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

Nguyễn Thuỳ Dương - 240202006

Tóm tắt

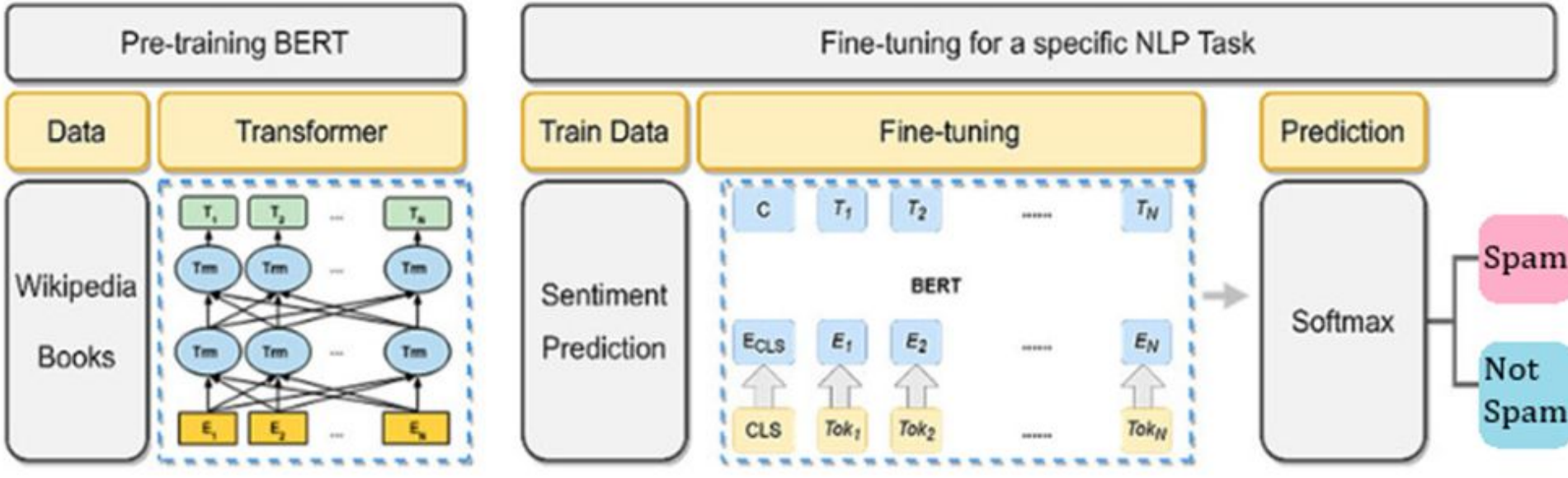
- Lớp: CS2205.NOV2024
- Link Github: https://github.com/NTD1810/CS2205.NOV2024
- Link YouTube video: https://www.youtube.com/watch?v=BhY8QaaLWhY
- Ảnh + Họ và Tên: Nguyễn Thuỳ Dương

Giới thiệu

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 trở nên tinh vi và khó phát hiện hơn.

Giới thiệu

Hướng giải quyết: Transfer Learning - một kỹ thuật trong học máy - đã mở ra một hướng tiếp cận mới đầy triển vọng.



Giới thiệu

Tính thực tiễn và khả thi: Phương pháp này đặc biệt phù hợp với 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, chúng ta có thể xây dựng các mô hình phát hiện spam hiệu quả hơn, ngay cả khi dữ liệu tiếng Việt còn hạn chế.

Mục tiêu

- Đề xuất một mô hình Transfer Learning đa ngôn ngữ kết hợp mBERT và XLM-R với các kỹ thuật học sâu để phát hiện spam email tiếng Việt một cách hiệu quả.
- Xây dựng một bộ dữ liệu song ngữ Anh-Việt chất lượng cao về spam email phục vụ cho nghiên cứu và đánh giá.
- 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.

Nội dung và Phương pháp

Thu thập và xử lý dữ liệu:

- Xây dựng crawler tự động thu thập email
- Sử dụng bộ dữ liệu chuẩn có sẵn
- Tiền xử lý và gán nhãn dữ liệu

Phát triển mô hình:

- Thiết kế kiến trúc dựa trên mBERT và XLM-R
- Tích hợp lớp Attention đa đầu
- Fine-tuning theo từng giai đoạn
- Áp dụng kỹ thuật regularization

Nội dung và Phương pháp

Đánh giá và triển khai:

- Đánh giá toàn diện mô hình
- So sánh với các phương pháp baseline
- Xây dựng API cho hệ thống
- Phát triển dashboard giám sát
- Triển khai trong môi trường thực tế

Kết quả dự kiến

- 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

- [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