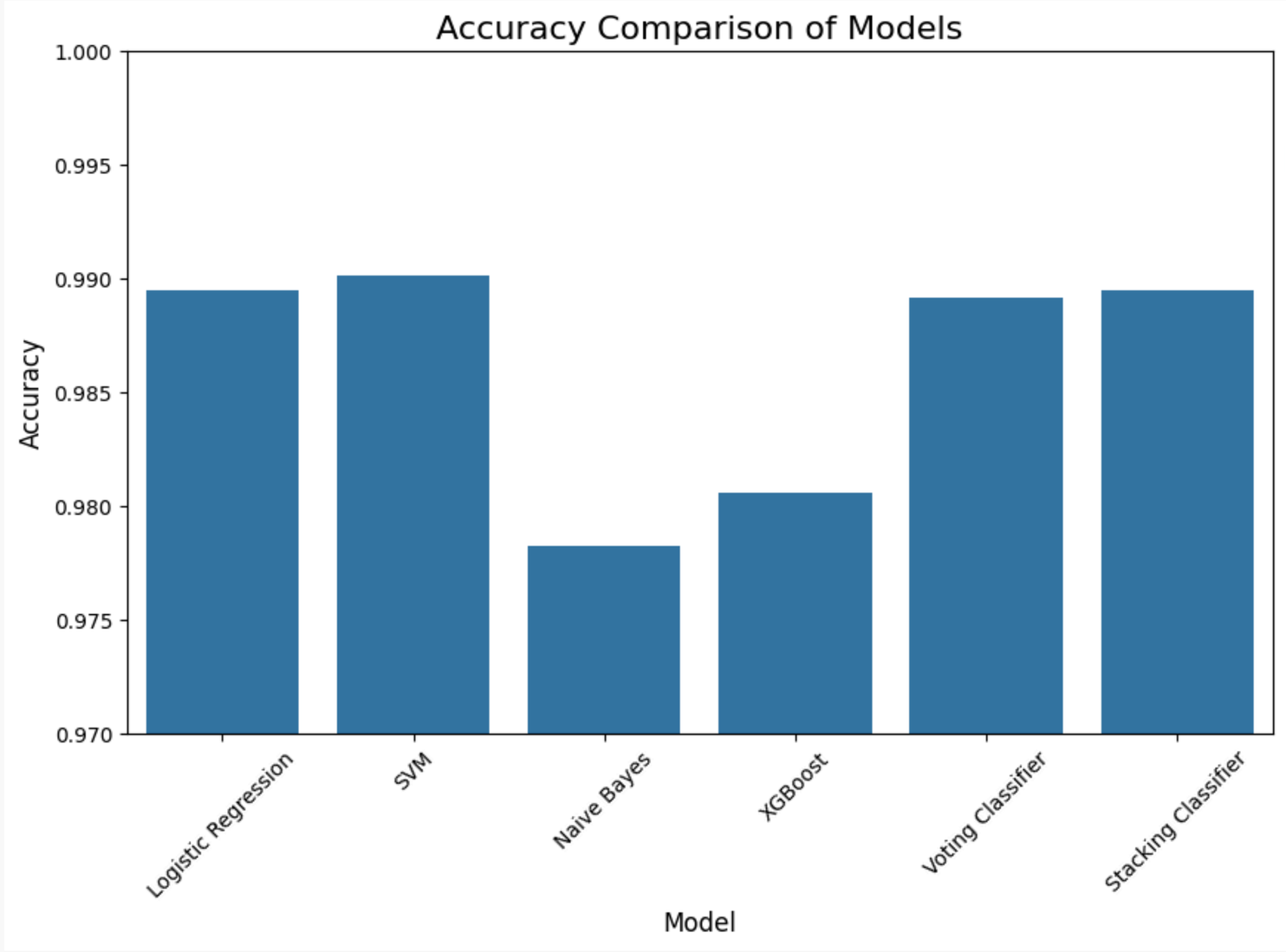
04 Demo

Đánh giá kết quả

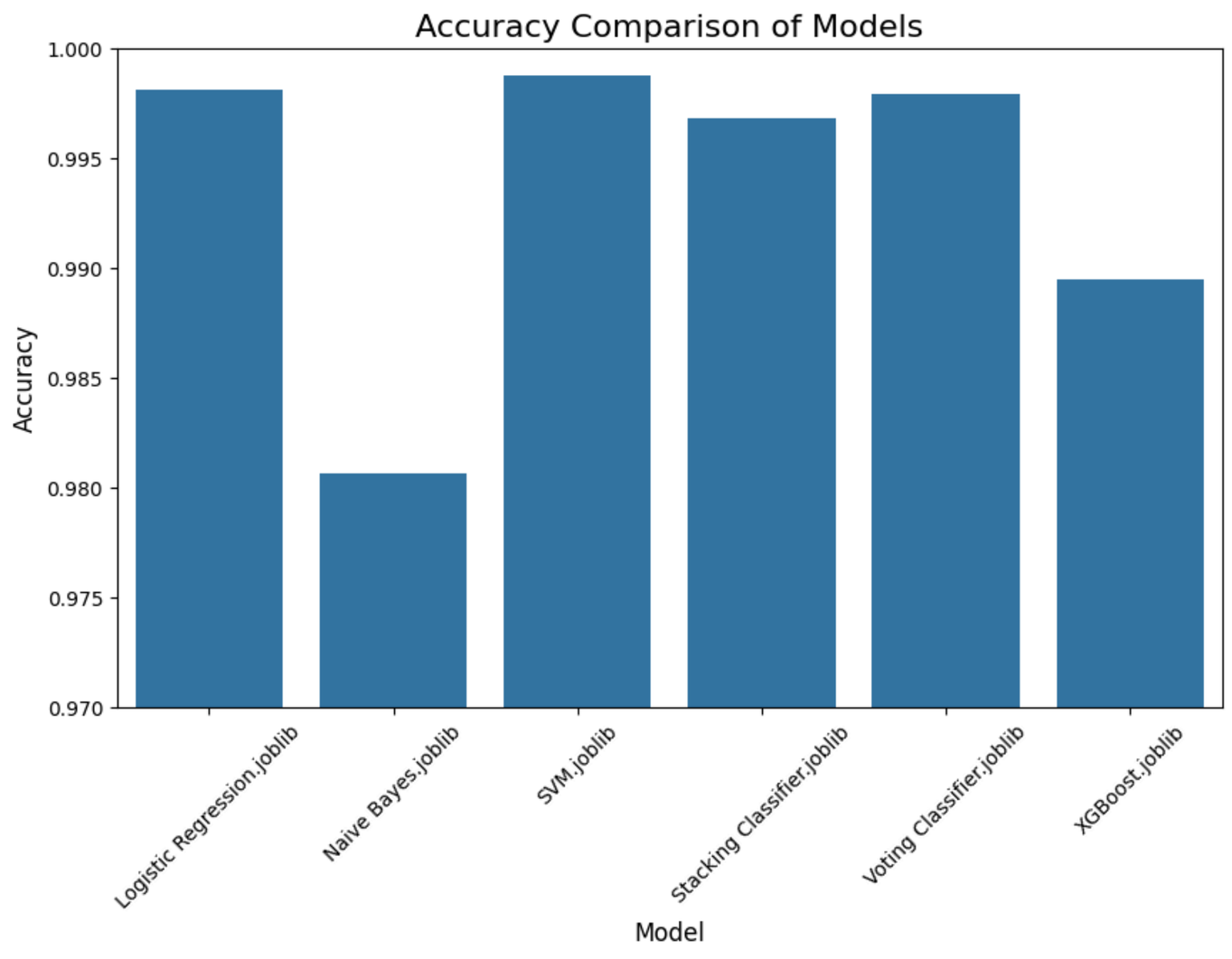
Đánh giá kết quả

STT	Đặc tả yêu cầu	Mức độ hoàn thành
1	Tải và xuất dữ liệu Enron-Spam	Hoàn Thành
2	Tiền xử lý dữ liệu	Hoàn Thành
3	Train mô hình và đánh giá kết quả	Hoàn Thành
4	Viết chương trình kiểm thử tính năng	Hoàn Thành

Kết quả tập Val

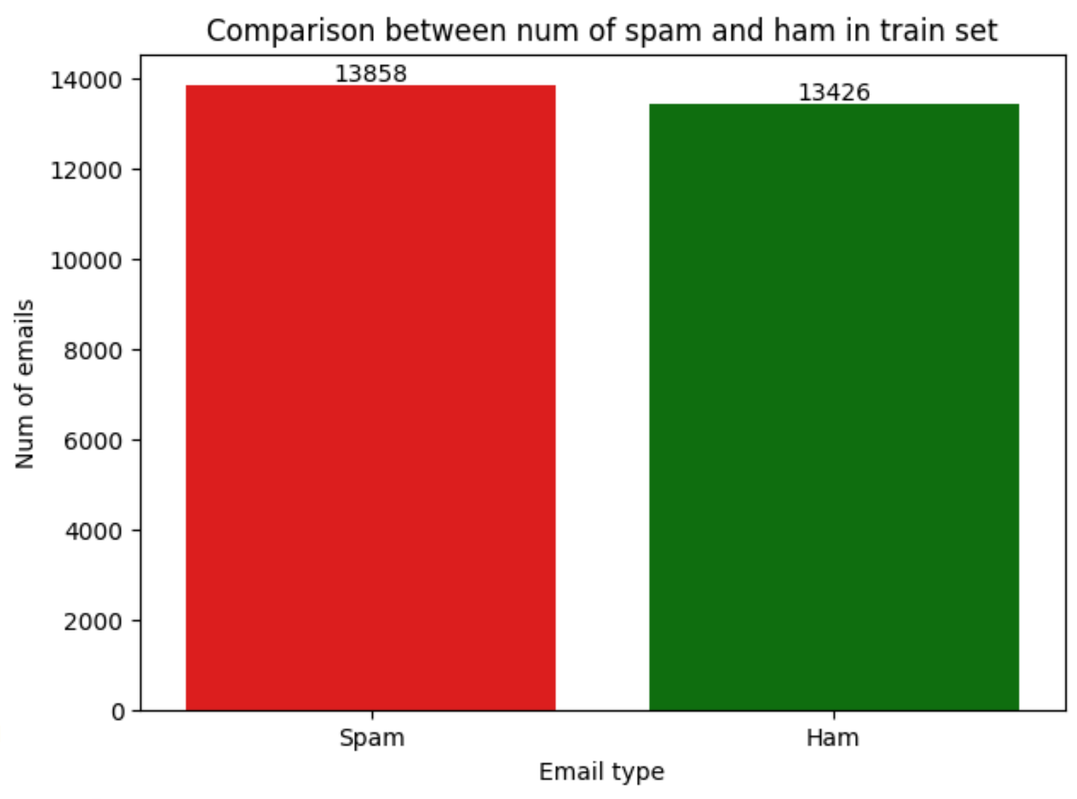


Kết quả tập Train

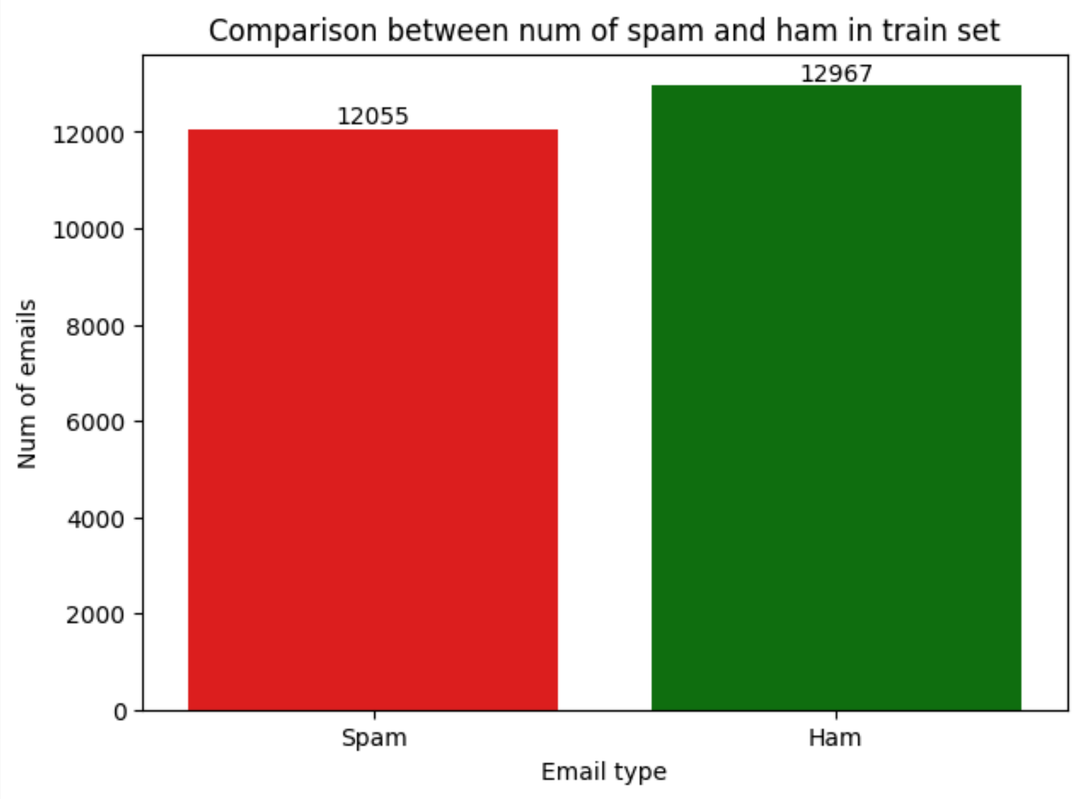


Phân tích dữ liệu

Bộ dữ liệu cân bằng ở cả tập train và val



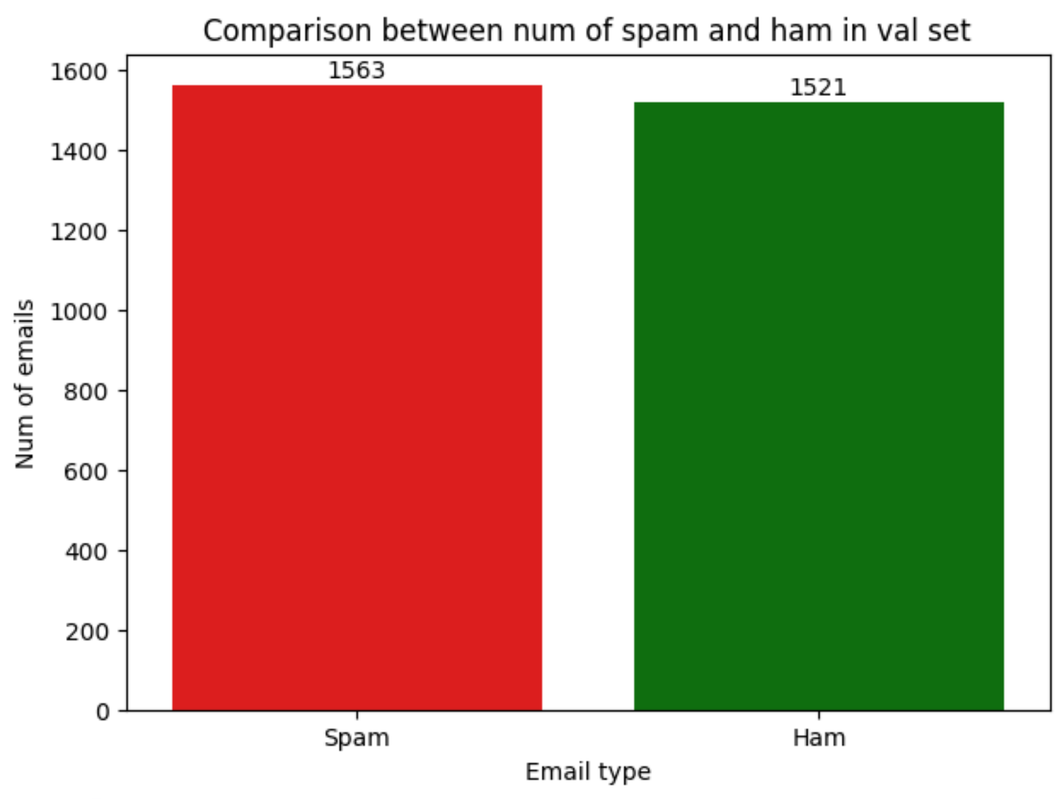
Trước khi làm sạch data



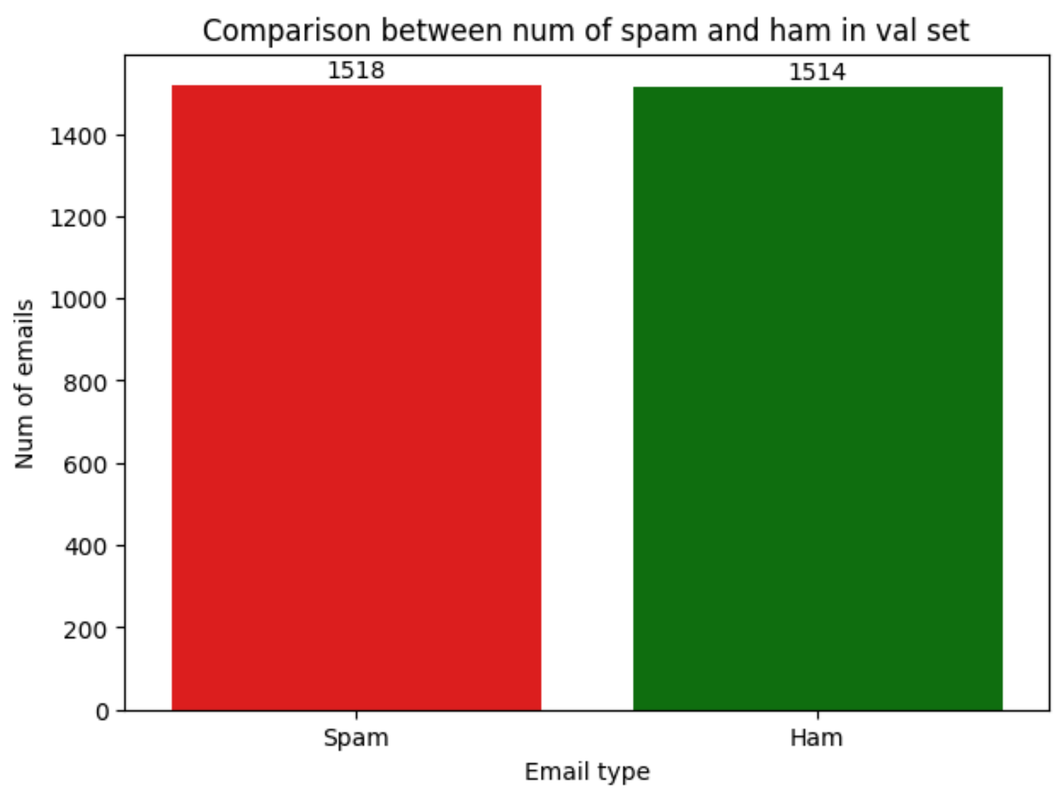
Sau khi làm sạch data

Phân bố nhãn trong tập huấn luyện

Bộ dữ liệu cân bằng ở cả tập train và val



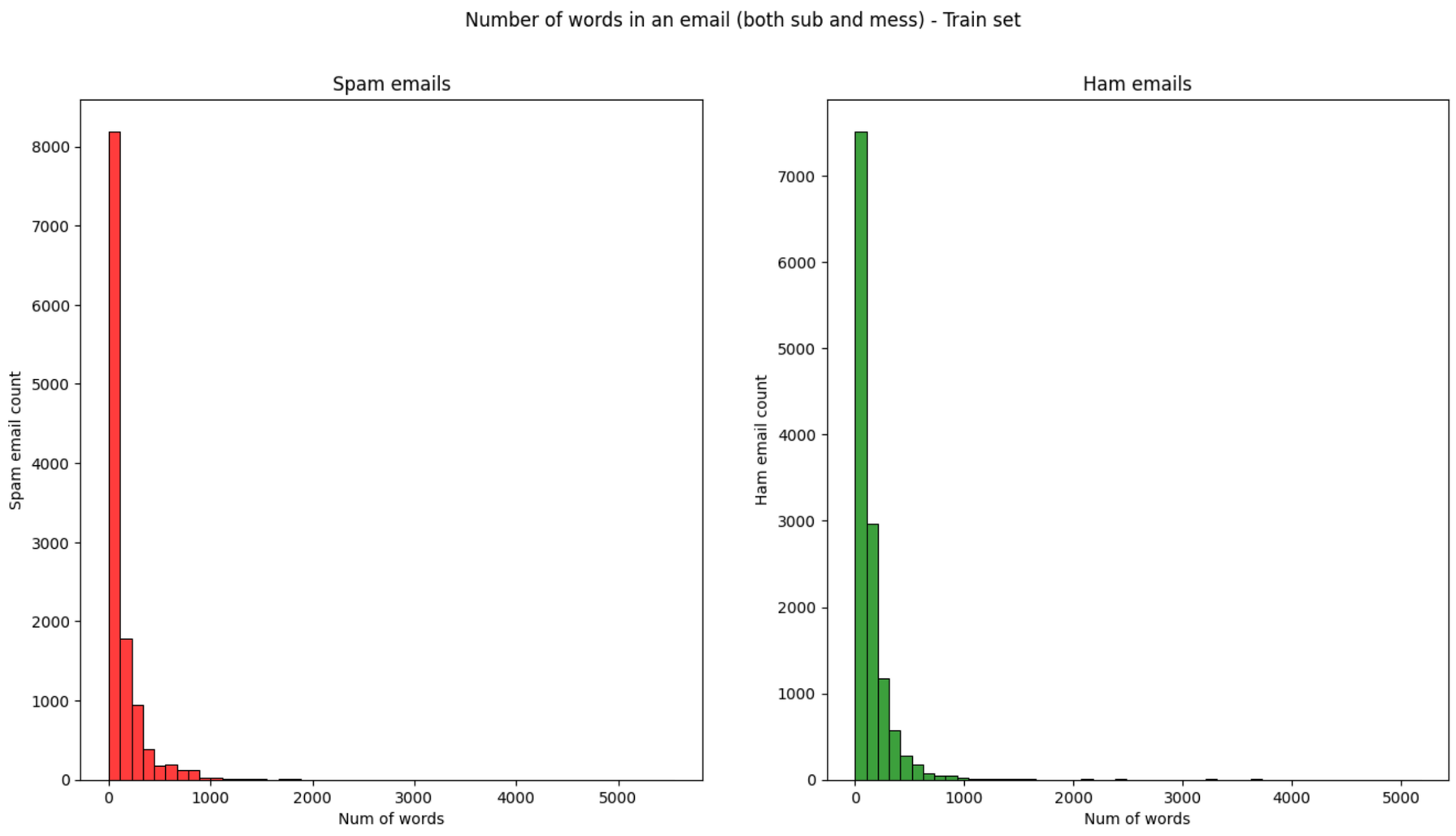
Trước khi làm sạch data



Sau khi làm sạch data

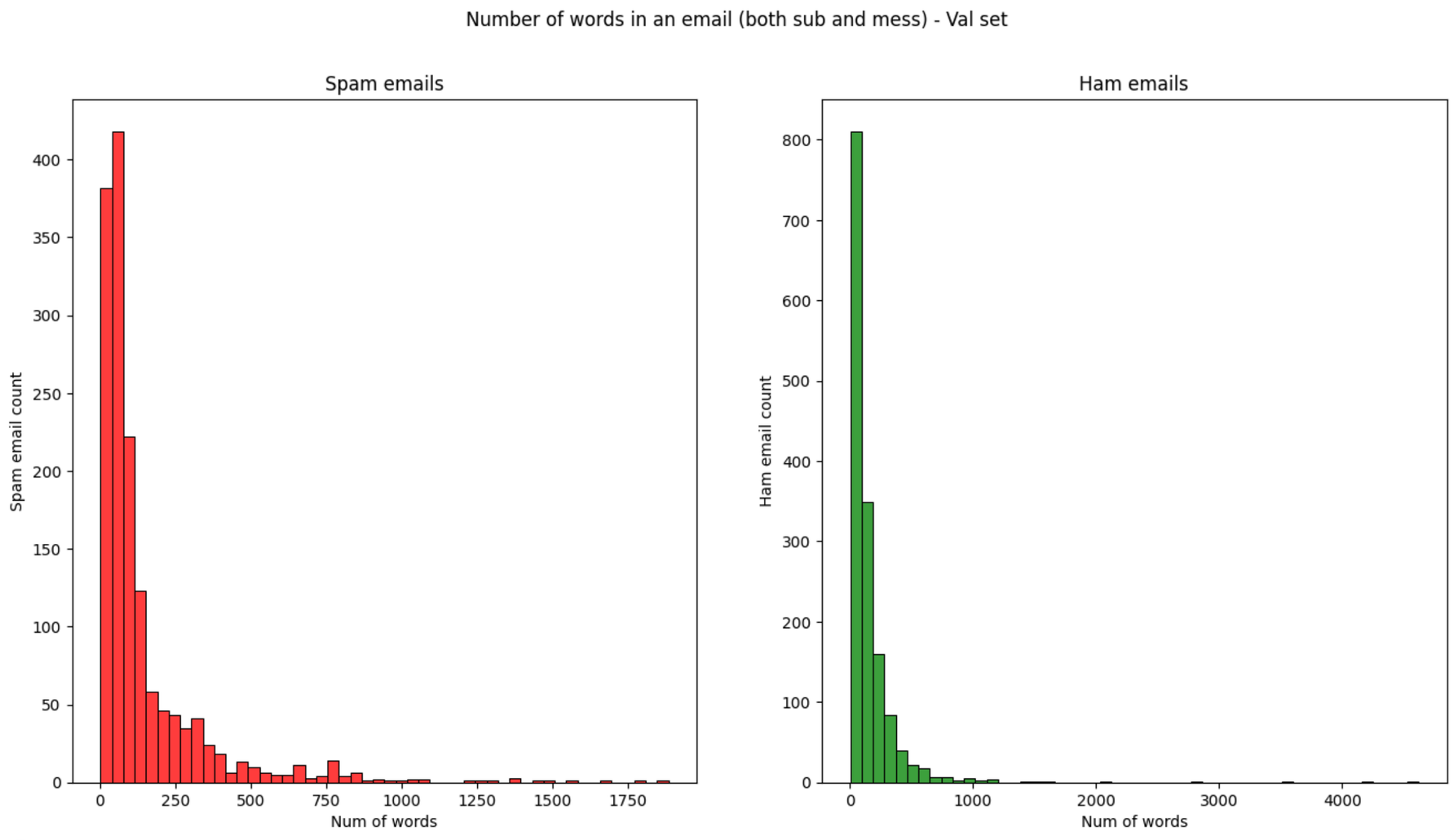
Phân bố nhãn trong tập kiểm tra lại

Số lượng từ trong một email đa phần ở mức từ 2 đến 200 từ

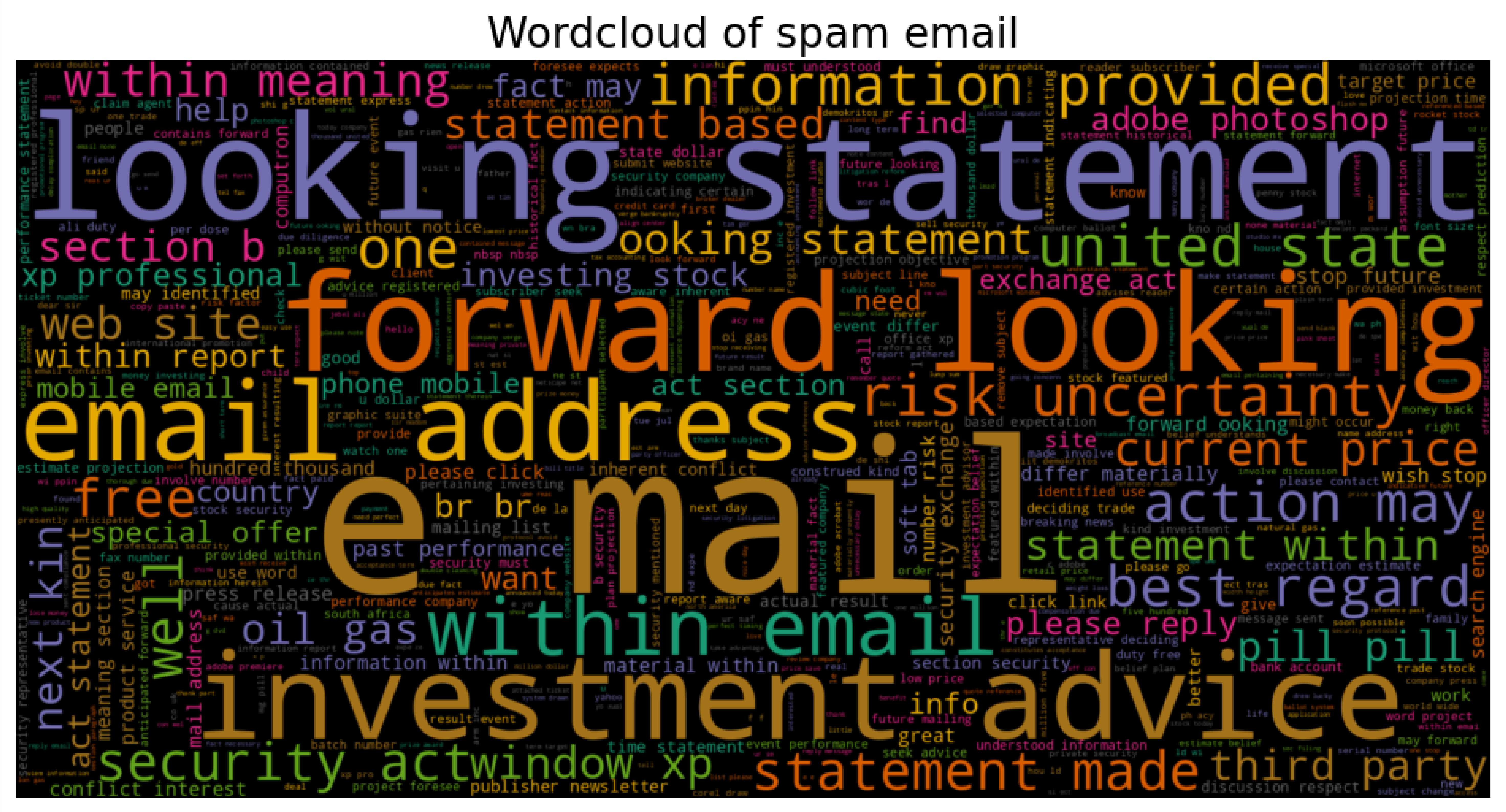


Phân bố số từ trong email của tập huấn luyện

Số lượng từ trong một email đa phần ở mức từ 2 đến 200 từ

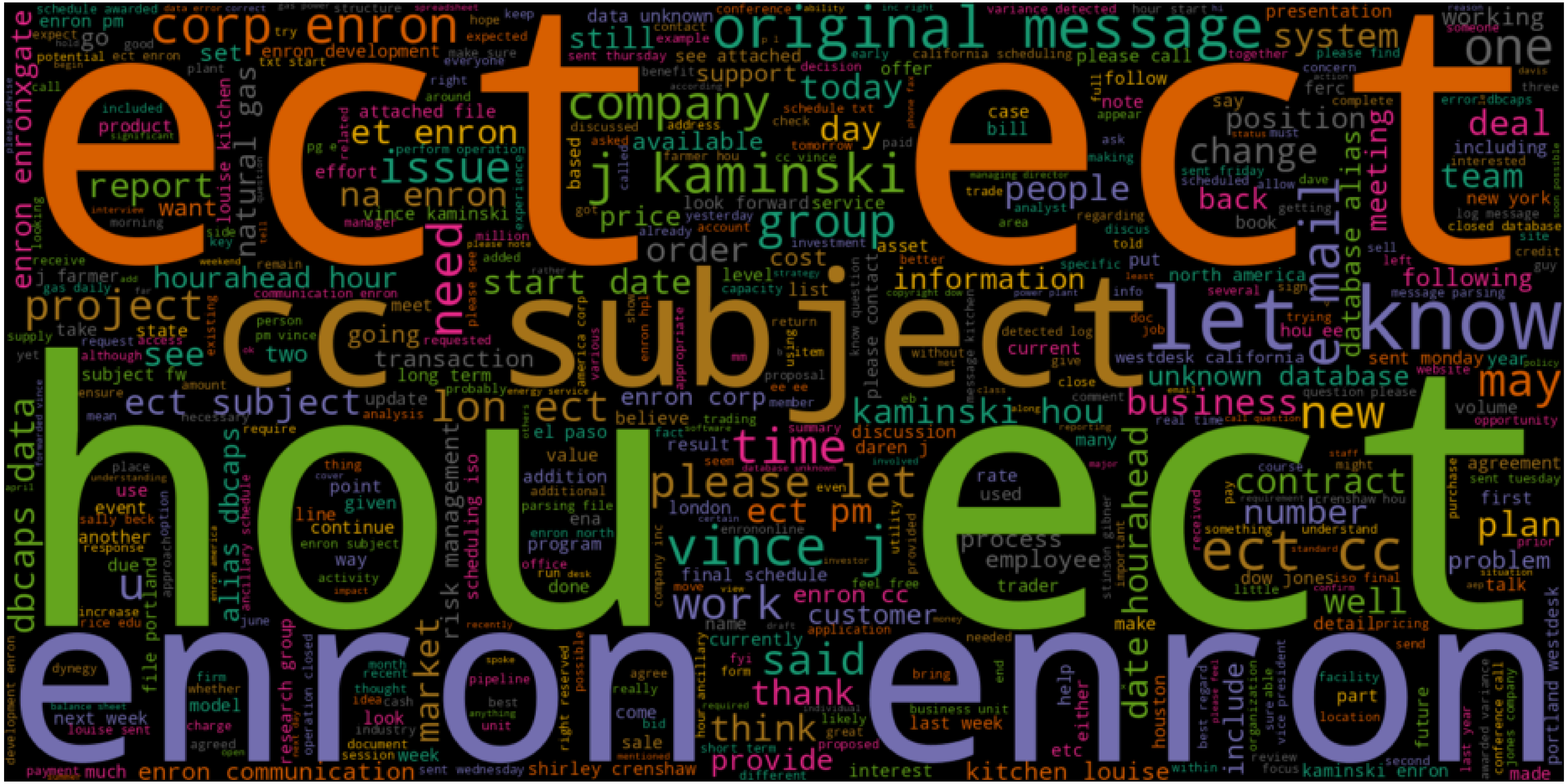


Phân bố số từ trong email của tập kiểm tra lại



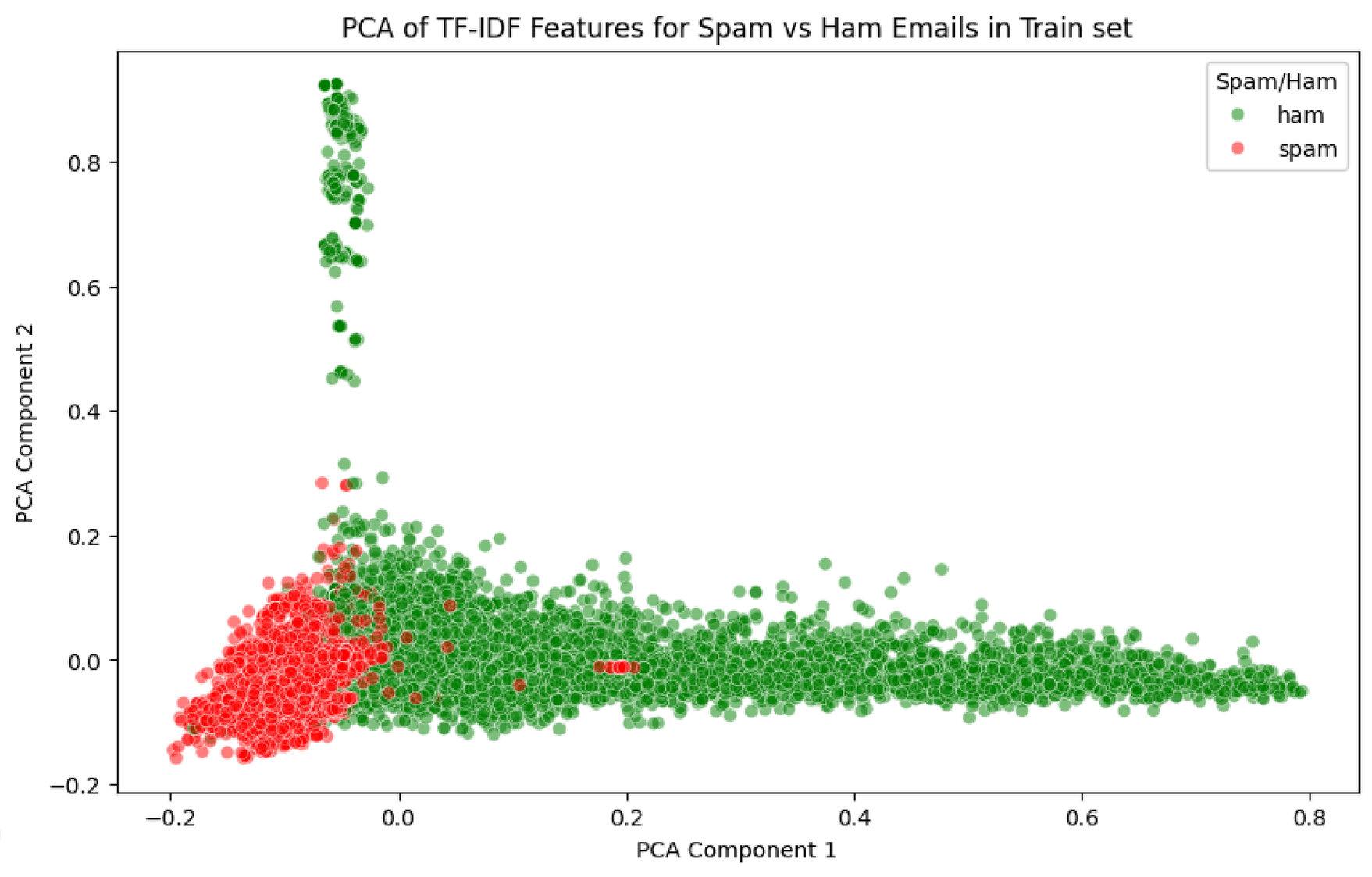
Các từ, cụm từ thường xuất hiện trong một spam email

Wordcloud of ham email



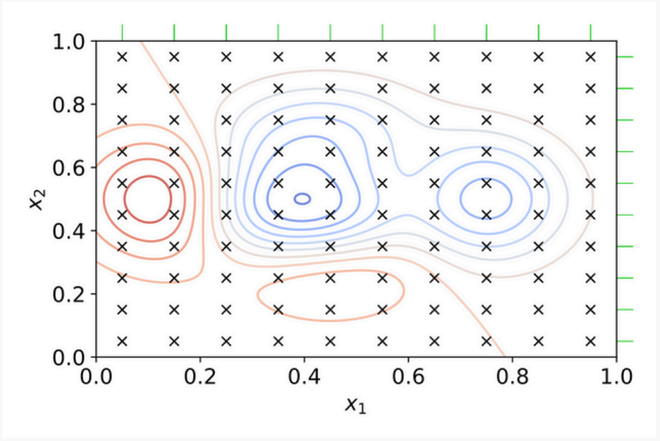
Các từ, cụm từ thường xuất hiện trong một ham email

Minh họa sự phân tách dữ liệu thông qua đặc trưng TF-IDF



Phương pháp đề xuất

Phương pháp đề xuất



Mô hình

Tiến hành thử nghiệm phân lớp với nhiều mô hình, phương pháp khác nhau để có nhiều lựa chọn

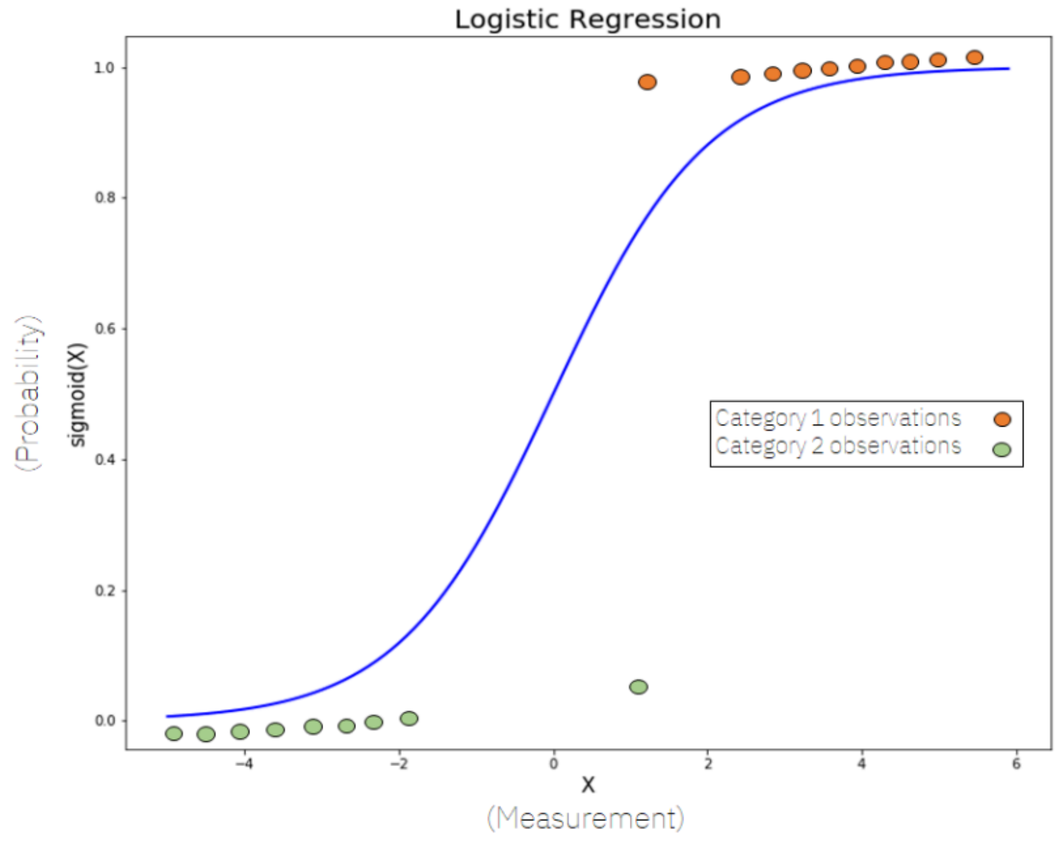
Ensemble

Phương pháp kết hợp dự đoán của nhiều mô hình để cải thiện độ chính xác và độ ổn định của quá trình phân loại.

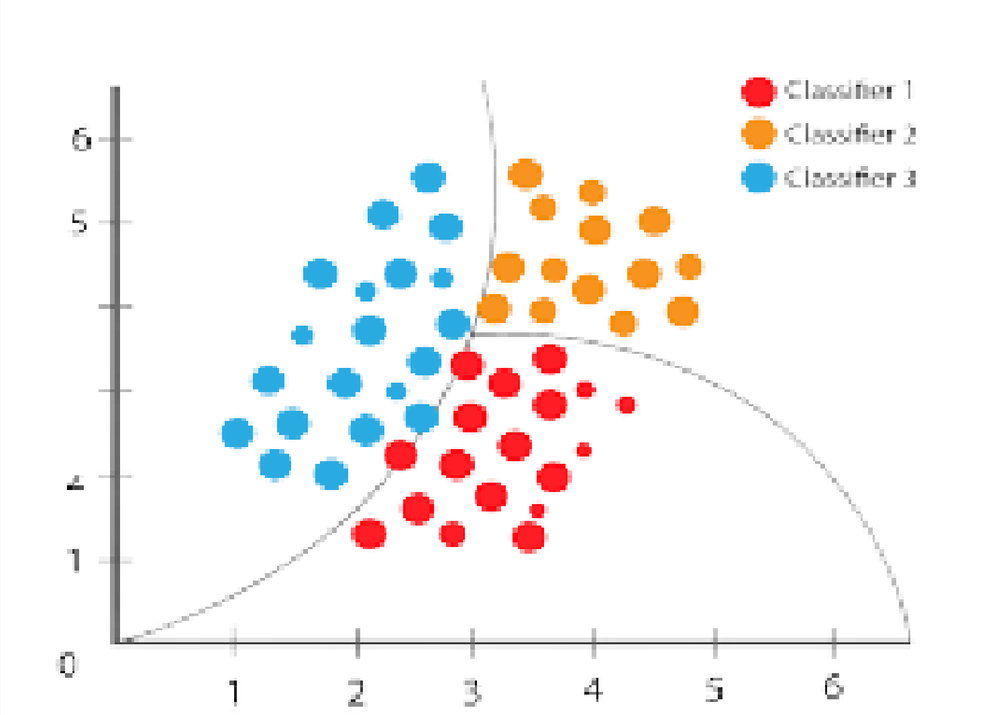
Hyperparameter tuning

Sử dụng phương pháp Grid Search để thử nghiệm và chọn ra các siêu tham số tốt nhất cho mô hình.

Thử nghiệm mô hình

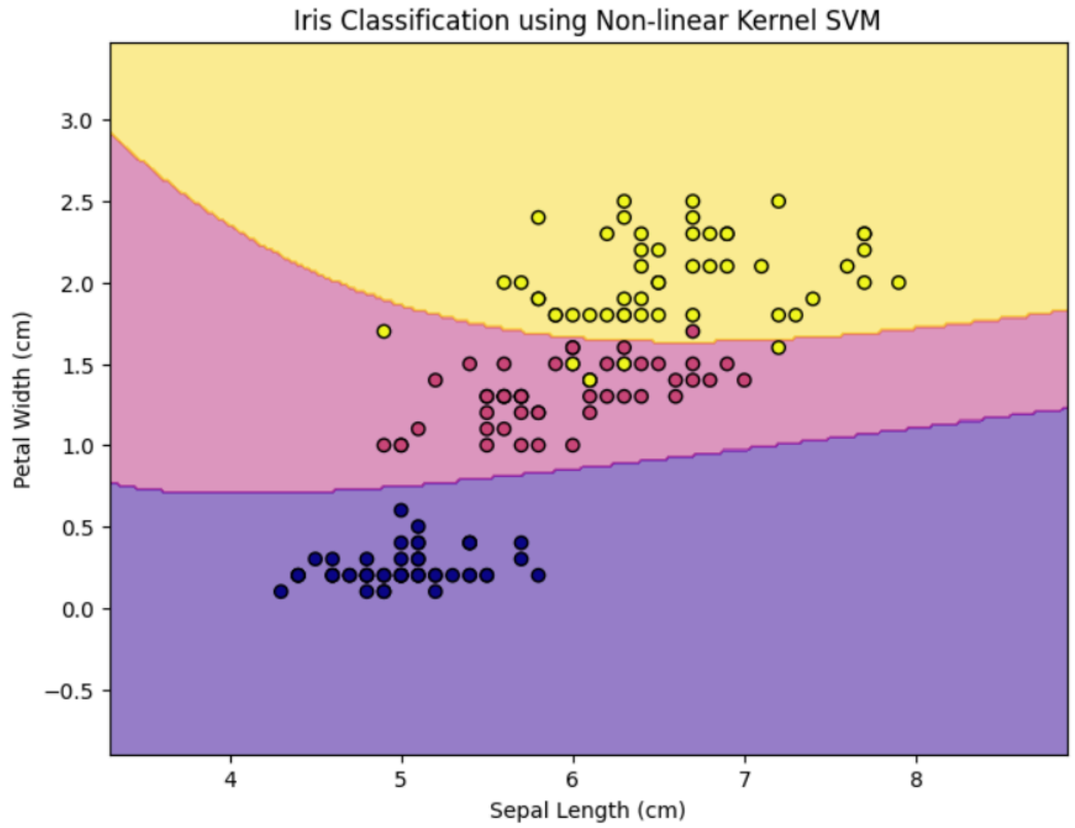


Phân lớp bằng Hồi quy logistic

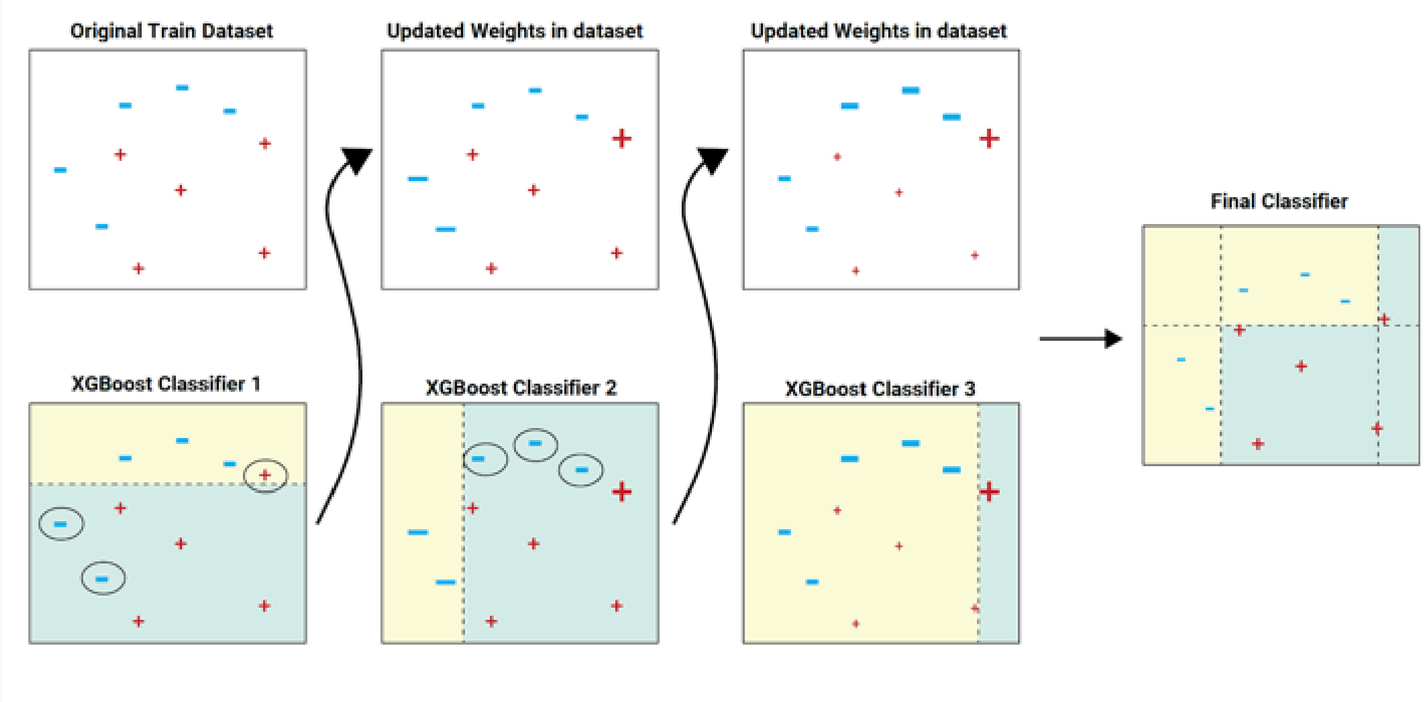


Phân lớp bằng Naive Bayes Classifier

Thử nghiệm mô hình

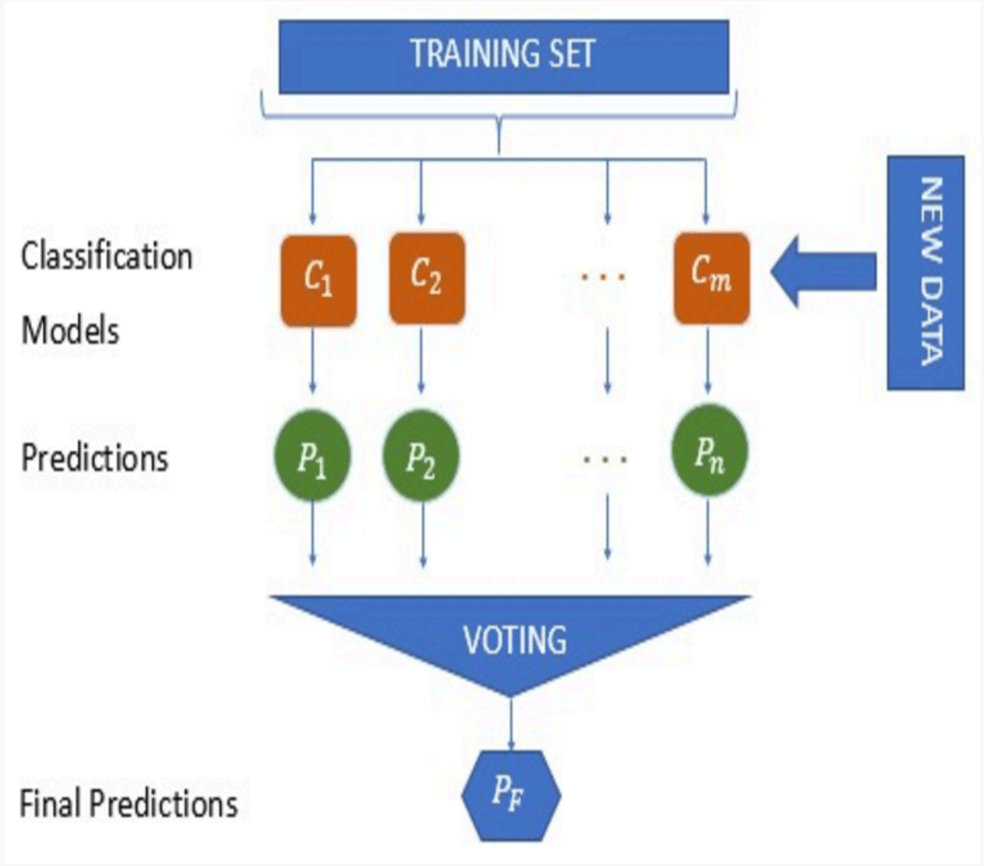


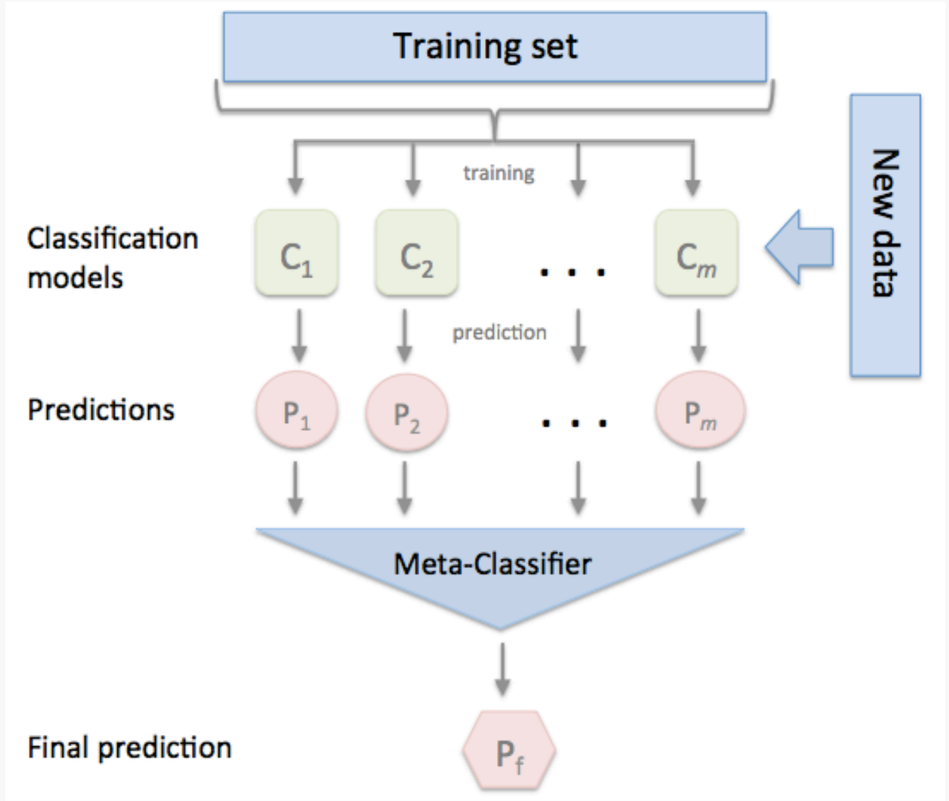
Phân lớp bằng SVM



Phân lớp bằng XGBoost (Gradient Boosting)

Phương pháp tổng hợp mô hình



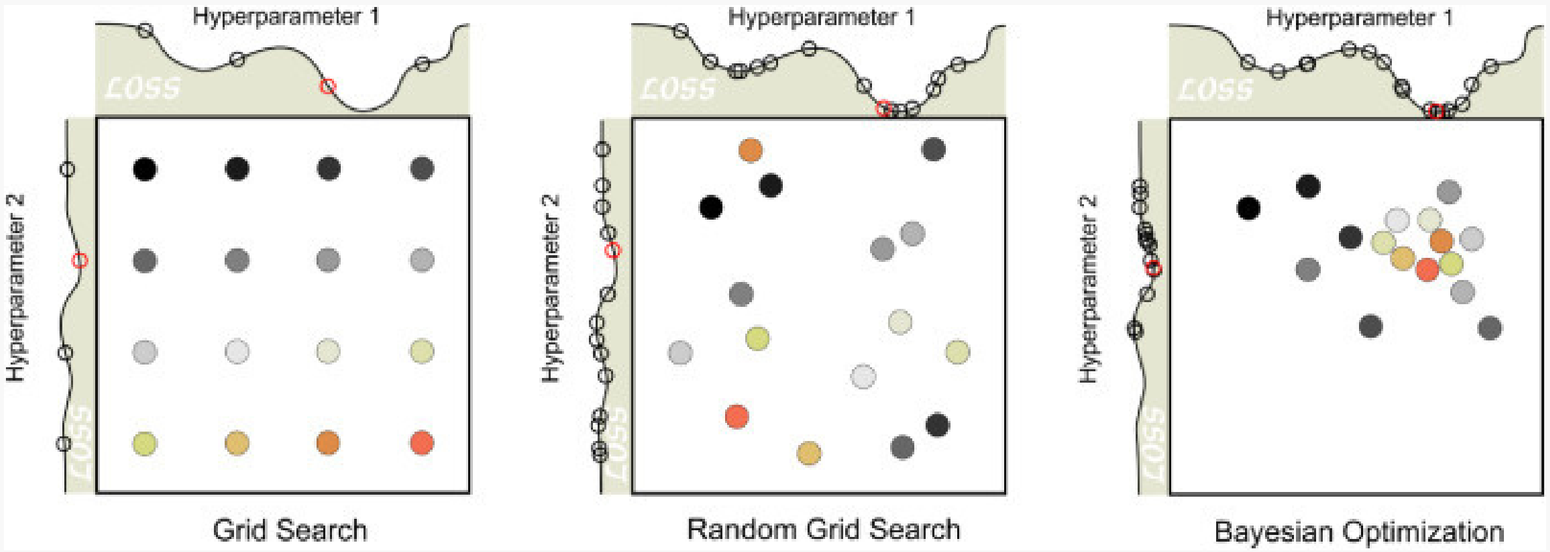


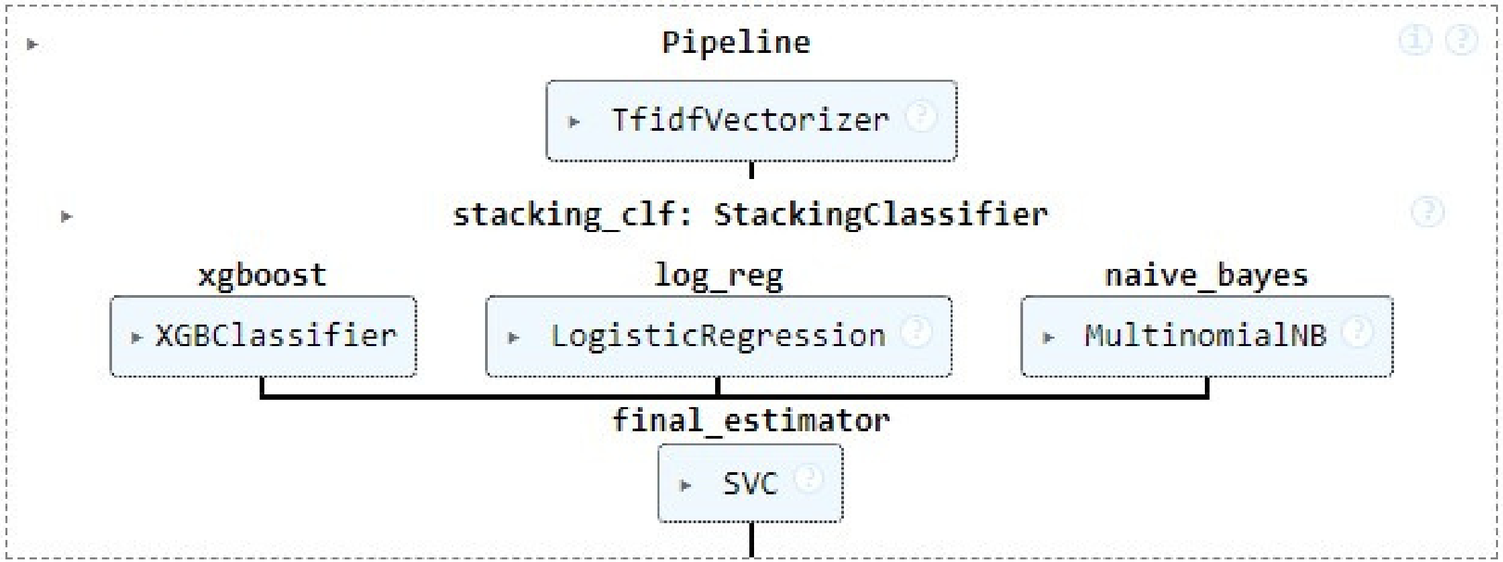
Phương pháp tổng hợp bằng bầu chọn

Phương pháp tổng hợp xếp chồng

Hyperparameter Tuning

Phương pháp tìm kiếm trên lưới





Tổng quan kiến trúc pipeline của nhóm

Demo

Thanks for listening!