TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NHẬP MÔN HỌC MÁY VÀ KHAI PHÁ DỮ LIỆU

ĐỀ TÀI : LỌC THƯ RÁC

Giảng viên hướng dẫn: TS. Thân Quang Khoát

Nhóm sinh viên thực hiện:

DE PUNLEU	20190150
BRAK LIHOU	20200836
SREY SOVANRITH	20200845
VEN CHHUT	20200844

HÀ NỘI, Tháng 06, 2023

Thông tin nhóm :

Họ và Tên	Mã số sinh viên	Email
DE PUNLEU	20190150	punleu.d190150@sis.hust.edu.vn
BRAK LIHOU	20200836	Lihou.b200836@sis.hust.edu.vn
SREY SOVANRITH	20200845	sovanrith.s200845@sis.hust.edu.vn
VEN CHHUT	20200844	chhut.v200844@sis.hust.edu.vn

MỤC LỤC

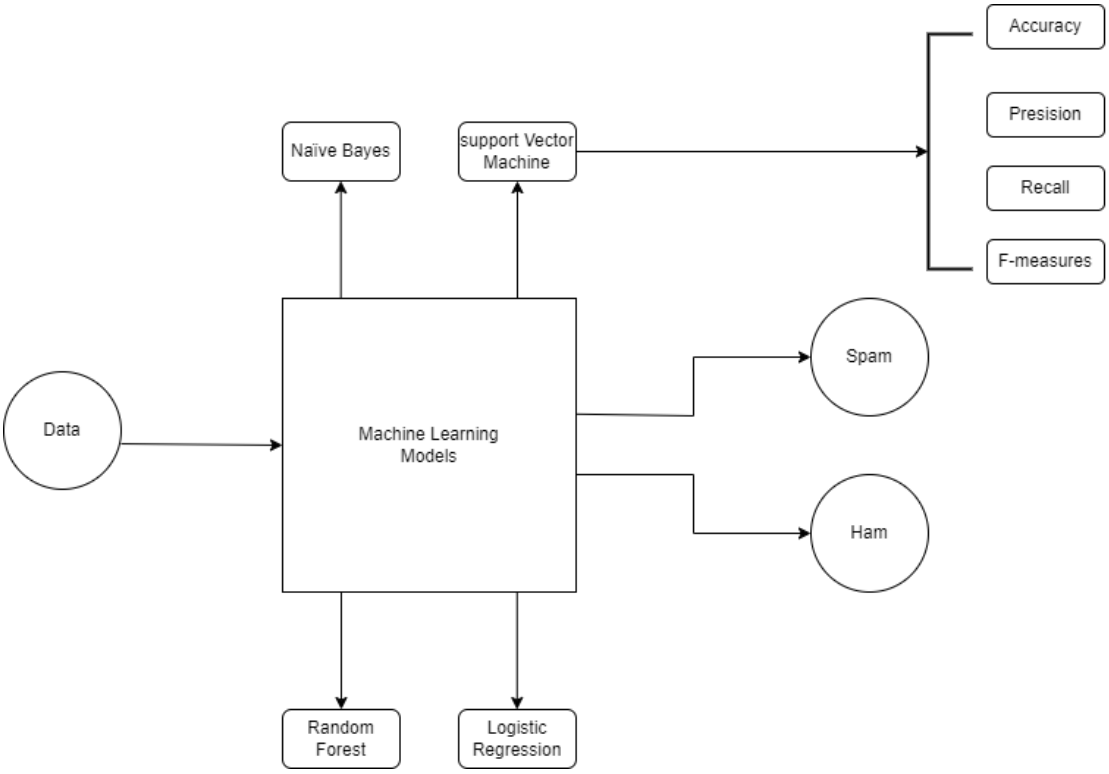
I. GIỚI THIỆU

Mạng internet đã trở thành một phần không thể thiếu trong cuộc sống con người, nơi hơn 4 tỷ người dùng internet tìm thấy nó thuận tiện để sử dụng cho mục đích của họ. Hơn nữa, email được coi là một hình thức giao tiếp đáng tin cậy bởi người dùng internet.

Trong nhiều thập kỷ, các dịch vụ e-mail đã phát triển thành một công cụ mạnh mẽ để trao đổi các loại thông tin khác nhau. Việc sử dụng e-mail ngày càng nhiều cũng kéo theo nhiều cuộc tấn công thư rác hơn đối với người dùng Internet. Thư rác có thể được gửi từ mọi nơi trên hành tinh từ những người dùng có ý định lừa đảo có quyền truy cập Internet. Thư rác là những email không mong muốn và không mong muốn được gửi đến những người nhận không muốn hoặc không cần chúng. Những email spam này có nội dung giả mạo với hầu hết các liên kết dành cho các cuộc tấn công lừa đảo và các mối đe dọa khác, đồng thời những email này được gửi hàng loạt đến một số lượng lớn người nhận. Mục đích đằng sau của chúng là đánh cắp thông tin cá nhân của người dùng và sau đó sử dụng chúng trái với ý muốn của họ để đạt được những lợi ích vật chất. Những email này chứa nội dung độc hại hoặc có URL dẫn đến nội dung độc hại. Những email như vậy đôi khi còn được gọi là email lừa đảo.

Bất chấp sự tiến bộ của các ứng dụng và dịch vụ lọc thư rác, không có cách nào chắc chắn để phân biệt giữa email hợp pháp và email độc hại do nội dung của những email đó luôn thay đổi. Thư rác đã được gửi trong hơn ba hoặc bốn thập kỷ nay và với sự sẵn có của các dịch vụ chống thư rác khác nhau, thậm chí ngày nay, những người dùng cuối không phải là chuyên gia vẫn bị mắc kẹt trong cạm bẫy ghê tởm đó. Trong trình quản lý e-mail, bộ lọc thư rác phát hiện thư rác và chuyển tiếp thư rác đó đến một không gian dành riêng, thư mục thư rác, cho phép người dùng chọn có truy cập chúng hay không. Các công cụ lọc thư rác như hệ thống email công ty, cổng lọc email, dịch vụ chống thư rác theo hợp đồng và đào tạo người dùng cuối có thể xử lý email rác bằng tiếng Anh hoặc bất kỳ ngôn ngữ nào khác.

Tuy nhiên, không hiệu quả trong việc lọc email spam bằng các ngôn ngữ khác gần đây đã được số hóa, chẳng hạn như tiếng Anh. Báo cáo này mô tả cách thức hoạt động của các mô hình học máy (ML) như Máy vectơ hỗ trợ (SVM), Naive Bayes, Rndom Forest , Logistic Regression , một mạng thần kinh hồi quy, có thể được đào tạo để phát hiện email rác tiếng Anh . Hơn nữa, vì không có tập dữ liệu cho email spam, bài viết này cũng giải thích việc tạo và đào tạo các mô hình máy học khác nhau.



II. Literature Review

Trước khi triển khai mô hình phát hiện thư rác bằng cách sử dụng máy học cho e-mail được viết bằng tiếng Anh.

III. Methodology

3.1. Dataset

Đối với nghiên cứu này, dữ liệu thô được thu thập từ kaggle tài nguyên trực tuyến, dữ liệu này sẽ được sử dụng để huấn luyện các mô hình máy học. Dữ liệu ban đầu có sẵn bằng tiếng Anh và được lấy ở định dạng giá trị được phân tách bằng dấu phẩy (CSV). Cột đầu tiên, được gắn nhãn là 'loại' có hai giá trị có thể là thư rác hoặc ham, được dùng để phân loại email. Cột thứ hai được gắn nhãn 'Văn bản e-mail' và chứa nhiều nội dung e-mail.

	Unnamed: 0	label	text	label_num
0	605	ham	Subject: enron methanol ; meter # : 988291\r\n...	0
1	2349	ham	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	3624	ham	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	4685	spam	Subject: photoshop , windows , office . cheap ...	1
4	2030	ham	Subject: re : indian springs\r\nthis deal is t...	0

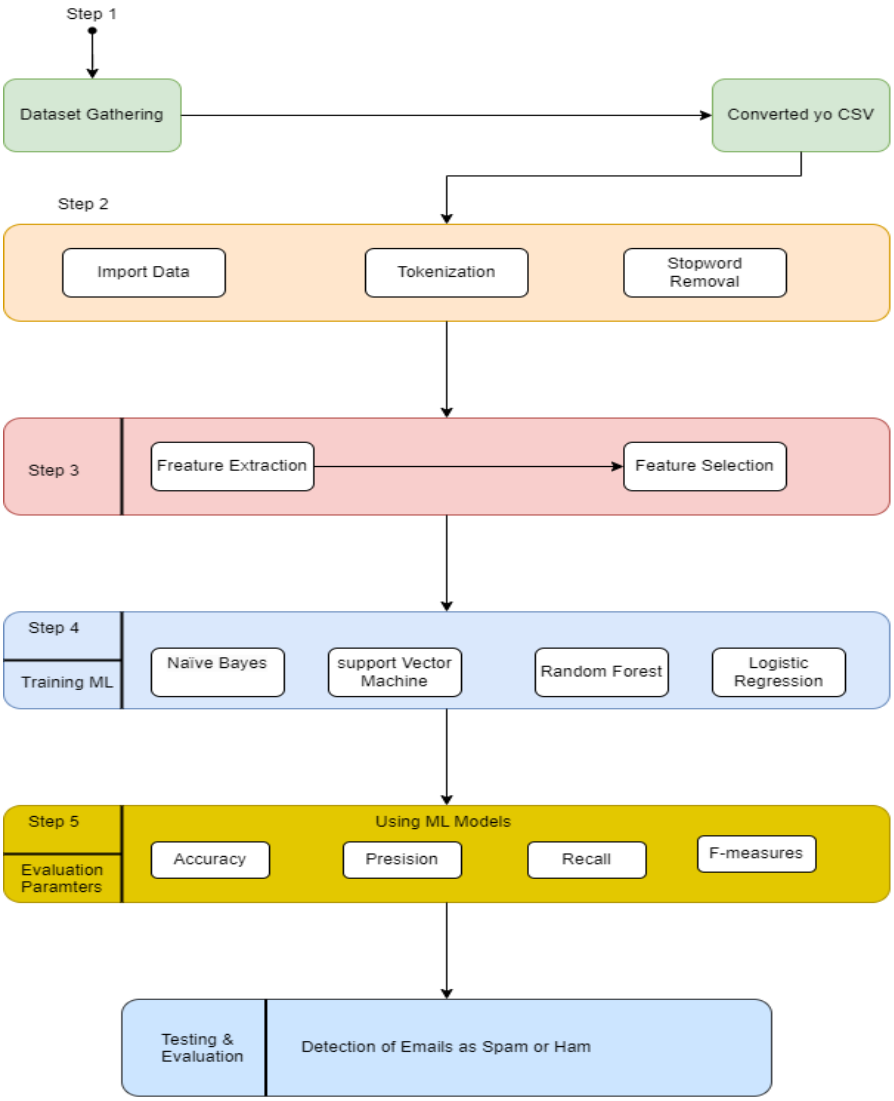
```
Columns Meaning

- **label**: This column represents whether the email is spam or not spam.
- **message**: This column contains the text of the emails.
```

3.2. Data Preprocessing

Trong học máy (ML), cụm từ tiền xử lý đề cập đến việc tổ chức và quản lý dữ liệu thô trước khi sử dụng nó để đào tạo và thử nghiệm các mô hình học tập khác nhau. Nói một cách đơn giản, tiền xử lý là một phương pháp khai thác dữ liệu ML để biến dữ liệu thô thành một cấu trúc có thể sử dụng được và tháo vát.

Đầu tiên trong quá trình xây dựng mô hình ML là tiền xử lý, trong đó dữ liệu từ thế giới thực, thường không đầy đủ, không chính xác và không chính xác do sai sót và thiếu sót, được biến thành một biến và xu hướng đầu vào chính xác, chính xác và có thể sử dụng được.

Phần phụ được đề cập dưới đây sẽ làm nổi bật từng bước liên quan đến giai đoạn tiền xử lý dữ liệu



3.2.1. Nhập dữ liệu

Giai đoạn đầu tiên là nhập tập dữ liệu được tải xuống từ 'Kaggle' và sau đó được chuyển đổi sang định dạng CSV. Bộ dữ liệu chứa 5171 email đã được phân loại là thư rác và ham.

3.2.2. Tokenization

Là một giai đoạn tiền xử lý quan trọng, trong bước này, tất cả các từ trong email được thu thập và số lần mỗi từ xuất hiện cũng như vị trí xuất hiện được tính. Với sự trợ giúp của Count Vectorizer, chúng em có thể tìm thấy sự lặp lại của các từ trong tập dữ liệu của mình. Mỗi từ được cấp một số duy nhất và do đó, chúng được gọi là mã thông báo, cũng mô tả số lần xuất hiện .

Python3

```
# Step 1: Load the input text
text = "The quick brown fox jumps over the lazy dog."

# Step 2: Define the tokenization rules (split on whitespace)
tokens = text.split()

# Step 4: Output the tokens
print(tokens)
```

Output:

```
['The', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy', 'dog.']
```

Mã thông báo bao gồm một trong những giá trị đặc trưng mà sau này sẽ giúp tạo ra các vectơ đặc trưng. Trong giai đoạn mã thông báo, mỗi từ được gán một mã thông báo duy nhất.

3.2.3. Stop word removal

Khi tập dữ liệu đã được chuyển đổi thành các mã thông báo duy nhất, bước tiếp theo là xóa mọi từ không cần thiết và không có ý nghĩa, ví dụ: khoảng trắng, dấu phẩy, dấu chấm, dấu hai chấm, dấu chấm phẩy và dấu chấm câu.

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

3.2.4. Lemmatization

Lemmatization (Lemmatization hay còn gọi là "hệ thống hoá từ điển") là quá trình nhóm các hình thức từ khác nhau của một từ lại với nhau để có thể phân tích như một đơn vị duy nhất. Lemmatization tương tự như stemming nhưng mang lại ngữ cảnh cho các từ. Điều này giúp liên kết các từ có ý nghĩa tương tự thành một từ duy nhất.

Trong quá trình tiền xử lý văn bản, cả stemming và lemmatization đều được sử dụng. Rất nhiều lần, người ta thấy hai thuật ngữ này gây nhầm lẫn. Một số người coi chúng là như nhau. Trên thực tế, lemmatization được ưu tiên hơn stemming vì nó thực hiện phân tích hình thái từ của các từ.

Các ứng dụng của lemmatization bao gồm:

1. Được sử dụng trong các hệ thống truy xuất toàn diện như các công cụ tìm kiếm.

2. Được sử dụng trong các hệ thống chỉ mục gọn nhẹ.

```
-> rocks : rock
-> corpora : corpus
-> better : good
```

3.2.5 Feature Extraction

Bag-of-Words (BoW) được sử dụng để biểu diễn văn bản và trích xuất đặc trưng trong các nhiệm vụ xử lý ngôn ngữ tự nhiên và truy xuất thông tin. Nó biểu diễn một văn bản dưới dạng tập hợp đa tập của các từ trong đó, bỏ qua ngữ pháp và thứ tự từ, nhưng giữ lại tần suất của từng từ. Biểu diễn này hữu ích cho các nhiệm vụ như phân loại văn bản, đo độ tương đồng giữa các tài liệu và phân cụm văn bản.

Bag-of-Words là một trong những phương pháp cơ bản nhất để chuyển đổi các từ thành một tập hợp các đặc trưng. Mô hình BoW được sử dụng trong phân loại văn bản, trong đó mỗi từ được sử dụng như một đặc trưng để huấn luyện bộ phân loại.

```
document = [ "One Geek helps Two Geeks", "Two Geeks help Four Geeks", "Each Geek helps many other Geeks at GeeksforGeeks."]
```

	at	each	four	geek	geeks	geeksforgeeks	help	helps	many	one	other	two
document[0]	0	0	0	1	1	0	0	1	0	1	0	1
document[1]	0	0	1	0	2	0	1	0	0	0	0	1
document[2]	1	1	0	1	1	1	0	1	1	0	1	0

Trong CountVectorizer, các từ này không được lưu trữ dưới dạng chuỗi. Thay vào đó, chúng được gán một giá trị chỉ mục cụ thể. Trong trường hợp này, từ 'at' có chỉ mục là 0, 'each' có chỉ mục là 1, 'four' có chỉ mục là 2 và cứ tiếp tục như vậy. Biểu diễn thực tế được hiển thị trong bảng dưới đây:

0	1	2	3	4	5	6	7	8	9	10	11
0	0	0	1	1	0	0	1	0	0	0	1
0	0	1	0	2	0	1	0	0	0	0	1
1	1	0	1	1	1	0	1	1	0	1	0

A Sparse Matrix

3.3. Trình diễn thuật toán từng bước

Bước 1 : Chọn một thư ngẫu nhiên từ bộ sưu tập cho mục đích thử nghiệm.

Bước 2 : Email được đề cập ở trạng thái chưa được xử lý. E-mail phải được xử lý sơ bộ trước khi quy trình phân loại và trích xuất tính năng có thể bắt đầu. Tokenization, and Stop Word Elimination là tất cả các bước trong quy trình tiền xử lý

i. Để bắt đầu, hãy chia e-mail thành các từ riêng biệt và mã hóa nó. Tokenization tách từng từ thành mã thông báo riêng của nó.
ii. Xóa tất cả các dấu chấm câu khỏi các ký tự bạn có được thông qua mã thông báo.
iii. Kiểm tra xem liệu có bất kỳ mã thông báo nào có sẵn trong văn bản nhập cơ sở hay không.

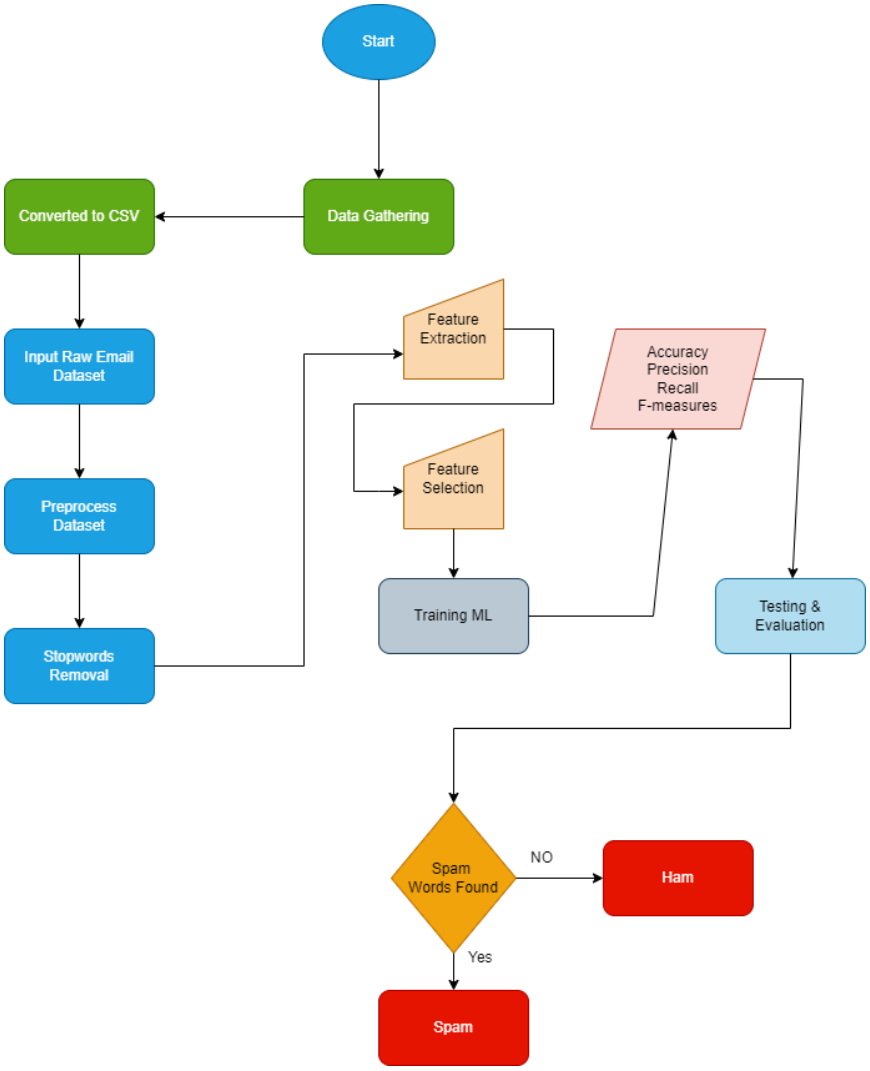
Bước 3 : Để sử dụng kỹ thuật trích xuất tính năng, hãy chọn các từ thuộc tính phù hợp từ bộ xác thực. Chỉ tập hợp các tính năng được kết nối gần nhất với danh mục được chọn.

Bước 4 : Sử dụng các tính năng được trích xuất và mã thông báo đã tạo để đào tạo ML. Mô hình đó có thể dễ dàng phân biệt giữa thư rác và thư rác.

Bước 5 : Tokens được phân loại là spam hoặc ham dựa trên tính năng tương tự của chúng khi các mô hình ML xác định.

Bước 6 : Cuối cùng, khả năng phân biệt spam hoặc ham trong một câu được đánh giá để phân loại cuối cùng.

Bước 7 : Đánh dấu e-mail là thư rác hoặc ham và tiếp tục với phần còn lại của email.



IV. Các mô hình học máy được sử dụng cho thử nghiệm

4.1. Naive Bayes

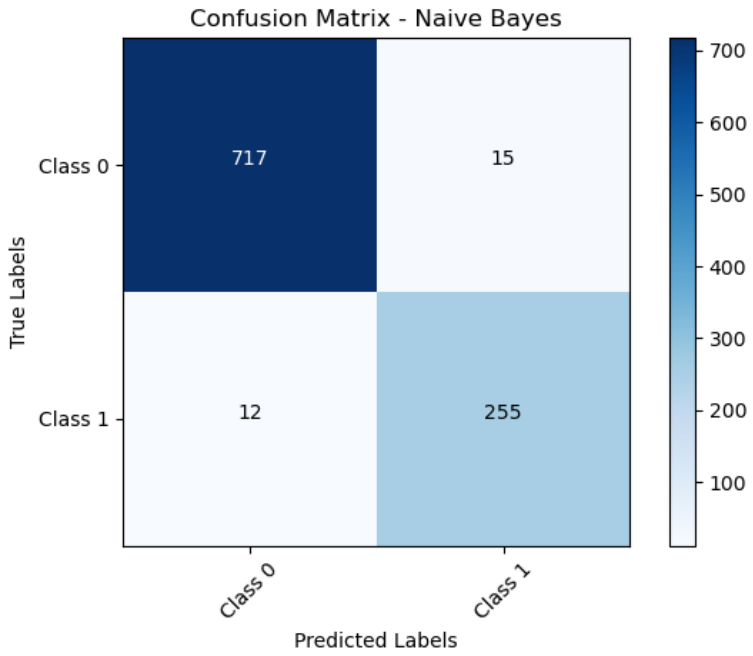
Từ năm 1998, Naive Bayes đã được sử dụng trong học máy có giám sát để xác định thư rác. Nó chủ yếu dựa vào cơ hội phân biệt giữa các thực thể khác nhau dựa trên các đặc điểm được xác định trước. Naive Bayes cảm nhận một từ hoặc sự kiện đã xảy ra trong ngữ cảnh trước đó và tính toán khả năng từ hoặc sự kiện đó xảy ra

lần nữa trong tương lai. Ví dụ: nếu một từ xuất hiện trong e-mail spam nhưng không xuất hiện trong e-mail ham, thuật toán rất có thể sẽ phân loại từ đó là thư rác.

$$P(c/x) = (P(x/c)\,P(c)) / (P(x)),$$
$$P(x) = \sum_{y} P(x/c)\,P(c).$$

Naive Bayes (Test Set):

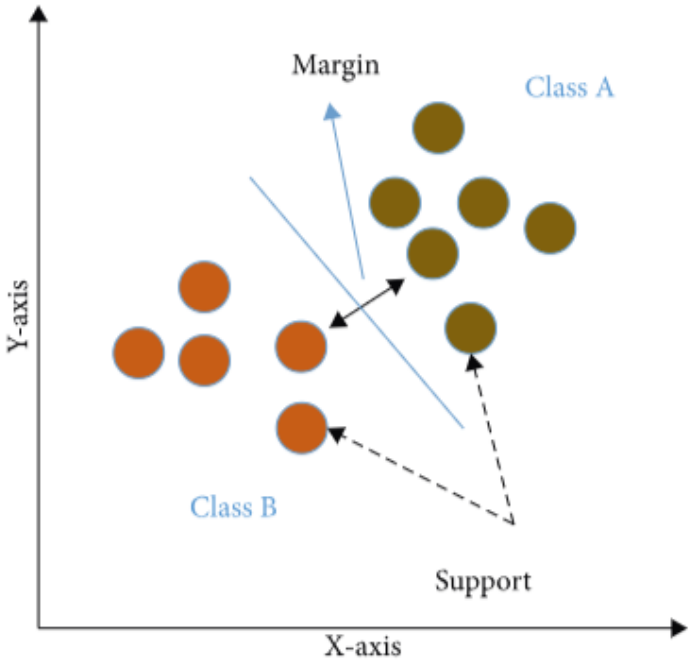
	precision	recall	f1-score	support
0	0.98	0.98	0.98	732
1	0.94	0.96	0.95	267
accuracy			0.97	999
macro avg	0.96	0.97	0.97	999
weighted avg	0.97	0.97	0.97	999



4.2. Support Vector Machine (SVM)

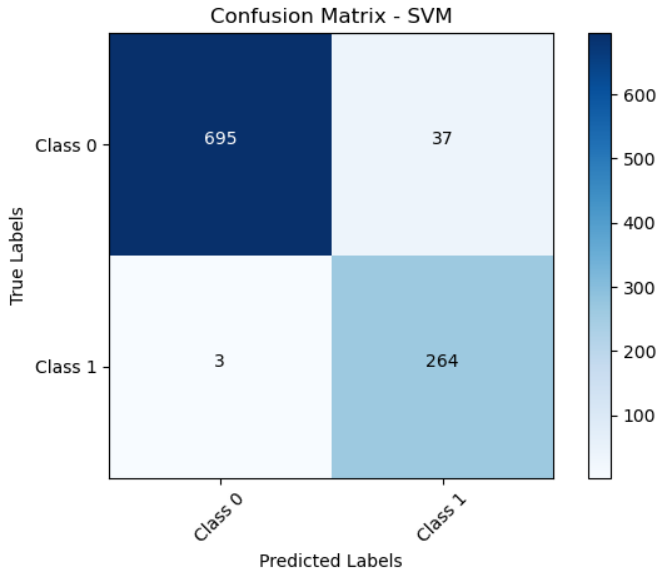
Support Vector Machine (SVM) là một thuật toán học máy được giám sát khác. Nó chỉ hoạt động cho các bộ dữ liệu đã được phân loại. Đối với mục đích đào tạo, SVM thường sử dụng cả bộ dữ liệu tích cực và tiêu cực. Các bộ dữ liệu phủ định không được sử dụng trong quá trình chuẩn bị cho bất kỳ mô hình máy học nào khác. SVM là mô hình phân loại và hồi quy được sử dụng phổ biến nhất. Đối với việc phân loại dữ liệu, nó đáng tin cậy hơn bất kỳ mô hình nào khác. SVM là mô hình phân loại nhanh nhất và đáng tin cậy nhất khi chúng ta chỉ có một lượng nhỏ dữ liệu được dán nhãn. Mô hình SVM sử dụng một siêu phẳng để phân tách các giá trị dương và âm

(thư rác và ham) khỏi tập dữ liệu. Sau đó, tìm ra các giá trị đủ gần với bề mặt quyết định



SVM (Test Set):

	precision	recall	f1-score	support
0	1.00	0.95	0.97	732
1	0.88	0.99	0.93	267
accuracy			0.96	999
macro avg	0.94	0.97	0.95	999
weighted avg	0.96	0.96	0.96	999

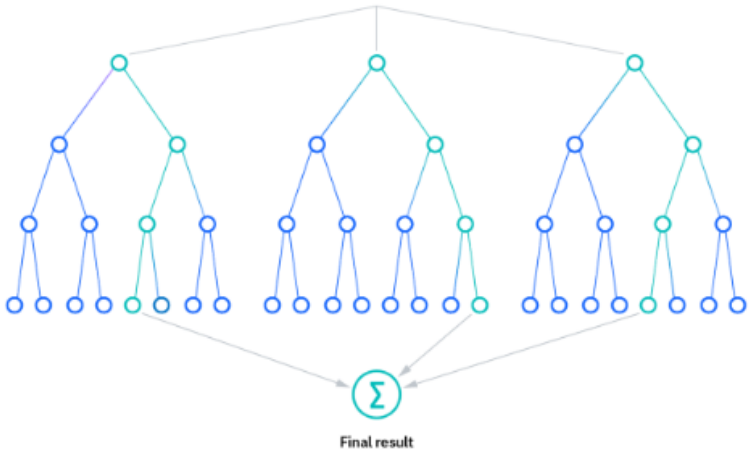


4.3. Random Forest

Random Forest (Rừng ngẫu nhiên) là một thuật toán học máy được sử dụng phổ biến trong các bài toán phân loại và dự đoán. Nó là một mô hình dựa trên Ensemble Learning (học tập từ tập hợp), tức là kết hợp các mô hình nhỏ để tạo ra một mô hình lớn và tốt hơn.

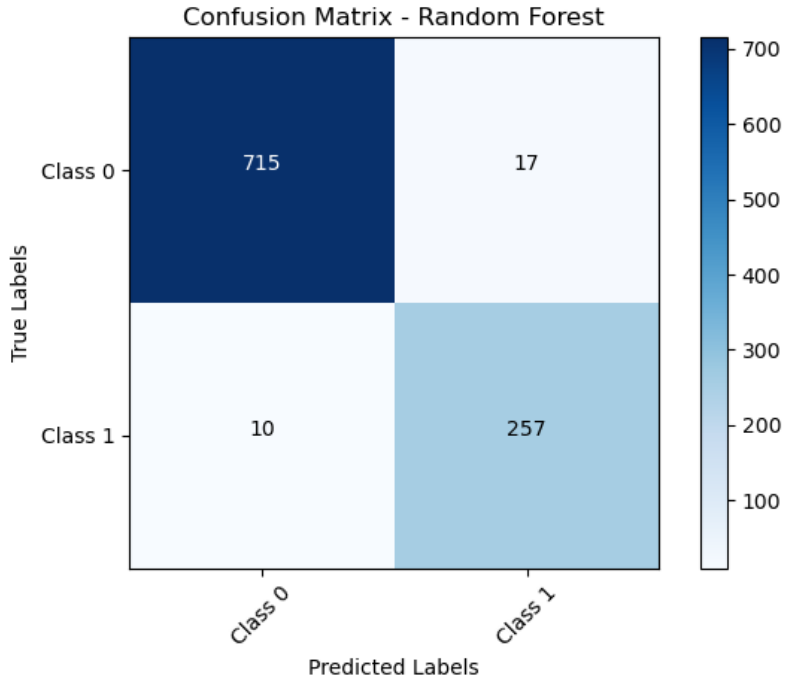
Thuật toán Random Forest sử dụng nhiều cây quyết định (decision trees) trong quá trình huấn luyện. Mỗi cây quyết định được xây dựng bằng cách chọn ngẫu nhiên một tập hợp con của các đặc trưng và các điểm dữ liệu trong tập huấn luyện để xây dựng một cây quyết định. Quá trình này được thực hiện nhiều lần để tạo ra nhiều cây quyết định khác nhau.

Khi cần phân loại hoặc dự đoán một điểm dữ liệu mới, thuật toán Random Forest sẽ truyền điểm dữ liệu đó qua tất cả các cây quyết định và tính toán đầu ra của mỗi cây. Kết quả cuối cùng được tính bằng cách lấy trung bình hoặc đa số phiếu bầu của các đầu ra của các cây. Kết quả này sẽ được coi là kết quả dự đoán cuối cùng.



Random Forest (Test Set):

	precision	recall	f1-score	support
0	0.99	0.98	0.98	732
1	0.94	0.96	0.95	267
accuracy			0.97	999
macro avg	0.96	0.97	0.97	999
weighted avg	0.97	0.97	0.97	999



Các ưu điểm của thuật toán Random Forest bao gồm:

- Khả năng xử lý dữ liệu có tính tương quan cao và nhiễu.
- Khả năng xác định độ quan trọng của các đặc trưng trong mô hình.
- Khả năng tạo ra các dự đoán chính xác và phổ biến trong các bài toán phân loại và dự đoán.

Tuy nhiên, các nhược điểm của thuật toán Random Forest bao gồm:

- Chi phí tính toán cao khi số lượng cây quyết định lớn.
- Khó khăn trong việc giải thích các quyết định của mô hình.

4.4. Logistic regression

Logistic Regression là một thuật toán học máy dùng để phân loại các điểm dữ liệu vào hai hoặc nhiều nhóm khác nhau. Nó được sử dụng rộng rãi trong các bài toán phân loại như xác định một email là spam hay không, hoặc xác định một bệnh nhân có mắc bệnh hay không dựa trên các đặc trưng y tế.

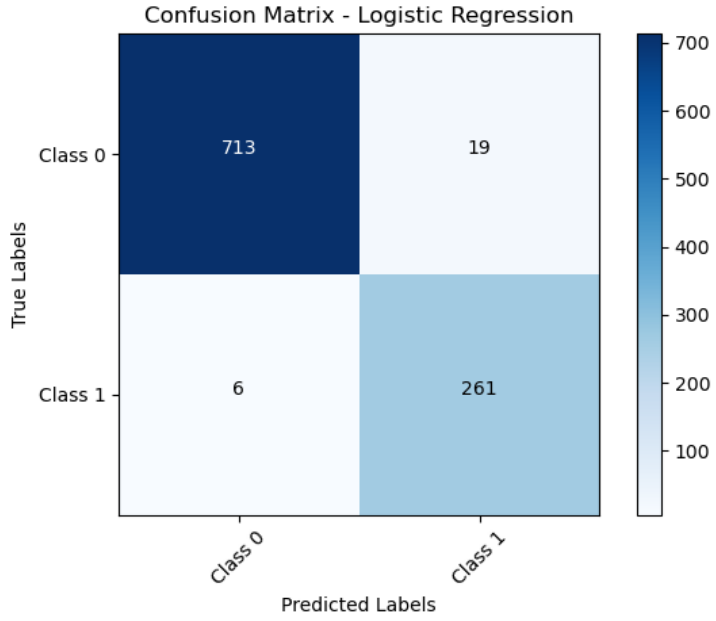
Thuật toán Logistic Regression dựa trên hàm sigmoid, một hàm số phi tuyến có giá trị giới hạn từ 0 đến 1. Hàm sigmoid được sử dụng để ước tính xác suất của một điểm dữ liệu thuộc về một nhóm cụ thể nào đó.

Cụ thể, trong Logistic Regression, các đặc trưng của dữ liệu được sử dụng để tính toán một giá trị đầu ra, được đưa qua hàm sigmoid để tạo ra một giá trị dự đoán nằm trong khoảng từ 0 đến 1. Giá trị này được coi là xác suất của điểm dữ liệu đó thuộc về một lớp cụ thể. Nếu giá trị dự đoán lớn hơn một ngưỡng (threshold) được xác định trước, điểm dữ liệu được phân loại vào lớp đó, ngược lại, nó sẽ được phân loại vào lớp còn lại.

Quá trình huấn luyện mô hình Logistic Regression bao gồm tối ưu hoá các tham số của hàm sigmoid thông qua thuật toán Gradient Descent, để tìm ra các tham số phù hợp nhất với dữ liệu huấn luyện. Các tham số này được sử dụng để dự đoán lớp của các điểm dữ liệu mới.

Logistic Regression (Test Set):

	precision	recall	f1-score	support
0	0.99	0.97	0.98	732
1	0.93	0.98	0.95	267
accuracy			0.97	999
macro avg	0.96	0.98	0.97	999
weighted avg	0.98	0.97	0.98	999



V. Các tham số đánh giá học máy

Độ chính xác (Accuracy), độ chính xác nhớ (Precision), độ chính xác gọi lại (Recall) và F-measure được sử dụng để kiểm tra hiệu suất của thuật toán đề xuất. Các giá trị True Positive (TP), False Positive (FP), True Negative (TN) và False Negative (FN) có thể được sử dụng để tính toán các tham số này.

Các tham số sau được tính toán bằng cách sử dụng mô hình Naive Bayes , SVM , Logisstic Regression , Random Forest .

5.1. Độ chính xác (Accuracy)

Đây là tỷ lệ phần trăm dữ liệu từ toàn bộ tập dữ liệu được dự đoán đúng. Độ chính xác mô tả độ chính xác tổng thể của việc dự đoán của một bộ phân loại.

$$\text{Accuracy} = \frac{TP + TN}{\text{total sample}}$$

5.2. Độ chính xác nhớ (Precision)

Đây là chỉ số đo lường hiệu suất của một bộ phân loại. Nó đại diện cho tổng số giá trị thực sự đúng được phân loại bởi một bộ phân loại.

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positive}}{\textbf{True Positive+False Positive}}$$

5.3. Độ chính xác gọi lại (Recall)

Đây là đo lường độ chính xác của dự đoán của một bộ phân loại.

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positive}}{\textbf{True Positive+False Negative}}$$

5.4. Độ đo F-Measure

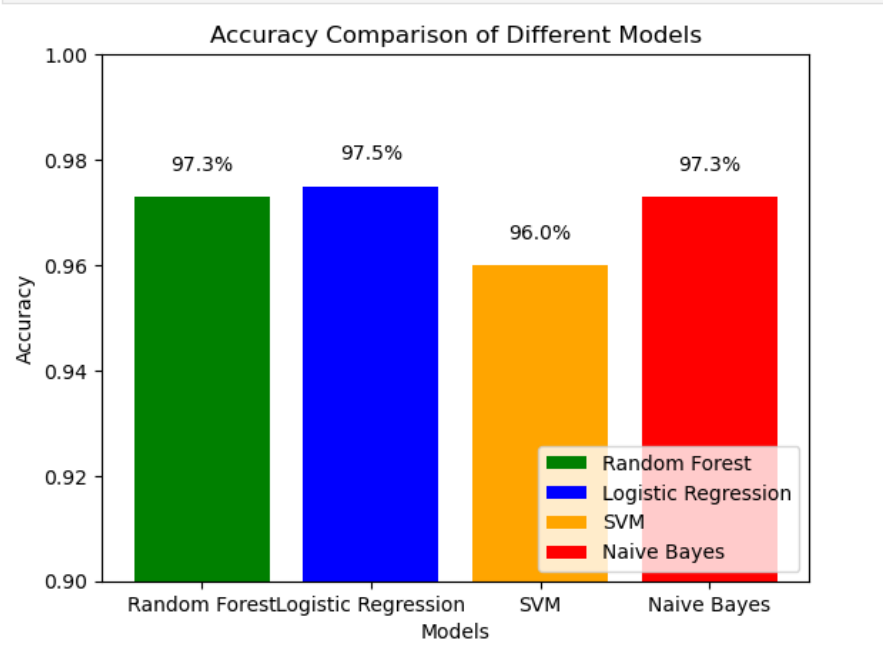
Đây là chỉ số cho thấy độ chính xác của dự đoán của bộ phân loại.

$$\boldsymbol{F1 - score} = \frac{\boldsymbol{2*Precision*Recall}}{\textbf{Precision+Recall}}$$

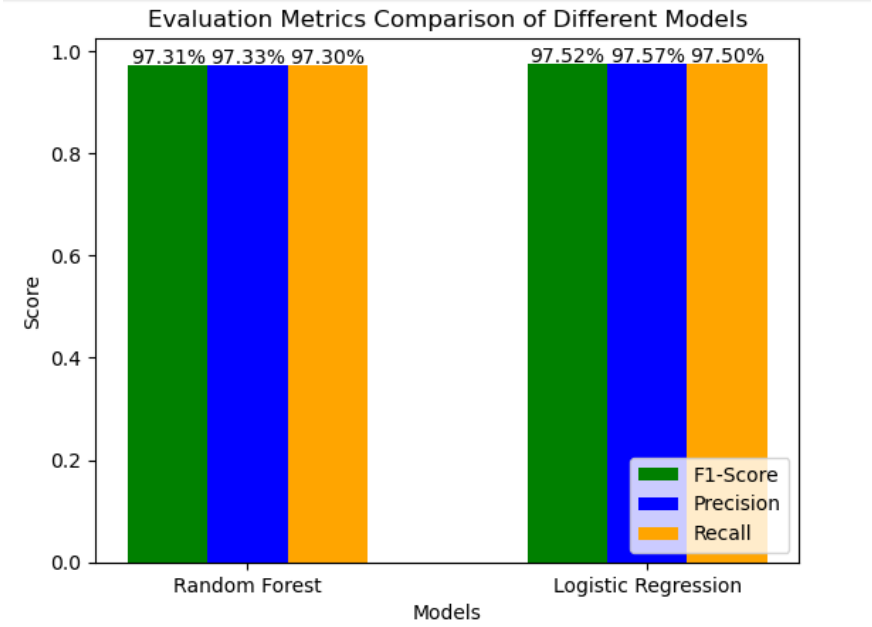
VI. Kết quả và Phân tích

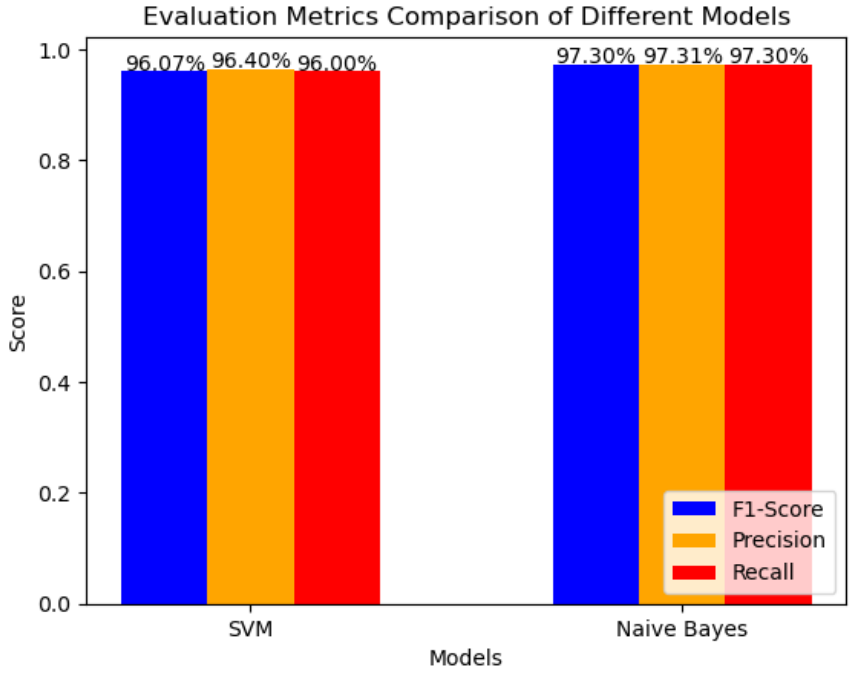
Kết quả và so sánh giữa các thuật toán khác nhau sau quá trình huấn luyện và kiểm tra dữ liệu được trình bày trong phần trên. Chúng em đã thu thập 5170 email

từ nguồn tài nguyên trực tuyến 'kaggle' . 3994 email được sử dụng để huấn luyện các mô hình ML khác nhau. 999 email được sử dụng cho việc kiểm tra để đo lường độ chính xác và các độ đo đánh giá. Như đã được giải thích về các độ đo đánh giá trong phần trênt chúng em đã đánh giá Độ chính xác (Accuracy) , Độ chính xác nhớ (Precision), Độ chính xác gọi lại (Recall) và độ đo F-measure được đo bằng cách sử dụng SVM , Naive Bayes , Logisstic Regression và Random Forest . Cuối cùng bằng cách sử dụng các biểu đồ khác nhau, so sánh các mô hình được trình bày dưới đây. Các kết quả trong Bảng dưới cho thấy thuật toán học máy (LOGISTIC) là một phương pháp mạnh hơn để phát hiện các email rác tiếng Anh với độ chính xác cao đạt 97,5%.



Trong Bảng này đã đề cập, chúng em đã so sánh độ chính xác của bốn mô hình ML khác nhau. Chúng em có thể thấy rằng mô hình ML (LOGISTIC) là mô hình có độ chính xác cao nhất trong số tất cả các mô hình. Các mô hình ML như SVM ,Naive Bayes và , Random Forest có độ chính xác tương đương nhau và thấp hơn so với Logisstic Regression.





các model như (Random Forest , Logistic Regression) , theo bar chart trên chúng em đã nhận thấy rằng F-score của Logistic Regrssion tốt hơn so với Random Forest , giá trị precision cũng gần bằng nhau . còn các model như (SVM , Naive Bayes) , giá trị Precision , Recall , F-score của Naive Bayes đều tốt hơn so với SVM ,nhung nó lại không tốt mấy do với Logistic Regression.

VII. Kết luận

Với sự gia tăng trong việc sử dụng email, nghiên cứu này tập trung vào việc sử dụng các phương pháp tự động để phát hiện email rác được viết bằng tiếng Anh. Nghiên cứu sử dụng các thuật toán học máy khác nhau để phát hiện . Trong nghiên cứu, một tập dữ liệu email đã được dịch sang phân loại, bao gồm email rác và email hợp lệ, được tạo ra từ Kaggle và được tiền xử lý với các phương pháp khác nhau. Accuracy, precision, recall, F-measure , được sử dụng làm các chỉ số so sánh để đánh giá hiệu suất. Nghiên cứu kết luận rằng các mô hình học máy (LOGISSTIC) hiệu quả hơn trong việc phân loại email rác tiếng Anh. So sánh, thuật toán LOGISSTIC có tỷ lệ chính xác cao khoảng 97.5% và các gia trị Precision , Recall , F-score cũng tốt hơn với 3 model khác nữa mà chúng em lấy làm model train.

References

https://www.hindawi.com/journals/sp/2021/6508784/#acknowledgments

https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv/code

https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

https://www.geeksforgeeks.org/