THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): https://youtu.be/fsxHMHfPZ6I
- Link slides (dạng .pdf đặt trên Github của nhóm):
- Link Poster

• Họ và Tên: Nguyễn Duy Phương • MSSV: 18521276 	• Lớp: CS2205.APR2023 • Tự đánh giá (điểm tổng kết môn): 8.5/10 • Số buổi vắng: 1 • Link Github:

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

NÂNG CAO HIỆU SUẤT PHÁT HIỆN THƯ RÁC VỚI BERT: XỬ LÝ NGÔN NGỮ TỰ NHIÊN

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

IMPROVING SPAM DETECTION PERFORMANCE WITH BERT: A NATURAL LANGUAGE PROCESSING APPROACH

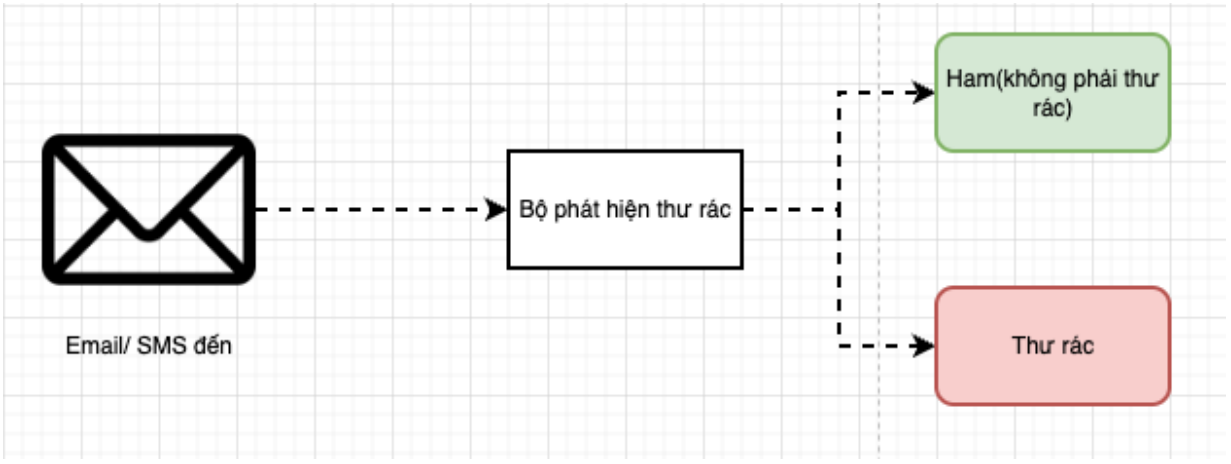
TÓM TẮT *(Tối đa 400 từ)*

Emails và tin nhắn SMS là công cụ phổ biến nhất trong giao tiếp hiện nay, và với sự gia tăng người dùng email và SMS, số lượng thư rác cũng tăng lên. Thư rác là bất kỳ loại giao tiếp kỹ thuật số không mong muốn nào được gửi đi hàng loạt, email và tin nhắn SMS rác gây lãng phí tài nguyên. Mặc dù hầu hết thư rác bắt nguồn từ các nhà quảng cáo muốn giới thiệu sản phẩm của mình, nhưng ngoài ra còn có một số khác chứa thông tin độc hại với ý đồ như: các email dùng để lừa đảo nạn nhân, cung cấp thông tin nhạy cảm như tên đăng nhập trang web hoặc thông tin thẻ tín dụng,.... Trong nghiên cứu này, chúng tôi xây dựng một bộ phát hiện thư rác bằng cách sử dụng mô hình đã được huấn luyện trước BERT, phân loại email và tin nhắn dựa trên việc hiểu văn bản của chúng. Chúng tôi huấn luyện mô hình phát hiện thư rác của mình bằng cách sử dụng nhiều nguồn tài liệu như tập dữ liệu SMS: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset, tập dữ liệu Enron: https://www.kaggle.com/datasets/wcukierski/enron-email-dataset, tập dữ liệu SpamAssassin: https://www.kaggle.com/datasets/cesaber/spam-email-data-spamassassin-2002 và tập dữ liệu Ling-Spam:https://www.kaggle.com/datasets/mandygu/lingspam-dataset.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Email là công cụ quan trọng nhất trong giao tiếp và được sử dụng rộng rãi trong hầu hết các lĩnh vực như kinh doanh, doanh nghiệp và ngay cả cho người dùng cá nhân.

Để giao tiếp một cách hiệu quả, việc phát hiện thư rác là một trong những tính năng quan trọng nhằm nâng cao trải nghiệm người dùng và bảo mật. Thư rác là một loại thư điện tử được gửi cho một số lượng lớn người dùng cùng một lúc, thường chứa những tin nhắn khó hiểu, lừa đảo hoặc nguy hiểm nhất là nội dung lừa đảo thông tin cá nhân. Bộ phát hiện thư rác là một chương trình được sử dụng để phát hiện thư điện tử không được yêu cầu, không mong muốn và nhiễm virus và ngăn những tin nhắn đó đến hộp thư đến của người dùng.



Bộ phát hiện thư rác sử dụng các phương pháp khác nhau để phân loại thư điện tử. Một phương pháp là lọc thư dựa trên địa chỉ người gửi được gọi là bộ lọc danh sách chặn, trong đó có một danh sách email của những kẻ gửi thư rác và danh sách này được cập nhật thường xuyên để bắt kịp với những kẻ gửi thư rác khi thay đổi địa chỉ email của họ, tuy nhiên khi kẻ gửi thư rác thay đổi tên miền email của mình, phương pháp đó sẽ không hoạt động cho đến khi email của kẻ gửi thư rác mới được liệt kê trong danh sách chặn, do đó chúng ta không thể chỉ tin vào phương pháp này. Một phương pháp khác cho bộ phát hiện thư rác là lọc theo nội dung email, trong đó bộ phát hiện thư rác kiểm tra nội dung của từng email và phân loại nó, và loại bộ phát hiện thư rác này có hiệu suất rất cao vì nội dung của thư rác thường dễ dự đoán, đề xuất giao dịch, quảng cáo nội dung rõ ràng hoặc nhắm vào những cảm xúc cơ bản của con người, như mong muốn và sợ hãi. Và một phương pháp khác là có một bộ lọc email dựa trên quy tắc, cho phép người dùng cấu hình thủ tục đặc biệt cho tất cả các email đến và phân loại chúng là thư rác hoặc không phải là thư rác, ví dụ về các quy

tắc này là phân loại email từ người gửi cụ thể của chủ đề email chứa một cụm từ cụ thể hoặc thậm chí là nội dung của thông điệp chứa một số từ và mỗi khi một email đến khớp với một trong các quy tắc, nó tự động chuyển tiếp email đó vào thư mục thư rác.

Trong nghiên cứu này, chúng tôi sử dụng mô hình đã được huấn luyện trước NLP "BERT" để xây dựng một bộ phát hiện thư rác hiệu suất cao bằng cách huấn luyện mô hình trên các tập dữ liệu khác nhau. BERT [1] là một mô hình đã được huấn luyện trước máy học được sử dụng cho nhiệm vụ Hiểu Ngôn ngữ Tự nhiên để giúp máy móc hiểu ngữ cảnh của các câu được phát triển bởi Google AI vào năm 2018.

Input:

- Email và SMS chưa được phân loại.

Output:

- Email và SMS đã được phân loại.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

- Sử dụng BERT để phát hiện thư rác hiệu quả: Tận dụng sức mạnh của BERT, để xây dựng một hệ thống phát hiện thư rác hiệu quả và chính xác.

- Cải thiện hiệu suất phân loại hiệu quả: Mục tiêu thứ hai là nâng cao hiệu suất phân loại của trình phát thư rác. Bằng cách sử dụng BERT context và tinh chỉnh mô hình bằng tác vụ phân loại nhị phân, mục đích là để đạt được độ chính xác và độ chính xác cao trong việc phân biệt giữa thư rác và email (ham). Mục đích là để giảm thiểu thông tin xác thực sai và phủ định sai, từ đó tối ưu hóa hiệu suất tổng thể của hệ thống phát hiện thư rác.

- Tăng cường mã hóa và đại diện đầu vào: Mục tiêu thứ ba là cải thiện quy trình mã hóa và đại diện đầu vào cho BERT.

NỘI DUNG VÀ PHƯƠNG PHÁP

- Tiền xử lý - Dữ liệu email từ SpamAssassin, SMS Spam Collection v.1, Enron corpus và Ling-Spam corpus được làm sạch bằng cách xóa URL, ký tự không mong muốn và thẻ HTML.

- Tạo tập dữ liệu đào tạo - Tập hợp dữ liệu được xử lý trước được chia thành tập dữ liệu đào tạo và đánh giá. Tập dữ liệu đào tạo được sử dụng để đào tạo mô hình phát hiện thư rác.

- Đào tạo mô hình - Mô hình cơ sở BERT được đào tạo bằng PyTorch. Email và nhãn tương ứng được cung cấp cho mô hình theo lô 16.

- Đánh giá - Bộ dữ liệu đánh giá được sử dụng để đánh giá hiệu suất của mô hình được đào tạo.

KẾT QUẢ MONG ĐỢI

- *Sử dụng BERT để phát hiện thư rác hiệu quả: Qua việc sử dụng BERT và tinh chỉnh mô hình, chúng ta đã xây dựng thành công một hệ thống phát hiện thư rác mạnh mẽ và chính xác. BERT giúp chúng ta hiểu ngữ cảnh(context) và nắm bắt được các đặc trưng quan trọng để phân loại email là thư rác hay không.*

- *Cải thiện hiệu suất phân loại hiệu quả: Kết quả đạt được là tăng cường hiệu suất phân loại của hệ thống phát hiện thư rác. Bằng cách sử dụng BERT và tinh chỉnh mô hình, chúng ta đã đạt được độ chính xác và độ chính xác cao trong việc phân biệt giữa thư rác và email hợp lệ (ham).*

- *Tăng cường mã hóa và đại diện đầu vào: Chúng ta đã cải thiện quy trình mã hóa và đại diện đầu vào cho BERT. Việc sử dụng tokenizer BERT và thực hiện quy trình mã hóa và đại diện cho email giúp chúng ta biểu diễn hiệu quả nội dung email và nắm bắt thông tin quan trọng.*