PHÁT HIỆN TIN GIẢ BẰNG CÁCH SỬ DỤNG MÁY HỌC VÀ MÔ HÌNH BERT

TRẦN TRỌNG NGỌC TÀI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Mục tiêu?

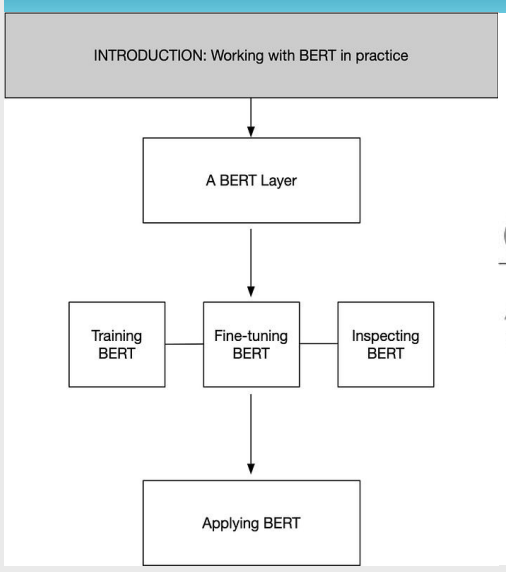
Chúng tôi ứng dụng mô hình trong việc phát hiện phân loại tin giả. Trong đó mục tiêu hướng đến:

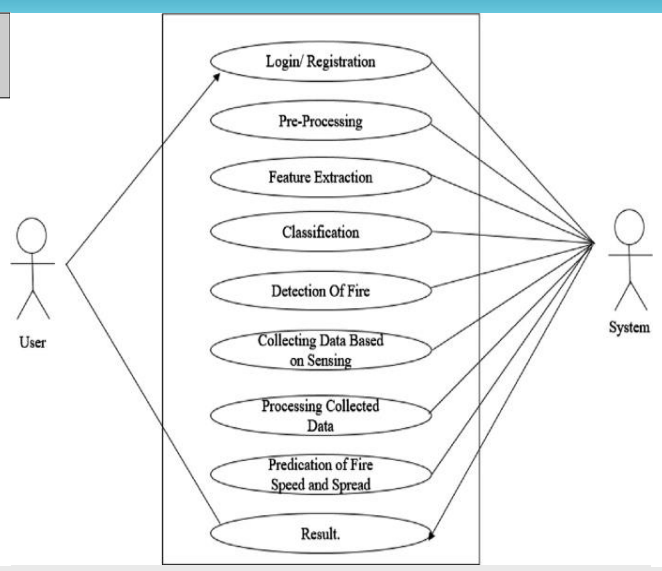
- Đề xuất mô hình (FakeBERT) kết hợp giữa BERT và mạng neural tích chập CNN để đối phó với sự lan truyền nhanh chóng của tin giả.
- Cải thiện độ chính xác trong việc phân loại tin thật/tin giả bằng cách phân tích ngữ cảnh hai chiều của văn bản.
- Giảm thiểu tác động tiêu cực của tin giả đến xã hội.

Lý do chọn đề tài ?

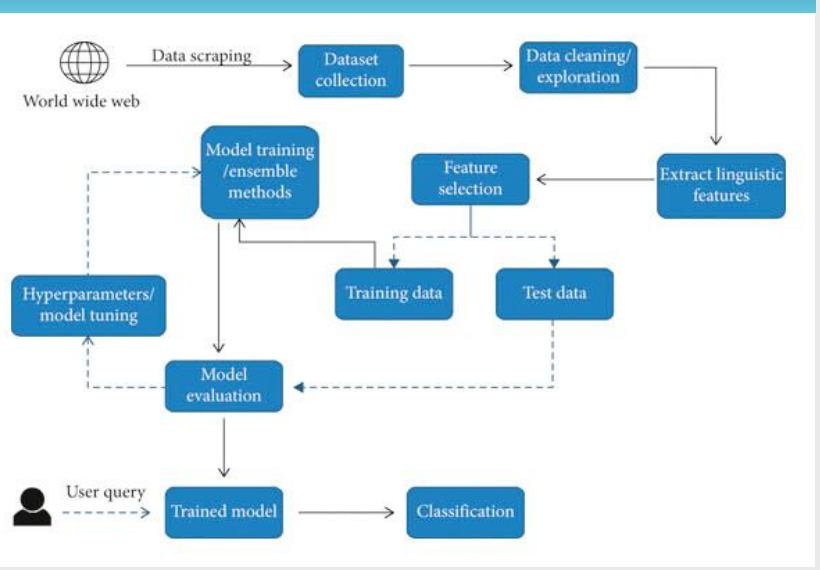
- Tin giả trên mạng xã hội và nhiều phương tiện truyền thông khác đang lan truyền rộng rãi và là vấn đề đáng quan ngại nghiêm trọng do khả năng gây ra nhiều thiệt hại về mặt xã hội và quốc gia với những tác động tiêu cực
- Khám phá các mô hình học máy truyền thống để lựa chọn mô hình tốt nhất, nhằm tạo ra mô hình sản phẩm có thuật toán học máy có thể giám sát, phân loại tin giả bằng cách sử dụng các công cụ như python scikit-learn, NLP để phân tích văn bản.

TỔNG QUAN





Phương pháp phát hiện tin giả và ngữ cảnh



Kiến trúc mô hình phân loại BERT

MÔ TẢ

Nội dung

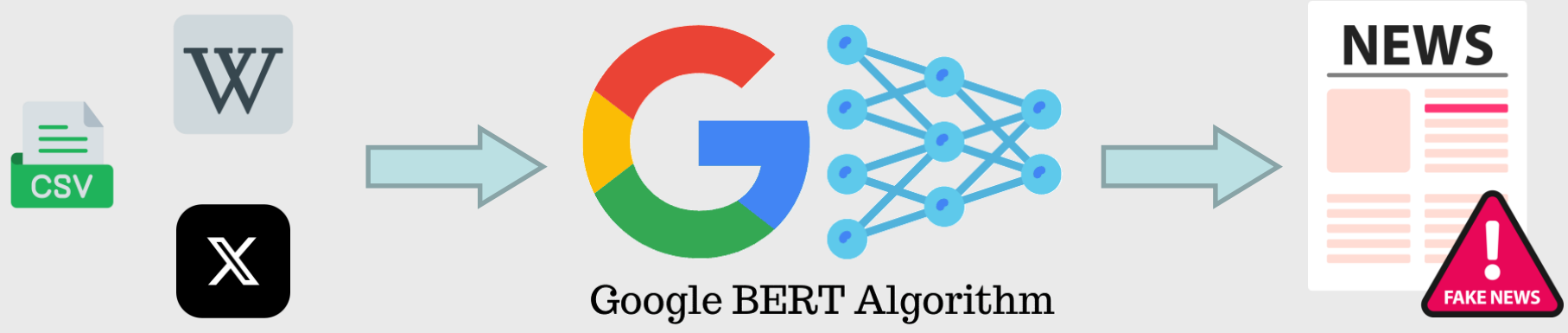
- Phương pháp học sâu dựa trên BERT (FakeBERT) bằng cách kết hợp các khối song song khác nhau của CNN với Bidirectional Encoder Representations from Transformers (BERT).
- Sử dụng các mô hình sau: Logistic Regression, Decision Tree, Gradient Booster, Random Forests, BERT

Phương pháp

- Thu thập dữ liệu từ nhiều nguồn tin tức
- Tiền xử lý và làm sạch dữ liệu
- Trích xuất đặc trưng ngôn ngữ bằng LIWC2015
- Chia dữ liệu training/testing tỷ lệ 70/30
- Huấn luyện các mô hình machine learning
- Đánh giá hiệu năng dựa trên nhiều tiêu chí:
  - ✓ Accuracy (Độ chính xác)
  - ✓ FPR (False Positive Rate - Tỷ lệ dương tính giả)
  - ✓ FNR (False Negative Rate - Tỷ lệ âm tính giả)
  - ✓ Cross-entropy loss (Độ mất mát entropy chéo)

Kết quả đạt được

- Mô hình FakeBERT đạt độ chính xác 98.90%
- Trích xuất 93 tính năng khác nhau từ bất kỳ văn bản nào để phát hiện tin giả
- Đề xuất hướng nghiên cứu tiếp về hiện tượng “echo chambers” trên mạng xã hội



TRẦN TRỌNG NGỌC TÀI - Trường Đại học Công nghệ thông tin TPHCM

TEL : 0378855787 Email : taittn.19@grad.uit.edu.vn