Ứng dụng Transfer Learning đa ngôn ngữ trong phát hiện Spam Mail Tiếng việt

Tác giả

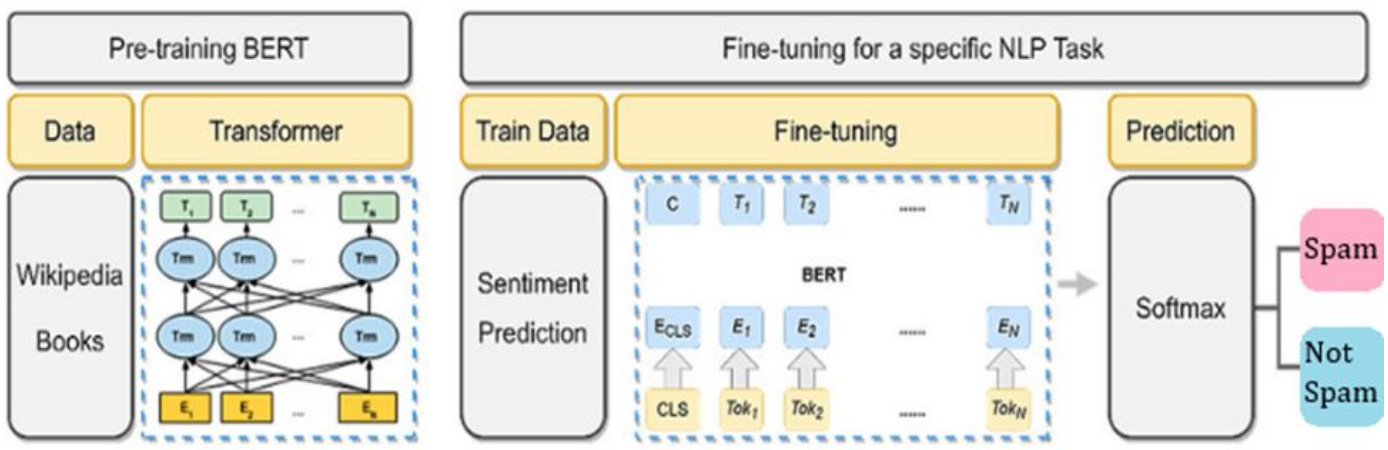
Nguyễn Thuỳ Dương
Trường Đại học Công nghệ thông tin TPHCM
Email: duongnt.19@grad.uit.edu.vn

Lý do chọn đề tài?

Với sự gia tăng của email spam tiếng Việt gây nhiều thiệt hại cho người dùng, việc nghiên cứu phương pháp phát hiện hiệu quả trở nên cần thiết. Transfer Learning đa ngôn ngữ mang lại lợi thế đáng kể khi có thể tận dụng kiến thức từ các mô hình đã được huấn luyện trên dữ liệu lớn từ các ngôn ngữ khác, qua đó giảm nhu cầu về dữ liệu huấn luyện tiếng Việt vốn còn hạn chế. Đây không chỉ là một hướng nghiên cứu mới trong lĩnh vực xử lý ngôn ngữ tự nhiên tiếng Việt, mà còn có tính ứng dụng thực tiễn cao khi có thể tích hợp trực tiếp vào các hệ thống email để bảo vệ người dùng khỏi các email lừa đảo và độc hại.

Giới thiệu

Transfer Learning - một kỹ thuật trong học máy - đã mở ra một hướng tiếp cận mới đầy triển vọng. Kỹ thuật này cho phép mô hình học được kiến thức từ một miền nguồn và áp dụng vào miền đích có ít dữ liệu hơn. Đặc biệt trong lĩnh vực xử lý ngôn ngữ tự nhiên, sự ra đời của các mô hình ngôn ngữ đa ngôn ngữ như mBERT và XLM-R đã tạo ra bước đột phá trong việc học chuyển giao kiến thức giữa các ngôn ngữ khác nhau.



MỤC TIÊU

1. 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.
2. Xây dựng bộ dữ liệu song ngữ Anh-Việt chất lượng cao về spam email phục vụ nghiên cứu và đánh giá.
3. Cải thiện độ chính xác trong việc phát hiện spam email tiếng Việt thông qua việc tận dụng kiến thức từ dữ liệu tiếng Anh.

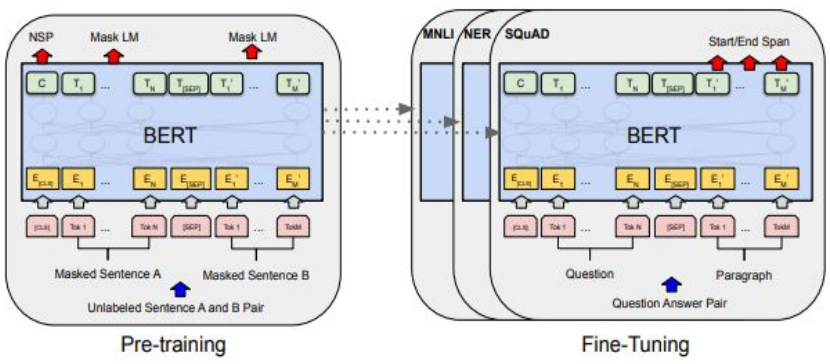
PHƯƠNG PHÁP

Thu thập và xử lý dữ liệu: Kết hợp nhiều phương pháp như xây dựng crawler tự động, thu thập thủ công và tận dụng các bộ dữ liệu chuẩn. Dữ liệu sau đó được tiền xử lý kỹ lưỡng và gán nhãn.
Phát triển mô hình: Thiết kế kiến trúc kết hợp giữa các mô hình nền tảng mBERT và XLM-R với các lớp Attention đa đầu. Quá trình huấn luyện được thực hiện theo phương pháp fine-tuning từng giai đoạn, áp dụng nhiều kỹ thuật regularization hiện đại.
Đánh giá và triển khai: Đánh giá toàn diện và so sánh với các phương pháp baseline. Xây dựng hệ thống hoàn chỉnh với API và dashboard để triển khai trong môi trường thực tế.

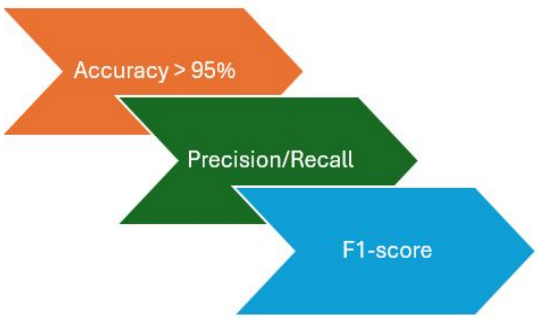
KẾT QUẢ

- Mô hình có độ chính xác trên 95% trong việc phát hiện spam email tiếng Việt
- Bộ dữ liệu song ngữ với hơn 100,000 mẫu đã được gán nhãn
- Hệ thống hoàn chỉnh có thể triển khai trong thực tế

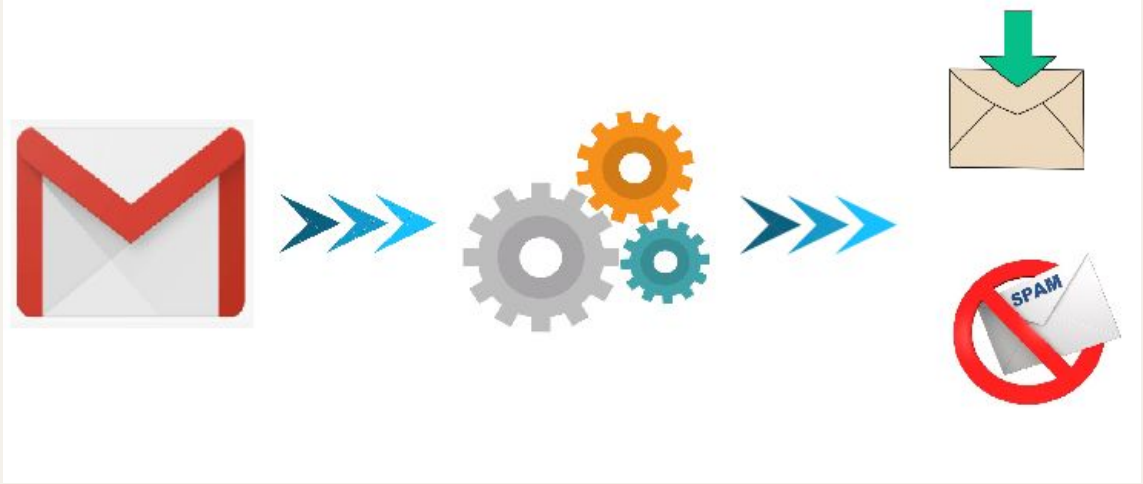
Thu thập và xử lý dữ liệu



Phát triển mô hình



Đánh giá mô hình



Kết luận

Với những đề xuất trên, nghiên cứu này không chỉ đóng góp một giải pháp hiệu quả cho bài toán phát hiện spam email tiếng Việt, mà còn mở ra hướng nghiên cứu mới trong việc ứng dụng Transfer Learning đa ngôn ngữ cho các bài toán xử lý ngôn ngữ tự nhiên khác tại Việt Nam.