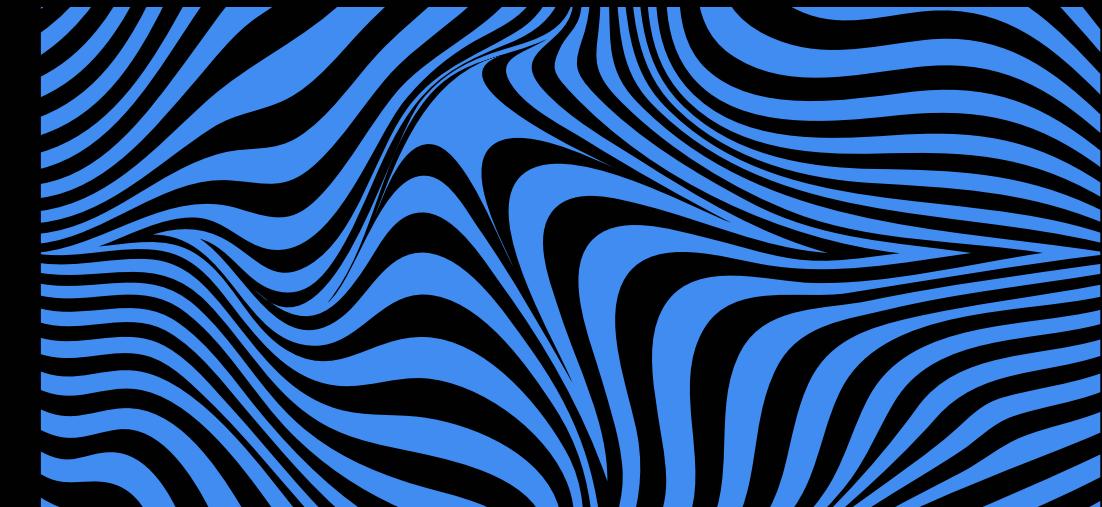
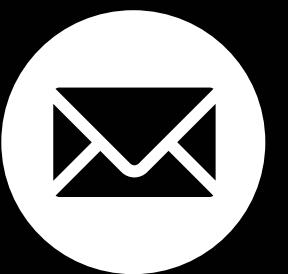


NHÓM 3



## ĐỀ TÀI

MÔ HÌNH HỌC MÁY LỌC THƯ SPAM DỰA TRÊN  
PHƯƠNG PHÁP XÂY DỰNG MẠNG NƠ-RON TUẦN TỤ VỚI PYTHON



NHÓM 3



# THÀNH VIÊN

PHẠM THỊ THANH HẰNG  
NGUYỄN ĐÌNH SƠN  
NGUYỄN MINH ĐỨC  
TỔNG QUANG NAM  
NGUYỄN XUÂN THÚC



HAM OR SPAM



# NỘI DUNG CHÍNH

- ★ 1. Giới thiệu
- ★ 2. Nội dung
- ★ 3. Kết luận
- ★ 4. Tài liệu liên quan



# 1. GIỚI THIỆU

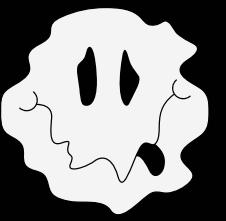
## a. Hiện trạng

- Việc sử dụng thư điện tử ngày càng phổ biến
- Thư rác, hay thư spam tăng, gây phiền hà và tiêu cực cho người nhận

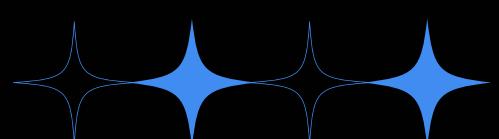
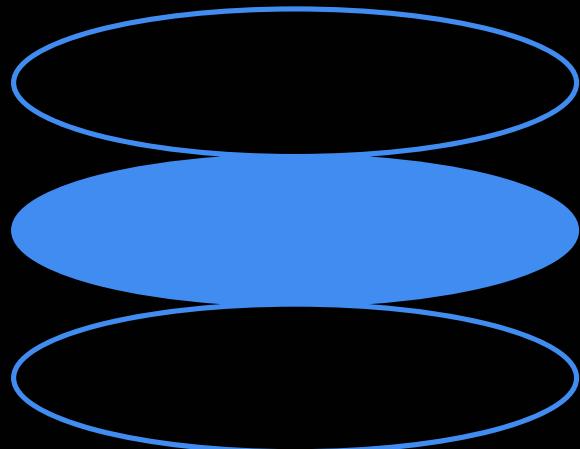
→ Nhu cầu loại bỏ thư spam ngày càng cao.



## b. Những phương pháp lọc thư rác đã và đang được sử dụng



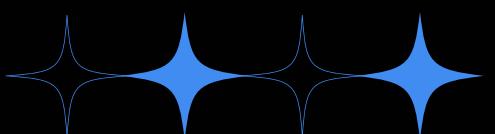
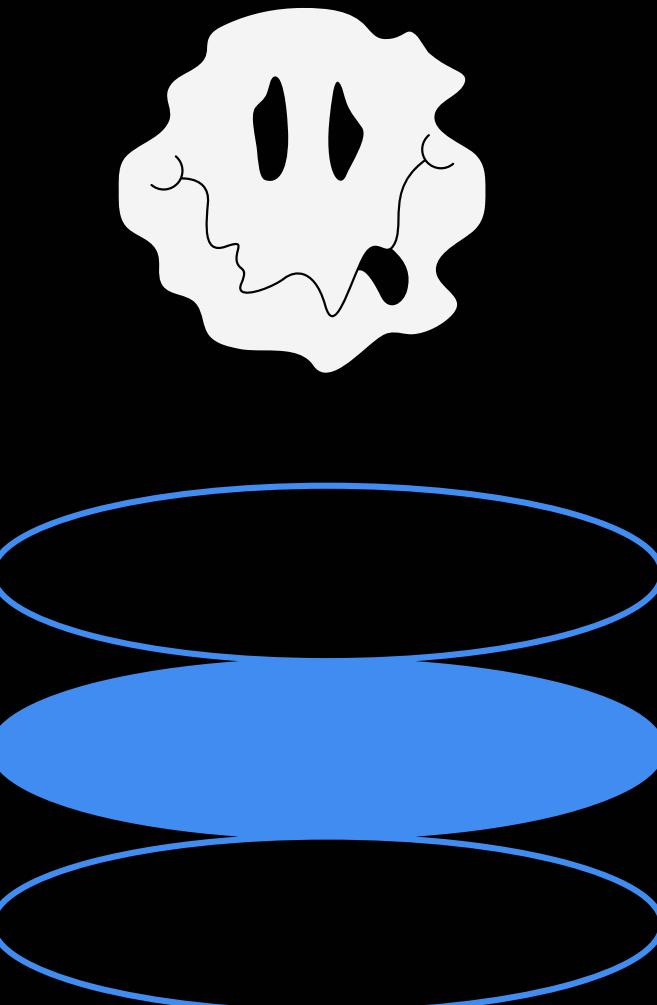
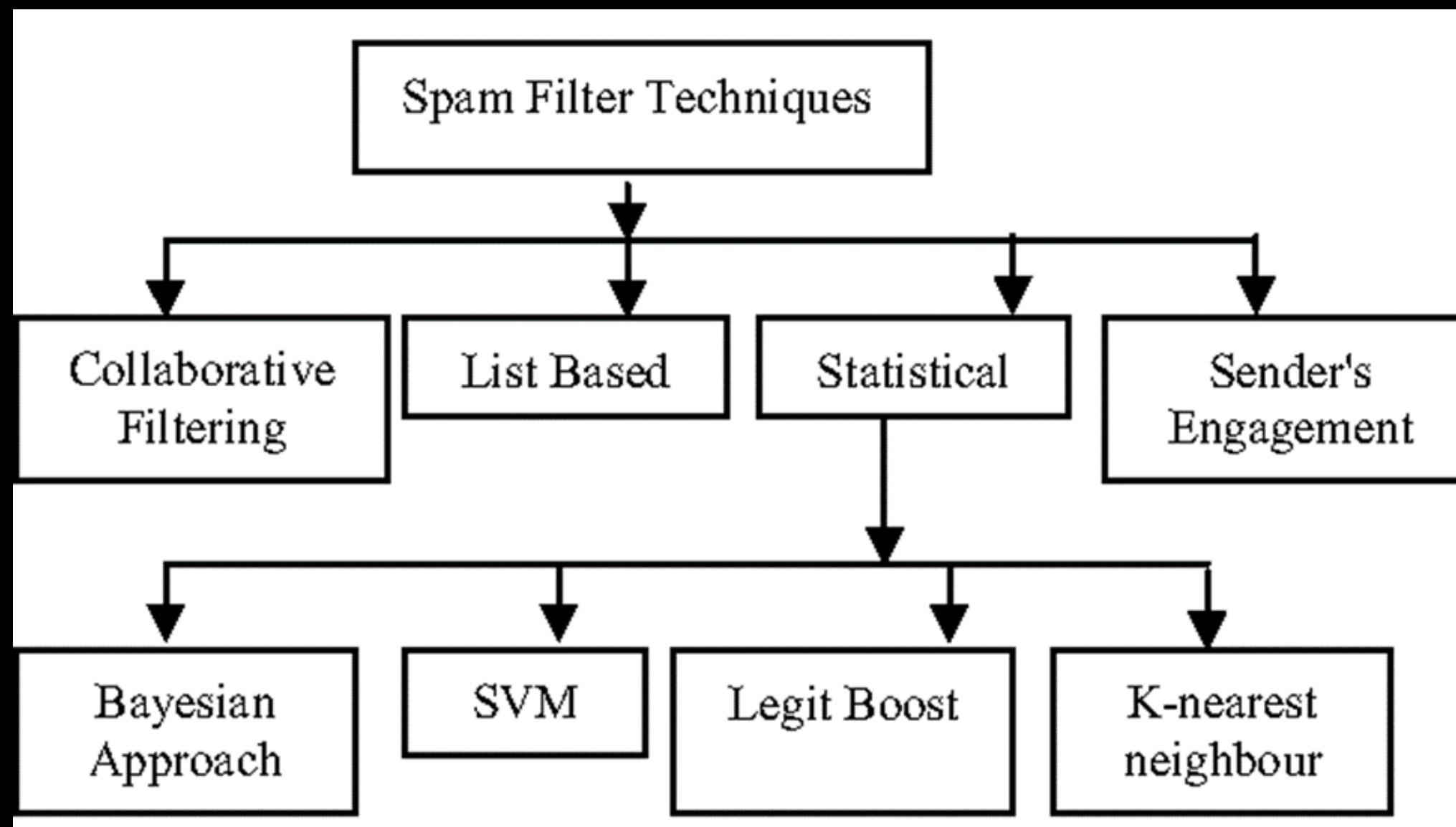
- Từng sử dụng nhiều kỹ thuật để lọc thư rác như Rule-based Filtering, Bayesian Filtering, Blacklist và Whitelist
- ▶ Không hiệu quả đối với thư spam ngày càng phức tạp và đa dạng



SPAM OR HAM

## b. Những phương pháp lọc thư rác đã và đang được sử dụng

- ▶ Thúc đẩy sử dụng học máy và học sâu để xử lý và lọc thư rác một cách hiệu quả hơn

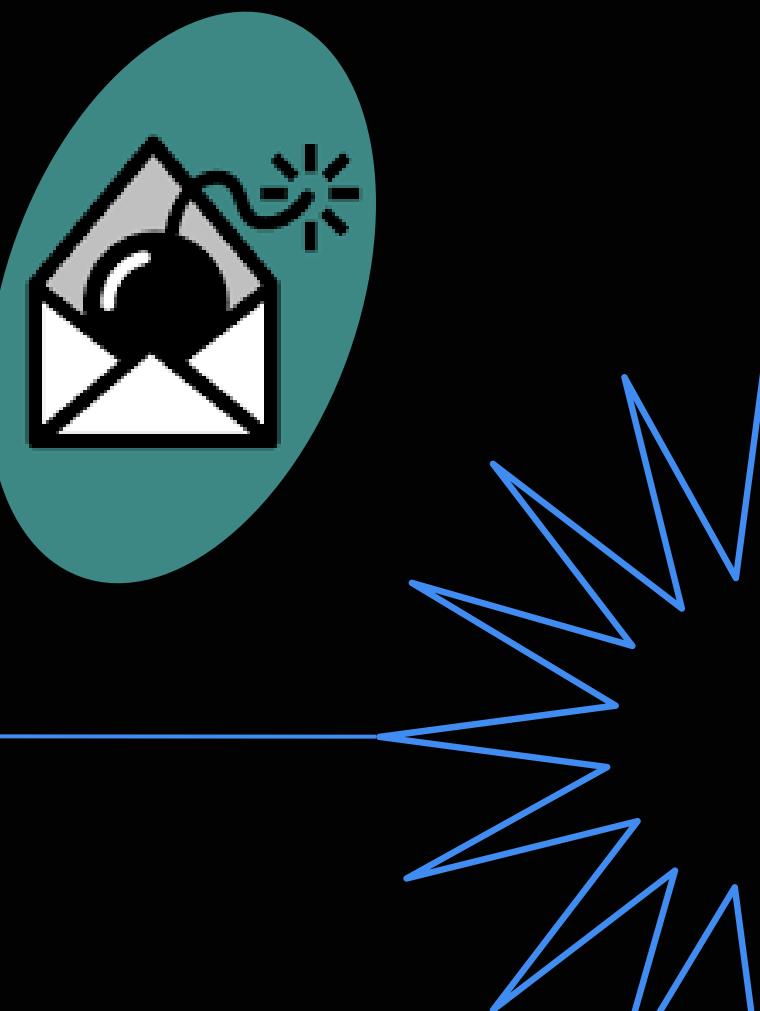
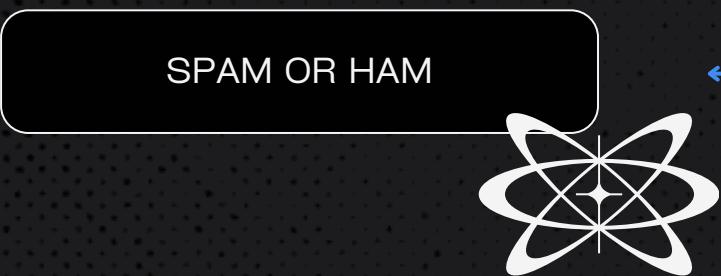


## c. Áp dụng học máy học sâu vào phân loại thư rác



- ★ Học máy và học sâu đã được áp dụng rộng rãi, bao gồm thị giác máy tính, xử lý âm thanh, ngôn ngữ tự nhiên, tự lái xe,...
- ★ Phân loại thư rác: "Phân loại nhị phân"
- ★ Thuật toán phổ biến: Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Tree, và Support Vector Machine

### ► Sử dụng Artificial Neural Network (ANN)

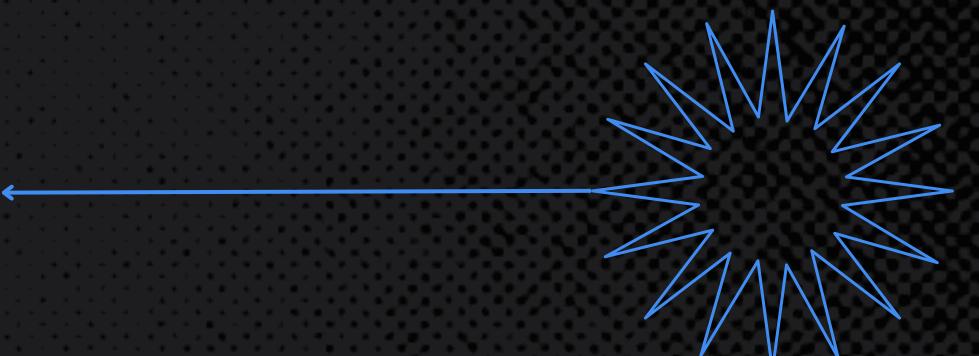


## c. Áp dụng học máy học sâu vào phân loại thư rác

- Tự động học và phát hiện các biểu hiện phức tạp của thư rác
- Điều chỉnh các tham số để cải thiện hiệu suất giúp thích nghi tốt hơn với các biểu hiện của thư rác
- Kết hợp các phương pháp tiền xử lý dữ liệu và khả năng xử lý dữ liệu phi cấu trúc giúp tăng cường hiệu suất của mô hình.



SPAM OR HAM



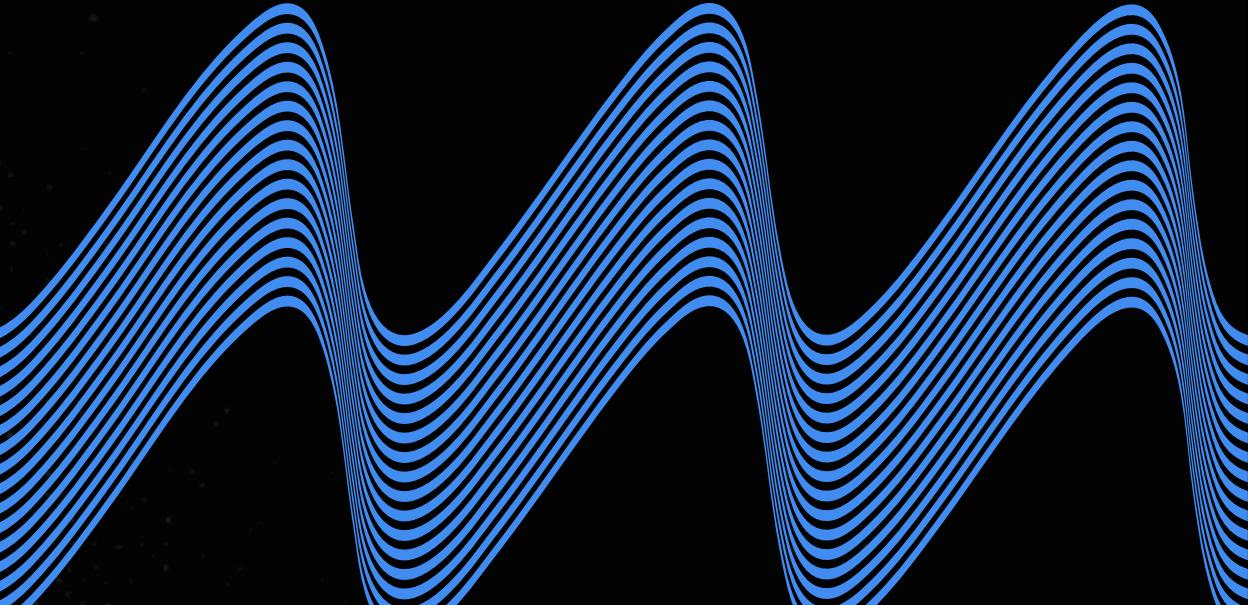
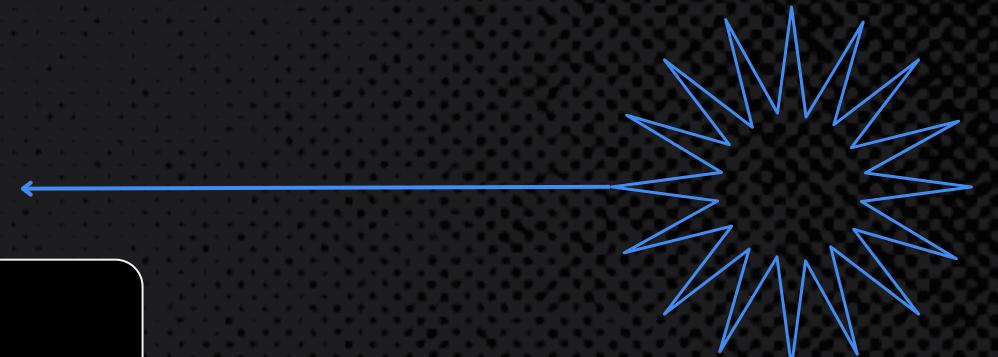
## c. Áp dụng học máy học sâu vào phân loại thư rác



### ► Chất lượng của mô hình

- Tập dữ liệu đầu vào phải cập nhật liên tục
- Mô hình điều chỉnh để đối phó với các kỹ thuật thay đổi của thư rác

SPAM OR HAM



# 2. NỘI DUNG

## a. Thu thập dữ liệu

- Tự động học và phát hiện các biểu hiện phức tạp của thư rác
- Điều chỉnh các tham số để cải thiện hiệu suất giúp thích nghi tốt hơn với các biểu hiện của thư rác
- Kết hợp các phương pháp tiền xử lý dữ liệu và khả năng xử lý dữ liệu phi cấu trúc giúp tăng cường hiệu suất của mô hình.

-----SPAM OR HAM-----

Các bước cần có





## b. Chuẩn bị phần mềm

- Cài đặt một số thư viện với lệnh: “**pip install**”
- Các thư viện bao gồm:
  - **Pandas:** Xử lý và phân tích dữ liệu
  - **Flask:** Framework phát triển ứng dụng web
  - **NLTK (Natural Language toolkit):** Xử lý ngôn ngữ tự nhiên
  - **Wordcloud:** Tạo đám mây từ khóa
  - **TensorFlow:** Hỗ trợ machine learning và deep learning
  - **Scikit-learn:** Xây dựng và huấn luyện mô hình học máy
  - **Matplotlib và Seaborn:** Tạo biểu đồ trực quan hóa dữ liệu

Thư viện quan trọng nhất TensorFlow



## c.Tiền xử lí dữ liệu

HAM OR SPAM

- Đọc tập dữ liệu
  - Cột 1 ‘spam’: là spam hay không
  - Cột 2 ‘text’: nội dung thư

`data_en.head()`

✓ 0.0s

	spam	text
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...

`data_vi.head()`

✓ 0.0s

	spam	text
0	0.0	Hãy đến cho đến Jurong Point, điên rồ .. Chỉ c...
1	0.0	Ok lar ... nói đùa wif u oni ...
2	1.0	Nhập viện miễn phí trong 2 WKLY Comp để giành ...
3	0.0	Bạn không nói quá sớm ... bạn đã nói ...
4	0.0	Không, tôi không nghĩ anh ấy đến USF, anh ấy s...

## c.Tiền xử lý dữ liệu

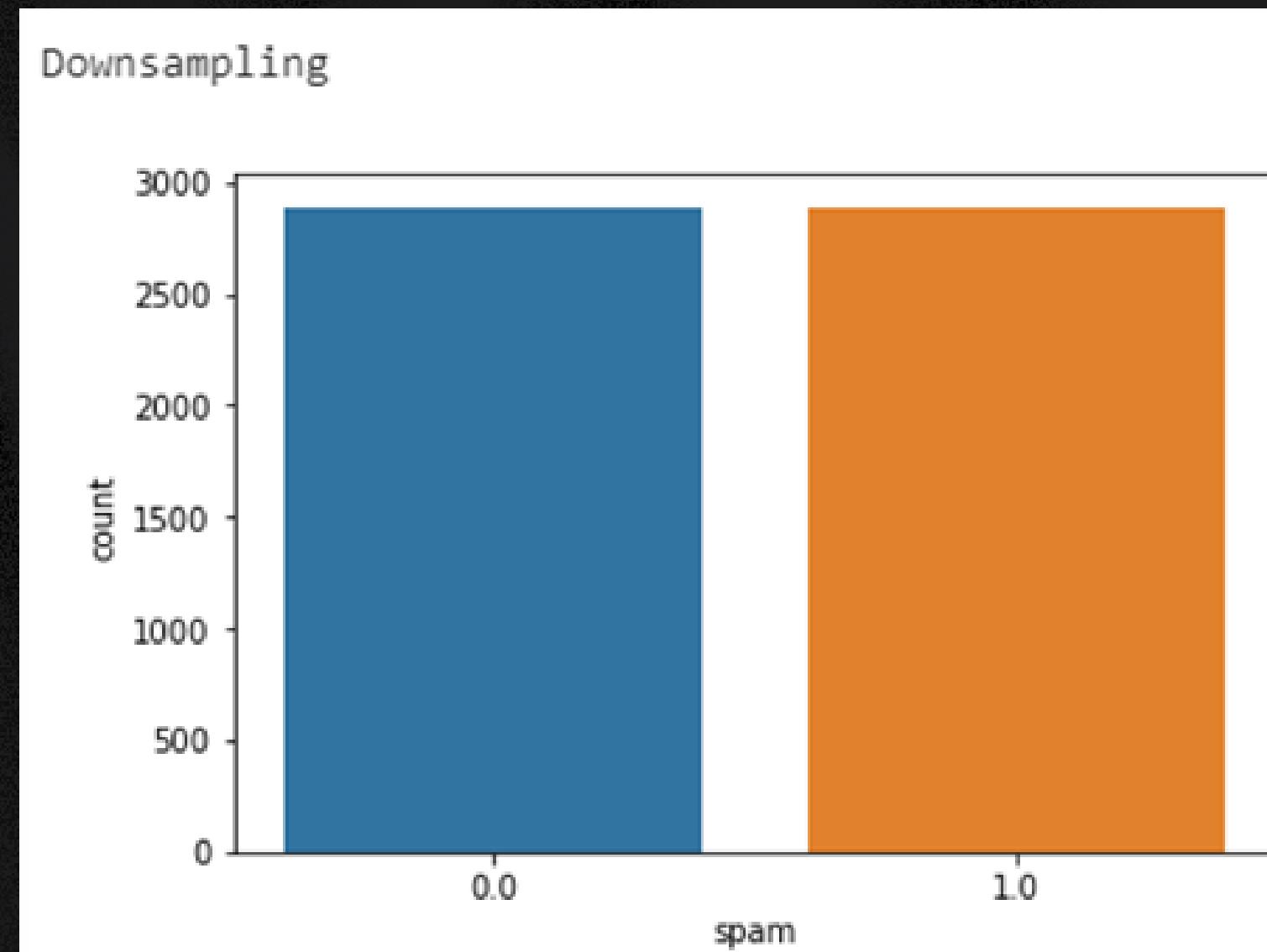
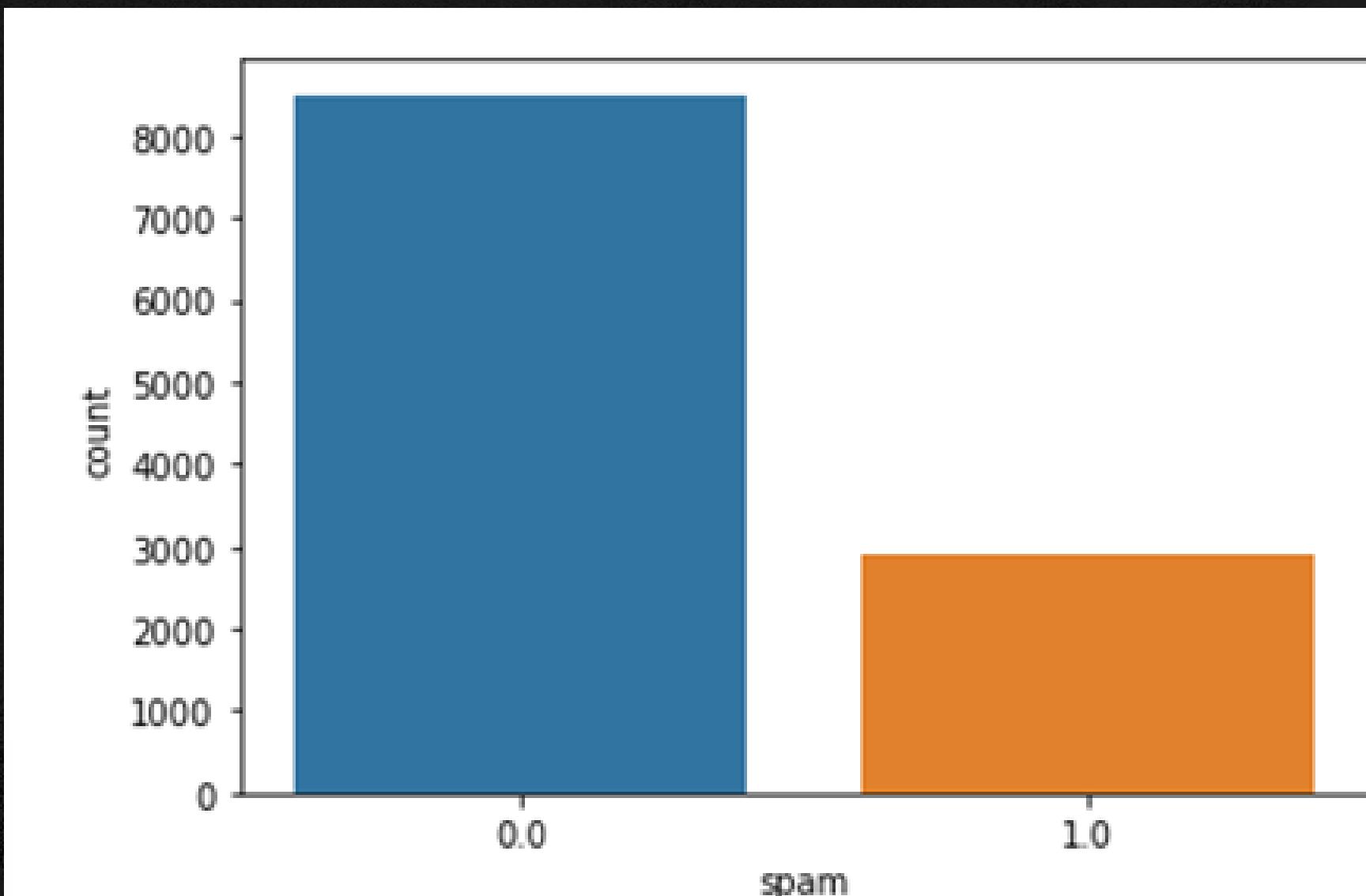
HAM OR SPAM



### STEP 1

#### Cân bằng tập dữ liệu

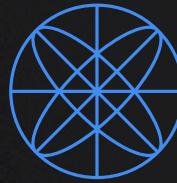
- Tránh hiện tượng mô hình học lệch về phân dữ liệu



## c.Tiền xử lý dữ liệu

HAM OR SPAM

### STEP 2



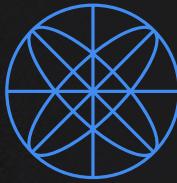
#### ► Xử lý văn bản

- Loại bỏ Stopword
  - Stopword: từ xuất hiện thường xuyên trong văn bản và không mang nhiều ý nghĩa về mặt nội dung
  - Các stopword tiếng việt vì có các từ ghép => loại bỏ khó khăn

## c.Tiền xử lí dữ liệu

HAM OR SPAM

### STEP 3



#### ► Chia tập, số hóa dữ liệu và Padding

- Chia tập:
  - Chia tập huấn luyện và tập kiểm tra: train x, text x, train y, test y
- Số hóa dữ liệu:
  - Chuyển các dữ liệu từ ngũ thành dữ liệu số để máy tính có thể sử dụng
- Padding:
  - Thêm vào cho các chuỗi số có độ dài bằng nhau



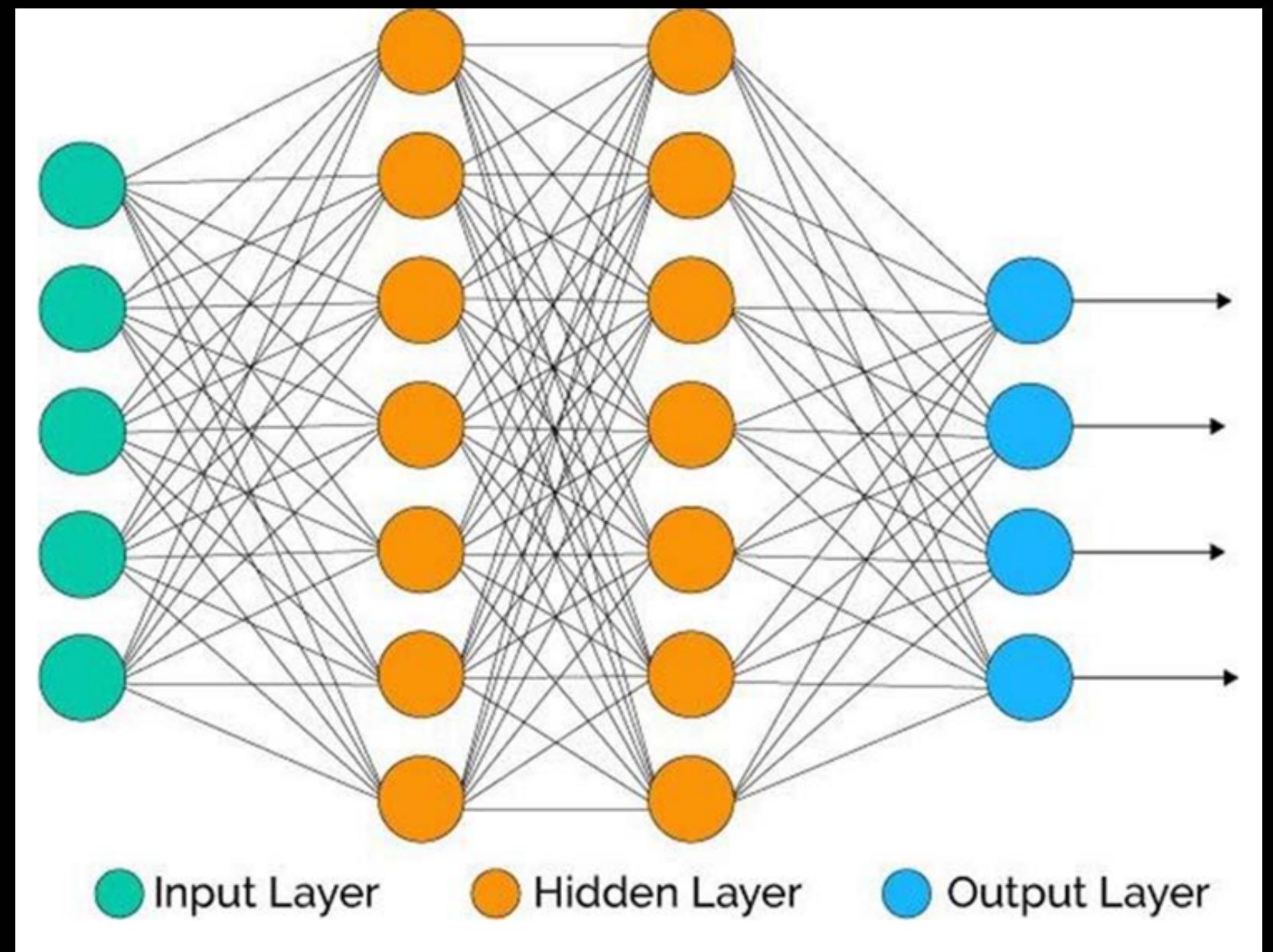
## d. Xây dựng mô hình

- Xây dựng Nơ – ron nhân tạo (ANN)
- Dựa trên cấu tạo của mạng Nơ – ron trong não người
  - Xây dựng mạng Nơ – ron tuần tự (SNN)
  - Cấu tạo của cấu trúc này là tuyến tính các lớp Nơ – ron được xếp chồng lên nhau trình tự.
  - Dữ liệu được truyền qua mạng từ đầu đến cuối theo một hướng

SPAM OR HAM



**Cấu trúc của một mạng  
Nơ-ron nhân tạo mẫu**





# XÂY DỰNG MÔ HÌNH HỌC MÁY LỌC THƯ RÁC, SỬ DỤNG MỘT KIẾN TRÚC MÔ HÌNH



## 4 LAYER

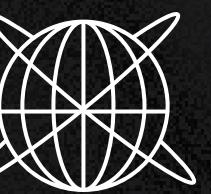
### 1.Embedding Layer

- Thêm một lớp nhúng
- Từng từ trong chuỗi số => vector nhúng
- Mỗi từ trong tập từ vựng ban đầu ánh xạ đến một điểm trong không gian đa chiều (32 chiều trong ví dụ này) bằng một vector
- Tối ưu hóa vector nhúng sao cho các từ có nghĩa gần nhau trong không gian vector nhúng cũng gần nhau về khoảng cách Euclidean

### 2.LSTM

- Thêm một lớp LSTM với 16 unit
- Học ngữ nghĩa thông qua chuỗi các vector nhúng
- Giữ trạng thái ẩn để lưu trữ thông tin ngữ nghĩa => hiểu và lưu trữ thông tin từ quá khứ
- Số 16 là số unit được sử dụng để học dữ liệu. Một số lớn có thể giúp mô hình hiểu các ngữ nghĩa phức tạp hơn, nhưng có thể gây ra overfitting và tăng thời gian chạy

SPAM OR HAM



# XÂY DỰNG MÔ HÌNH HỌC MÁY LỌC THƯ RÁC, SỬ DỤNG MỘT KIẾN TRÚC MÔ HÌNH



## 4 LAYER

### 3. Dense Layer

- Một lớp dense được thêm vào để học các mối quan hệ tuyến tính trong dữ liệu
- Dữ liệu được biểu diễn bằng các hàm tuyến tính
- Học mối quan hệ phi tuyến tính => hàm kích hoạt ReLU (Rectified Linear Unit)
- Có 32 unit nơ-ron ẩn kết nối với tất cả các unit nơ-ron đầu vào và đầu ra



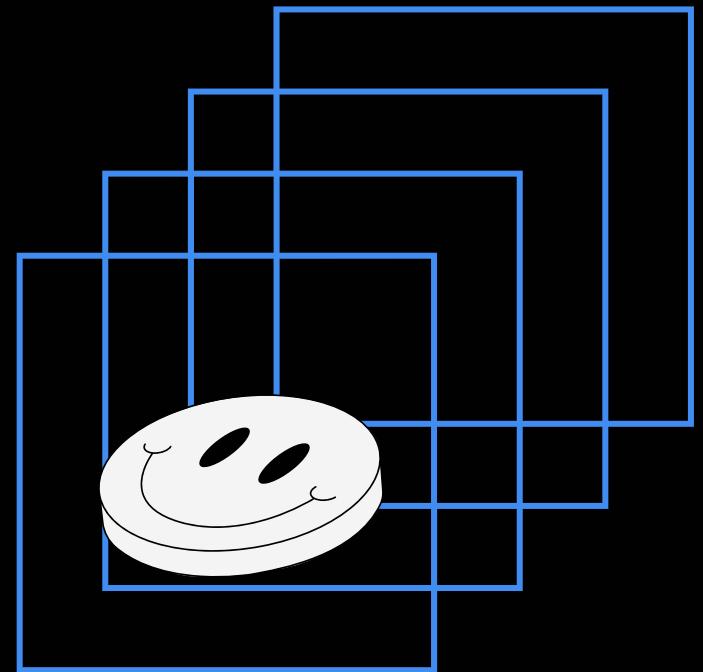
### 4. Output Layer

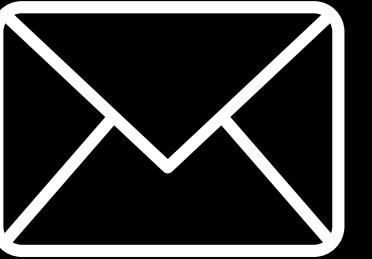
- Để tạo một giá trị dự đoán
- Lớp này có 1 unit và sử dụng hàm kích hoạt Sigmoid
- Với mục tiêu phân loại nhị phân, ngưỡng (threshold) là 0.5. Giá trị đầu ra từ lớp này được so sánh với ngưỡng để xác định lớp dự đoán (0 hoặc 1)



## CÂU HÌNH HUÂN LUYỆN

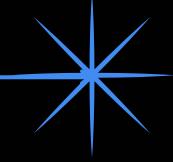
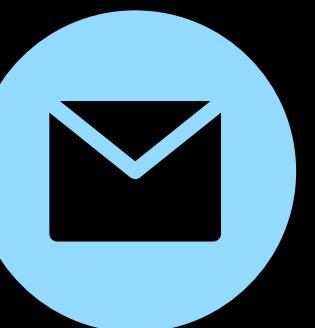
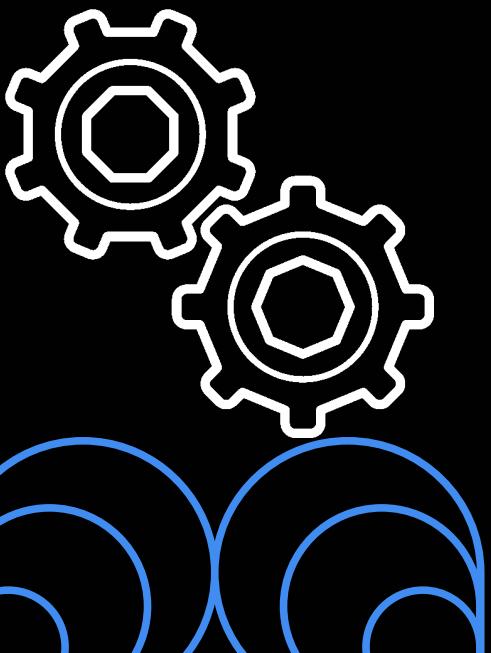
- Hàm `model.compile` : cấu hình mô hình
- `Optimizer` : thuật toán tối ưu được sử dụng cho mô hình
  - 'Adam' thuật toán hiệu quả cho việc điều chỉnh trọng số mô hình trong quá trình học
- `Metrics` : đánh giá mô hình thông qua một số độ đo, ở đây là 'accuracy'
- `Loss` : tính toán tỷ lệ lỗi giữa kết quả dự đoán và kết quả thực tế trong tập dữ liệu





# 3. KẾT LUẬN

- Tập trung vào mạng nơ-ron tuần tự và các phương pháp xử lý dữ liệu để giải quyết bài toán phân loại thư rác
  - Đặc biệt ứng dụng trong ngôn ngữ phức tạp như tiếng Việt



7255

74510

877  
96

059234

3

274410

901

7482

# KẾT LUẬN

Mô hình học máy

Phương pháp tiền xử lý dữ liệu

Tối ưu mô hình

Kết quả đáng tin cậy

So sánh với các phương pháp khác

Sử dụng một mô hình mạng nơ-ron tuần tự được xây dựng bằng các lớp mạng nơ-ron tuần tự để giải quyết bài toán

Tối ưu mô hình được thực hiện để đảm bảo mô hình có khả năng dự đoán chính xác

So sánh: KNN, Naïve Bayes, Logistic Regression và Support Vector Machine

Mô hình được kết hợp với các phương pháp tiền xử lý dữ liệu, đảm bảo rằng dữ liệu được chuẩn bị một cách phù hợp trước khi đưa vào mô hình

Khả năng phân loại thư rác ở mức độ đáng tin cậy, cả nội dung thư bằng tiếng Việt





# So sánh



SPAM OR HAM

KNN (K - Nearest Neighbor)

Accuracy on training data\_knn : 0.8502961175696425  
Accuracy on test data\_knn : 0.849561403508772

Logistic Regression

Accuracy on training data\_LgR : 0.9644658916429042  
Accuracy on test data\_LgR : 0.9508771929824561

Naïve Bayes

Accuracy on training data\_nb : 0.9520728229874973  
Accuracy on test data\_nb : 0.9236842105263158

SVM (Support Vector  
Machine)

Accuracy on training data\_svm : 0.9992322877824085  
Accuracy on test data\_svm : 0.9679824561403508

**Kết quả cho thấy mô hình mạng nơ-ron tuần tự hiệu quả hơn KNN,  
tương đương với Naïve Bayes và tiệm cận với Logistic Regression và  
Support Vector Machine**



THANKS FOR LISTENING

