THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://www.youtube.com/watch?v=BhY8QaaLWhY*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/NTD1810/CS2205.NOV2024/blob/main/Slide_Ung%20dung%20Transfer%20Learning%20da%20ngon%20ngu%20trong%20phat%20hienn%20Spam%20Email%20Tieng%20viet.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Nguyễn Thuỳ Dương
- MSSV: 240202006

- Lớp: CS2205.NOV2024
- Tự đánh giá (điểm tổng kết môn): 8/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 5
- Link Github: https://github.com/NTD1810/CS2205.NOV2024

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

ỨNG DỤNG TRANSFER LEARNING ĐA NGÔN NGỮ TRONG PHÁT HIỆN SPAM EMAIL TIẾNG VIỆT

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

APPLYING MULTILINGUAL TRANSFER LEARNING FOR VIETNAMESE EMAIL SPAM DETECTION

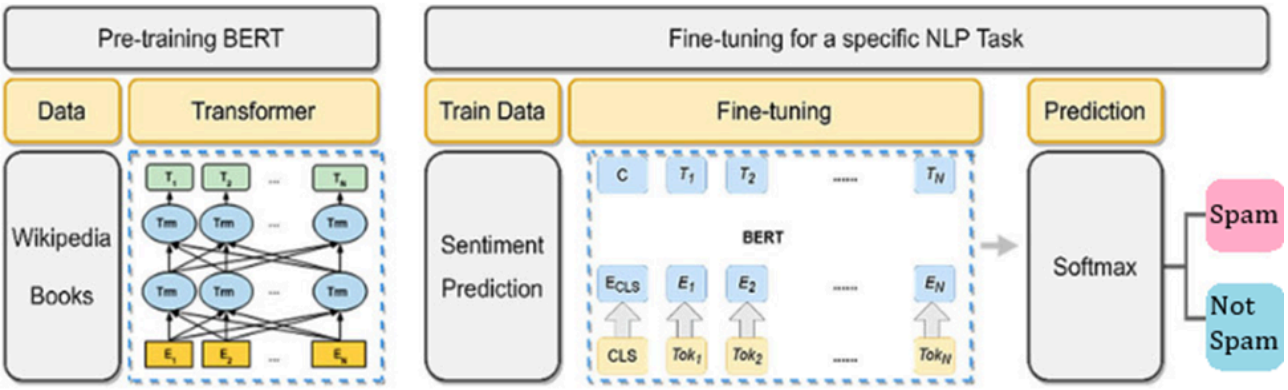
TÓM TẮT *(Tối đa 400 từ)*

Email spam đang trở thành một vấn đề nghiêm trọng đối với người dùng internet Việt Nam trong thời đại số hóa. Do đặc điểm của ngôn ngữ và sự thiếu hụt dữ liệu hướng dẫn tiếng Việt chất lượng cao, các phương pháp phát hiện spam truyền thống trở nên khó khăn. Các nghiên cứu gần đây chỉ ra rằng việc học đa ngôn ngữ có thể tận dụng kiến thức từ các ngôn ngữ giàu dữ liệu để cải thiện hiệu quả các tác vụ xử lý ngôn ngữ nói (NLP) trong các ngôn ngữ ít dữ liệu. Chúng tôi đề xuất một mô hình dựa trên kỹ thuật chuyển đổi đa ngôn ngữ để tìm spam email tiếng Việt bằng cách kết hợp các mô hình ngôn ngữ tiên tiến như mBERT và XLM-R. Một bộ dữ liệu song ngữ Anh-Việt được thu thập thủ công và một bộ dữ liệu tự động được thu thập từ email công khai. Theo thử nghiệm, mô hình đề xuất đạt được độ chính xác cao trong việc xác định spam email tiếng Việt so với các phương pháp cơ bản truyền thống.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Email spam đang là một trong những mối đe dọa lớn đối với an ninh mạng và trải nghiệm người dùng tại Việt Nam. Với sự phát triển của công nghệ, các email spam ngày càng tinh vi và khó phát hiện hơn. Các phương pháp truyền thống dựa trên luật và học máy cơ bản thường gặp khó khăn do đặc thù của tiếng Việt và sự thiếu hụt dữ liệu huấn luyện chất lượng cao.

Transfer Learning, một kỹ thuật trong học máy, cho phép mô hình học được kiến thức từ một miền nguồn và áp dụng vào miền đích có ít dữ liệu hơn. Trong lĩnh vực xử lý ngôn ngữ tự nhiên, Transfer Learning đã chứng minh hiệu quả vượt trội khi áp dụng cho các ngôn ngữ ít tài nguyên. Đặc biệt, sự ra đời của các mô hình ngôn ngữ đa ngôn ngữ như mBERT và XLM-R đã mở ra khả năng học chuyển giao kiến thức giữa các ngôn ngữ khác nhau một cách hiệu quả.



Transfer Learning đa ngôn ngữ mở ra một hướng tiếp cận mới, cho phép tận dụng kiến thức từ các ngôn ngữ giàu dữ liệu như tiếng Anh để cải thiện hiệu quả phát hiện spam cho tiếng Việt. Phương pháp này đặc biệt hữu ích trong bối cảnh tiếng Việt, khi mà việc thu thập và gán nhãn dữ liệu chất lượng cao là một thách thức lớn. Bằng cách tận dụng các đặc trưng chung và cấu trúc ngôn ngữ tương đồng giữa tiếng Anh và tiếng Việt trong email spam, Transfer Learning có thể giúp mô hình học được các mẫu phát hiện spam hiệu quả hơn, ngay cả khi dữ liệu tiếng Việt còn hạn chế.

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

- Đề xuất mô hình Transfer Learning đa ngôn ngữ kết hợp mBERT/XLM-R với các kỹ thuật học sâu để phát hiện spam email tiếng Việt hiệu quả.
- Xây dựng bộ dữ liệu song ngữ Anh-Việt chất lượng cao về spam email phục vụ nghiên cứu và đánh giá.
- Cải thiện, nâng cao độ chính xác trong việc phát hiện spam email tiếng Việt thông qua việc tận dụng kiến thức từ dữ liệu tiếng Anh.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung: Nghiên cứu tập trung vào việc phát triển mô hình Transfer Learning đa ngôn ngữ cho phát hiện spam email tiếng Việt. Mô hình được đề xuất kết hợp các kiến trúc transformer đa ngôn ngữ (mBERT, XLM-R) với các lớp neural network chuyên biệt cho bài toán phát hiện spam. Khác với các nghiên cứu trước chỉ sử dụng mô hình đơn ngôn ngữ, chúng tôi khai thác khả năng học biểu diễn chung giữa các ngôn ngữ của các mô hình pre-trained đa ngôn ngữ.

Phương Pháp: Nghiên cứu được tiến hành qua ba giai đoạn chính: thu thập và xử lý dữ liệu, phát triển mô hình, và đánh giá triển khai.

- Trong giai đoạn thu thập và xử lý dữ liệu, chúng tôi kết hợp nhiều phương pháp bao gồm xây dựng crawler tự động thu thập email từ các nguồn công khai, thu thập thủ công từ các hộp thư, và tận dụng các bộ dữ liệu email spam tiếng Anh chuẩn. Quy trình dịch tự động kết hợp kiểm tra chéo được áp dụng để tạo các cặp dữ liệu song ngữ chất lượng cao. Dữ liệu sau đó được tiền xử lý kỹ lưỡng thông qua việc loại bỏ các email trùng lặp, chuẩn hóa định dạng văn bản, tách từ cho cả tiếng Việt và tiếng Anh, đồng thời loại bỏ thông tin nhạy cảm. Để đảm bảo chất lượng gán nhãn, chúng tôi xây dựng hướng dẫn chi tiết và gán nhãn độc lập cho mỗi email, sử dụng điểm Cohen's Kappa để đánh giá độ nhất quán và giải quyết các trường hợp bất đồng thông qua thảo luận.
- Giai đoạn phát triển mô hình tập trung vào việc thiết kế một kiến trúc hiệu quả dựa trên các mô hình nền tảng mBERT/XLM-R. Kiến trúc này được tăng cường với các lớp Attention đa đầu để xử lý ngữ cảnh tốt hơn, kết hợp với lớp phân loại được tối ưu hóa bằng dropout và batch normalization. Quá trình huấn luyện được thực hiện thông qua phương pháp fine-tuning theo từng giai đoạn, áp dụng các kỹ thuật regularization như gradient clipping và early stopping. Việc tối ưu hóa mô hình được thực hiện thông qua thử nghiệm với nhiều hyperparameter khác nhau, áp dụng kỹ thuật học tỷ lệ, và tối ưu hóa bộ nhớ

cũng như tốc độ xử lý.

- Giai đoạn cuối cùng tập trung vào đánh giá và triển khai thực tế. Hiệu suất của mô hình được đánh giá toàn diện thông qua các metrics như Accuracy, Precision, Recall, F1-score, và được so sánh với các phương pháp baseline truyền thống. Chúng tôi cũng phân tích kỹ lưỡng hiệu suất trên các loại spam khác nhau và đánh giá khả năng mở rộng của hệ thống.

KẾT QUẢ MONG ĐỢI

- Mô hình Transfer Learning đa ngôn ngữ với độ chính xác cao hơn (trên 95%) so với các mô hình khác trong phát hiện spam email tiếng Việt.
- Bộ dữ liệu song ngữ Anh-Việt về spam email với hơn 100,000 mẫu đã được gán nhãn.
- Hệ thống hoàn chỉnh có thể triển khai trong thực tế.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL 2019

[2] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V. (2020) "Unsupervised Cross-lingual Representation Learning at Scale." ACL 2020

[3] Nguyen, D.Q., Nguyen, A.T. (2020) "PhoBERT: Pre-trained language models for Vietnamese." EMNLP 2020

[4] Liu, Y., Ott, M., Goyal, N., Du, J. (2019) "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv preprint

[5] Lample, G., Conneau, A. (2019) "Cross-lingual Language Model Pretraining." NeurIPS 2019