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): 9/10
- 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)*

Vấn nạn thư rác đang bủa vây người dùng mạng Việt Nam trong kỷ nguyên số. Rào cản ngôn ngữ và sự khan hiếm dữ liệu huấn luyện tiếng Việt chất lượng đã vô hiệu hóa các công cụ lọc spam cổ điển. Các phát kiến gần đây chỉ ra rằng transfer learning thông qua mô hình đa ngữ có khả năng chuyển giao kiến thức từ các ngôn ngữ dữ liệu dồi dào để tăng cường xử lý ngôn ngữ tự nhiên cho những ngôn ngữ ít tài nguyên. Nghiên cứu của chúng tôi khai thác kỹ thuật transfer learning đa ngữ, kết hợp sức mạnh của mBERT và XLM-R để săn lùng spam tiếng Việt. Chúng tôi đã xây dựng kho ngữ liệu song ngữ Anh-Việt theo hai hướng: thu thập thủ công và khai thác tự động từ email công khai. Kết quả thử nghiệm cho thấy mô hình transfer learning đề xuất vượt xa hiệu suất của các phương pháp truyền thống trong cuộc chiến chống spam tiếng Việt.

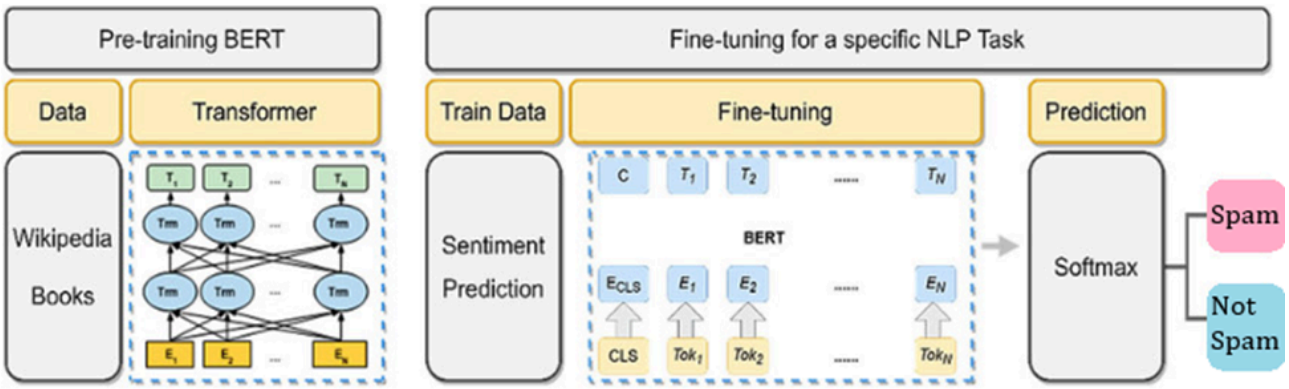
GIỚI THIỆU *(Tối đa 1 trang A4)*

Email spam đang đe dọa an ninh mạng và trải nghiệm người dùng Việt Nam, với thủ đoạn ngày càng tinh vi. Phương pháp truyền thống thường bất lực trước đặc thù tiếng Việt và cơn khát dữ liệu chất lượng.

Transfer learning - chiến lược "chắp cánh tri thức" từ miền nguồn sang miền đích

nghèo dữ liệu - đã chứng tỏ sức mạnh vượt trội trong xử lý ngôn ngữ ít tài nguyên. Sự xuất hiện của mBERT và XLM-R mở ra cánh cửa chuyển giao kiến thức liên ngôn ngữ hiệu quả.

Transfer learning đa ngôn ngữ tạo bước đột phá, cho phép "vay mượn" tri thức từ kho báu dữ liệu tiếng Anh để nâng cao hiệu suất lọc spam tiếng Việt. Phương pháp này đặc biệt quý giá khi việc thu thập và gán nhãn dữ liệu tiếng Việt còn nhiều chông gai. Bằng cách khai thác đặc điểm chung và cấu trúc tương đồng giữa hai ngôn ngữ, transfer learning giúp mô hình nhận diện mẫu spam hiệu quả dù trong điều kiện dữ liệu tiếng Việt còn khiêm tốn.



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

- Thiết kế mô hình transfer learning đa ngôn ngữ tích hợp mBERT/XLM-R cùng công nghệ học sâu tiên tiến để săn lùng thư rác tiếng Việt với độ chính xác vượt trội.
- 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 hiệu suất phát hiện spam tiếng Việt thông qua việc "mượn sức" từ kho báu tri thức ẩn trong 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. Đột phá giới hạn nghiên cứu cũ bó buộc trong khuôn khổ mô hình đơn ngữ, chúng tôi khai phá tiềm năng transfer learning qua việc nắm bắt biểu diễn chia sẻ liên ngôn ngữ từ các mô hình pre-trained đa ngôn ngữ tiên tiến.

Phương pháp: Nghiên cứu qua ba giai đoạn:

- Giai đoạn thu thập và xử lý dữ liệu: chúng tôi triển khai chiến lược thu thập dữ liệu đa chiều: xây dựng crawler săn lùng email từ không gian công khai, thu thập thủ công từ các hộp thư, và "mượn sức" từ kho dữ liệu email spam tiếng Anh chuẩn mực. Quy trình transfer learning bắt đầu từ giai đoạn dịch tự động kết hợp kiểm định chéo để tạo cặp song ngữ đỉnh cao. Dữ liệu sau đó trải qua quá trình "thanh lọc": loại bỏ bản sao, chuẩn hóa văn bản, tách từ song ngữ, và lọc thông tin nhạy cảm. Để đảm bảo độ tin cậy, chúng tôi thiết lập quy trình gán nhãn nghiêm ngặt với hướng dẫn cụ thể, thực hiện gán nhãn độc lập, ứng dụng thang đo Cohen's Kappa để đánh giá tính nhất quán, và giải quyết xung đột qua thảo luận chuyên sâu.
- Giai đoạn phát triển mô hình: đổi trọng tâm sang việc kiến tạo cấu trúc transfer learning đột phá, được xây dựng trên nền tảng mBERT/XLM-R. Kiến trúc này được nâng cấp với các lớp Attention đa đầu tinh xảo để nắm bắt bối cảnh sâu rộng, kết hợp lớp phân loại siêu việt nhờ dropout và batch normalization. Quá trình huấn luyện được điều phối qua chiến lược fine-tuning đa tầng, tích hợp các kỹ thuật regularization như gradient clipping và early stopping. Công cuộc tối ưu mô hình được thực hiện qua hành trình khám phá không gian hyperparameter đa chiều, áp dụng kỹ thuật học tỷ lệ tiên tiến, cùng giải pháp tối ưu bộ nhớ và tốc độ xử lý vượt trội.

- Giai đoạn đánh giá và triển khai: khả năng transfer learning được đo lường toàn diện qua bộ chỉ số Accuracy, Precision, Recall và F1-score, đối chiếu với các phương pháp baseline cổ điển để chứng minh ưu thế vượt trội. Chúng tôi còn mổ xẻ chi tiết hiệu suất trên từng loại thư rác đặc thù và khảo sát tiềm năng mở rộng của hệ thống trong môi trường thực tế.

KẾT QUẢ MONG ĐỢI

- Mô hình transfer learning đa ngôn ngữ đạt đỉnh cao với độ chính xác vượt ngưỡng 95%, áp đảo các đối thủ cạnh tranh trong lĩnh vực phát hiện thư rác 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