ĐẠI HỌC QUỐC GIA HÀ NỘI

Trường Đại học Công nghệ



**BÀI TIỂU LUẬN**

**MÔN:** Xử lý ngôn ngữ tự nhiên

**Đề tài:** Áp dụng Học máy và Xử lý ngôn ngữ tự nhiên trong phát hiện email lừa đảo

Thành viên : Nguyễn Nam Khánh 21020267

Nguyễn Ngọc Minh 21021519

Nhóm : 18

Lớp : QH-2021-I/CQ-C-A-CLC3

Giảng viên : Đặng Trần Bình.

# 

# Mở đầu

## Đặt vấn đề

Sự phát triển nhanh chóng của các dịch vụ trực tuyến và internet là một cơ hội lớn cho tất cả mọi người có thể tiếp cận thông tin từ mọi nơi trên thế giới. Tuy nhiên, đáng tiếc là, điều này cũng làm gia tăng các cuộc tấn công mạng, mà trong đó, lừa đảo qua mạng nói chung, và lừa đảo qua email là một trong những hình thức phổ biến nhất.

Sự thành công của email lừa đảo dựa trên việc thao túng cảm xúc của con người, dẫn đến sự bất cẩn trong những hành động của mình. Nó khai thác các lỗ hổng được tìm thấy trong quy trình từ phía người dùng cuối. Việc chống lại nó phụ thuộc vào kiến thức và nhận thức của mỗi người. Ngay cả khi hệ thống đủ an toàn chống lại tội phạm mạng, người dùng thiếu hiểu biết vẫn có thể bị lừa tiết lộ dữ liệu cá nhân. Tệ hơn nữa, nó có thể tạo điều kiện cho chúng truy cập vào dữ liệu của công ty, tổ chức. Tổn thất hằng năm do tội phạm mạng trên toàn cầu do McAfee ước tính lên tới hàng nghìn tỉ đô la Mĩ ([McAfee, 2020](https://www.mcafee.com/enterprise/en-us/assets/reports/rp-hidden-costs-of-cybercrime.pdf?5c22b0df_page=11)). Vào năm 2020, 75% số tổ chức được hỏi tới xác nhận là đã trải qua ít nhất một cuộc tấn công lừa đảo, và 95% trong số đó là đến từ email ([Phishing Statistics Report, 2020](https://www.tessian.com/blog/phishing-statistics-2020/)).

Trong tài liệu này, chúng tôi nghiên cứu một mô hình phát hiện email lừa đảo áp dụng Học máy và Xử lý ngôn ngữ tự nhiên. Chúng tôi sử dụng nội dung của email làm đặc trưng chính để phân tích. Nội dung email sẽ được xử lý bằng các kỹ thuật NLP, sau đó, áp dụng mô hình RNN để phân loại vào 1 trong 2 lớp là “Safe email” và “Phishing email”.

## Đối tượng và mục tiêu nghiên cứu

Mô hình này chỉ được huấn luyện với các email bằng tiếng Anh, không bao gồm các trang web hoặc email lừa đảo bằng các ngôn ngữ khác. Mô hình này cũng không xét đến các đặc trưng khác của email như tệp đính kèm hoặc URL. Tuy nhiên, nó có thể ứng dụng trong việc phát hiện lừa đảo thông qua phân tích văn bản từ các phương tiện lừa đảo khác nhau, chẳng hạn như tin nhắn, bài đăng hoặc tweet trên mạng xã hội.

Phần còn lại của tài liệu này sẽ mô tả chi tiết các kỹ thuật xử lý dữ liệu (Phần 2), cấu trúc chi tiết mô hình RNN (Phần 3), các tinh chỉnh khi huấn luyện mô hình (Phần 4), và đánh giá hiệu quả của mô hình với dữ liệu (Phần 5).

# Nội dung

## Tổng quan

### Email lừa đảo là gì?

Email lừa đảo (Phishing email) là một hình thức tấn công trực tuyến mà kẻ tấn công sử dụng email để lừa dối người nhận với mục tiêu chiếm đoạt thông tin cá nhân, tài sản hoặc gây hại trong nhiều hình thức khác. Bản chất của thuật ngữ “Phishing” là biến thể của từ “Fishing” trong tiếng Anh có nghĩa là “câu cá”. Thuật ngữ này được sử dụng lần đầu tiên vào năm 1987 và sau đó lan rộng ra toàn cầu.

### Các loại email lừa đảo

#### Email lừa đảo thông tin cá nhân:

Loại email này thường hướng đến lấy thông tin người dùng cuối, bao gồm họ tên, địa chỉ, số tài khoản ngân hàng, căn cước công dân, số điện thoại, mật khẩu mạng xã hội, và kể cả định vị vị trí của người dùng. Các hình thức của loại này:

* Email mang thông tin giật gân, cùng với đó là đính kèm những đường link dẫn đến những trang web lạ, yêu cầu nhập thông tin để tiếp tục.
* Email xác nhận mua hàng
* Email phần quà, phần thưởng, yêu cầu nhập thông tin cá nhân để nhận quà

#### Email lừa đảo tiền

Loại email này thường dùng để lấy tiền của người dùng cuối. Các hình thức của loại này là:

* Email giao hàng, yêu cầu chuyển tiền để nhận hàng
* Email giả mạo là người thân để vay tiền
* Email giả mạo chuyển nhầm tiền

### Phát hiện email lừa đảo

#### Dấu hiệu của email lừa đảo

Các dầu hiệu phổ biến của một email lừa đảo là:

* Địa chỉ gửi email không đáng tin cậy, gần giống với tên của một tổ chức quen thuộc
* Yêu cầu thông tin cá nhân nhạy cảm
* Liên kết hoặc các tệp đính kèm đáng ngờ
* Lỗi chính tả và ngữ pháp

#### Phương pháp phát hiện email lừa đảo

Có 2 cách thông thường để phát hiện email lừa đảo

* Danh sách đen: Phương pháp này so sánh địa chỉ email, địa chỉ IP hoặc địa chỉ DNS của người gửi với danh sách các địa chỉ lừa đảo được xác định trước và nếu dữ liệu khớp, email sẽ bị từ chối trước khi đến máy chủ SMTP
* Học máy: phát hiện email lừa đảo bằng tập liệu được huấn luyện

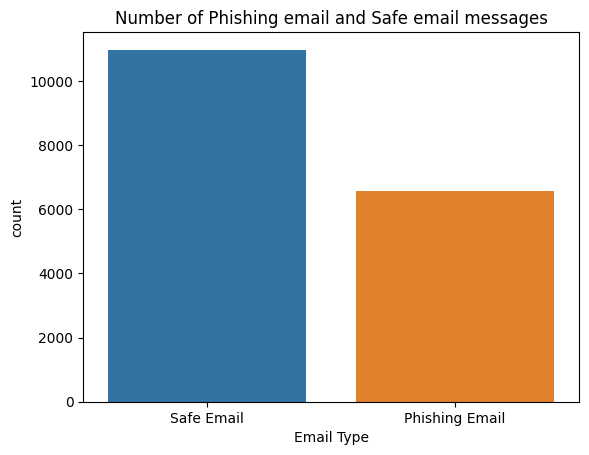
Báo cáo này tập trung nghiên cứu phương pháp học máy, nghiên cứu tập dữ liệu của các email lừa đảo để huấn luyện mô hình

## Xử lý dữ liệu

### Lựa chọn dữ liệu

Nhóm chúng tôi sử dụng bộ dữ liệu gán nhãn công khai trên Kaggle: <https://www.kaggle.com/datasets/subhajournal/phishingemails>

Bộ dữ liệu trong tệp csv được tổ chức dưới dạng bảng với 17538 cặp text - label. Các nhãn "Safe email" và "Phishing email" được đánh số y=1 và y=0 tương ứng. Phân phối của các nhãn được cho bằng biểu đồ như sau:



### Phân chia dữ liệu

Dữ liệu được phân chia thành 3 nhóm train\_set, val\_set, test\_set với tỉ lệ tương ứng 0.6, 0.2, 0.2.

### Vector hoá dữ liệu

Gồm 3 bước như sau;

* Bước 1: Xây dựng một từ điển với max\_words = 105 từ phổ biến nhất trong train\_set. Thứ tự các từ sẽ được sắp xếp theo số lần xuất hiện trong train\_set
* Bước 2: Mỗi từ trong train\_set sẽ được gán bằng 1 số tự nhiên là vị trí của từ đó trong từ điển. Bây giờ, mỗi văn bản trong tập dữ liệu sẽ biểu diễn bằng 1 vector từ có độ dài tương ứng.
* Bước 3: Chèn thêm 0 hoặc cắt bớt kích thước để các vector còn lại có độ dài max\_len = 150. Sử dụng các vector có kích thước giống nhau giúp xử lý batch dễ dàng hơn.

## Xây dựng mô hình

Cấu trúc RNN gồm 3 lớp được lắp ghép tuần tự:

* Một lớp Embedding: Lớp Embedding có nhiệm vụ biến mỗi từ thành một vector embedding. Vector embedding tập trung nhiều vào ngữ nghĩa của từ, 2 từ có nghĩa tương đồng nhau sẽ tạo ra 2 embedding tương đồng (khi được biểu diễn trong cùng 1 không gian thì các điểm này nằm gần nhau).
* Một lớp LSTM (Long Short-Term Memory): LSTM là một lớp phổ biến trong các mạng RNN xử lý ngôn ngữ tự nhiên. LSTM cùng với GRU (Gated Recurrent Unit) khắc phục tình trạng Gradient triệt tiêu ở mạng neuron sâu, từ đó giải quyết vấn đề phụ thuộc ngữ nghĩa giữa các từ cách nhau xa trong 1 câu.
* Một lớp Dense: Các vector embedding sau khi đi qua lớp LSTM sẽ đi vào lớp Dense 1 unit duy nhất với hàm kích hoạt sigmoid để tính xác suất P(Safe email | ) và P(Phishing email | ).

## Huấn luyện mô hình và tinh chỉnh tham số

Mô hình được huấn luyện với train\_set, tinh chỉnh các tham số bằng val\_set.

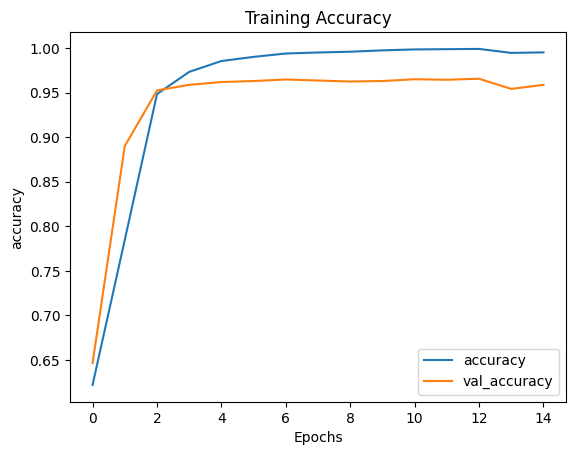
Chúng tôi định nghĩa thêm hàm Loss và hàm Optimizer phục vụ việc huấn luyện mô hình:

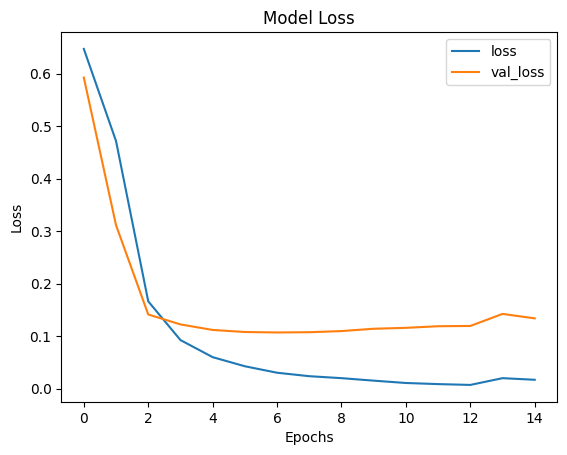
* Hàm optimizer Adam (Adaptive Moment Estimation): Thuật toán tối ưu sử dụng trong quá trình huấn luyện. Nó tự động cập nhật learning\_rate phù hợp với gradient, từ đó tăng tốc độ hội tụ của mô hình.
* Hàm loss Binary Cross Entropy: hàm tính toán sai số giữa đầu ra dự đoán với nhãn thực tế. Công thức của hàm loss:

Với t là nhãn thực tế, y là đầu ra dự đoán.

Thông qua thực nghiệm, chúng tôi nhận thấy mô hình được huấn luyện với learning rate = 10-3, và số epochs là 15 đem lại hiệu suất cao nhất.

* Learning rate bé làm chậm quá trình hội tụ, ngược lại learning rate cao hơn khiến gradient không thể hội tụ.
* Mô hình dần hội tụ với các epoch < 15. Khi epoch ≥ 15, mô hình có dấu hiệu overfitting.





## Đánh giá mô hình

Sau khi huấn luyện nhiều lần và tinh chỉnh tham số, chúng tôi sử dụng test\_set để đánh giá hiệu quả mô hình dựa trên các chỉ số Accuracy, Precision, Recall, F1-score. Kết quả thu được là:

* Accuracy: 96.18%
* Precision: 0.9767
* Recall: 0.9615
* F1-score: 0.9690