PHÁT HIỆN TIN GIẢ BẰNG CÁCH SỬ DỤNG MÔ HÌNH BERT

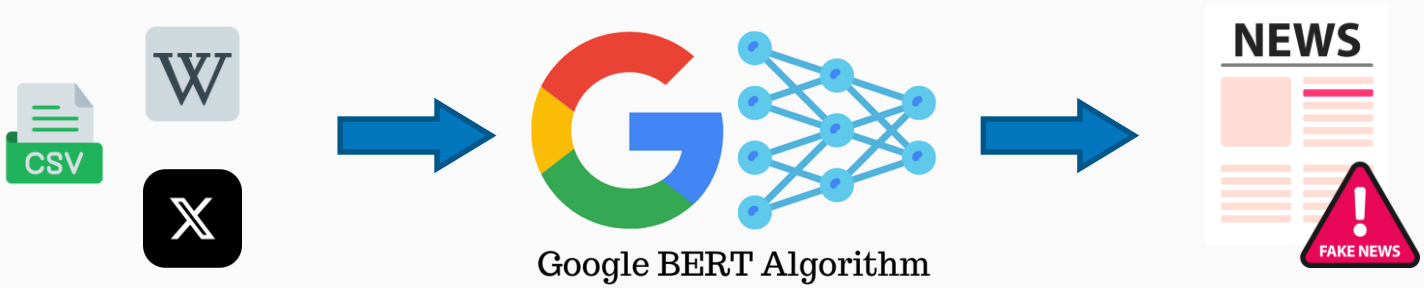
Trần Trọng Ngọc Tài - 240202011

Tóm tắt

- Lớp: CS2205.CH183
- Link Github: https://github.com/TaiTranTrongNgoc/CS2205.CH183
- Link YouTube video: https://youtu.be/MDBoQ3-xnJQ
- Ảnh + Họ và Tên: Trần Trọng Ngọc Tài - 240202011

Giới thiệu

- Vấn đề tin giả đang lan truyền nhanh chóng trên môi trường mạng và gây tác động tiêu cực đến xã hội
- Các phương pháp trước đây chủ yếu phân tích văn bản một chiều
- Xem xét các kỹ thuật và lĩnh vực khác nhau của máy học (ML), xử lý ngôn ngữ tự nhiên (NLP) và trí tuệ nhân tạo (AI) để phát hiện tin giả



Mục tiêu

- Đề xuất mô hình học sâu FakeBERT kết hợp giữa mô hình 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.

Nội dung và Phương pháp

NỘI DUNG:

Sử dụng mô hình FakeBERT kết hợp giữa:

- BERT (Bidirectional Encoder Representations from Transformers)
- Mạng CNN (Convolutional Neural Network) với nhiều khối song song
- Tạo một mô hình có thể phân loại tin tức giả là đúng hay sai bằng cách sử dụng các công cụ như Python scikit-learn và NLP để phân tích văn bản.

Có 5 mô hình machine learning được sử dụng:

- Regress Logistic
- Decision Tree
- Gradient Booster
- Random Forests và BERT

Nội dung và Phương pháp

PHƯƠNG PHÁP:

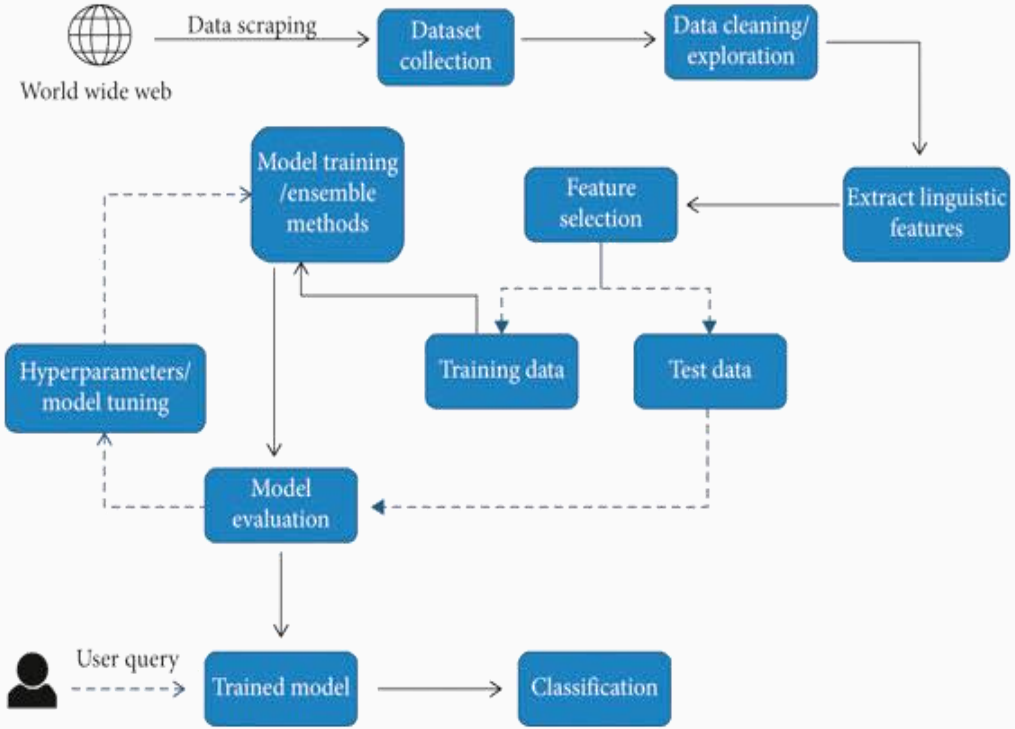
Kiến trúc và quy trình:

- 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)

Nội dung và Phương pháp



Kiến trúc mô hình phân loại BERT

Kết quả dự kiến

- 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

Tài liệu tham khảo

[1] Crestani F, Rosso P (2020) The role of personality and linguistic patterns in discriminating between fake news spreaders and fact-checkers.” I n Natural language processing and information systems: 25th international conference on applications of natural language to information systems, NLDB 2020, Saarbrucken, Germany. Proceedings, vol 181. Springer Nature

[2] Alkhodair S A, Ding S H.H, Fung B C M and Liu J 2020 Detecting breaking news rumors of emerging topics in social media” Inf. Process. Manag. 57 102018 2020

[3] Kaur Prabhjot et al 2019 Hybrid Text Classification Method for Fake News Detection Inf.” International Journal of Engineering and Advanced Technology (IJEAT) 2388-2392

[4] Bondielli A, Marcelloni F (2019) A survey on fake news and rumor detection techniques.” Inform Sci 497:38–55

[5] Chen W, Zhang Y, Yeo CK, Lau CT, Sung Lee B (2018) Unsupervised rumor detection based on users’ behaviors using neural networks.” ‘Pattern

[6] De S, Sohan FY, Mukherjee A (2018) Attending sentences to detect satirical fake news. ” In: Proceedings of the 27th international conference on computational linguistics, pp 3371–3380

[7] Ahmed H, Traore I, Saad S (2017) Detection of online fake news using N-gram analysis and machine learning techniques. systems.,” : International conference on intelligent, secure, and dependable systems in distributed and cloud environments. Springer, Cham, pp 127–138

[8] Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election.” Econ Perspect 31(2):211–36

[9] Granik Mykhailo and Mesyura Volodymyr 2017 First Ukraine Conference on Electrical and Computer Engineering (UKRCON) (Ukraine: IEEE) Fake news detection using naive Bayes classifier” Journal of Computational and Theoretical Nanoscience., 12. 6334-6342. 10.1166/jctn.2015.4675.

[10] Greff K, Srivastava RK, Koutn´ık J, Steunebrink BR, Schmidhuber J (2016) LSTM” IEEE Trans Neural Netw Learn Syst 28(10):2222–2232, [7] De S, Sohan FY, Mukherjee A (2018) Attending sentences to detect satirical fake news. ” In: Proceedings of the 27th international conference on computational linguistics, pp 3371–3380