TĂNG CƯỜNG PHÁT HIỆN TIN GIẢ BẰNG TINH CHỈNH MÔ HÌNH HỌC SÂU THEO MIỀN CHO TRÍCH XUẤT ĐẶC TRƯNG NGỮ NGHĨA

Lưu Nguyễn Công Minh - 240202024

Tóm tắt

- Lớp: CS2205.FEB2025
- Link Github của nhóm: https://github.com/P0uut
- Link YouTube video:
- Lưu Nguyễn Công Minh

Giới thiệu

- Trong những năm gần đây, sự lan truyền nhanh chóng của tin giả (fake news) trên các nền tảng kỹ thuật số đã gây ra những lo ngại nghiêm trọng trong lĩnh vực như y tế công cộng, chính trị và an ninh toàn cầu
- Với xu thế hiện nay với các mô hình ngôn ngữ lớn (LLM) đang ngày càng được phát triển, các mô hình LLM cũng được ứng dụng vào các nghiên cứu giúp hỗ trợ việc tự động phát hiện tin giả một cách hiệu quả và chính xác.

Giới thiệu

- LESS4FD (viết tắt của LLM Enhanced Semantics Mining for Fake News Detection), là một phương pháp được đề xuất, phát hiện tin giả bằng cách mô hình hóa các mối quan hệ giữa nội dung tin tức, các thực thể được đặt tên và các chủ đề trong một đồ thị không đồng nhất (heterogeneous graph)
- Tuy nhiên, các LLM được sử dụng trong LESS4FD được thực hiện theo phương pháp hộp đen zero-shot, nghĩa là các mô hình này không được tối ưu cho lĩnh vực về tin giả

Mục tiêu

- Tinh chỉnh mô hình LLM nhỏ hơn như Mistral-7B với dataset LIAR
- Tích hợp mô hình LLM đã được tinh chỉnh vào framework LESS4FD trong bài báo
- Thực nghiệm trên dataset LIAR đã được chia sẵn tập dùng để đánh giá

Nội dung và Phương pháp

- **Phase 1: Tinh chỉnh mô hình LLM**

❑ Phần 1: Thu thập và xử lí dữ liệu (LIAR)

❑ Phần 2: Tinh chỉnh mô hình (Mistral-7B + LoRA)

❑ Phần 3: Đánh giá và lưu mô hình

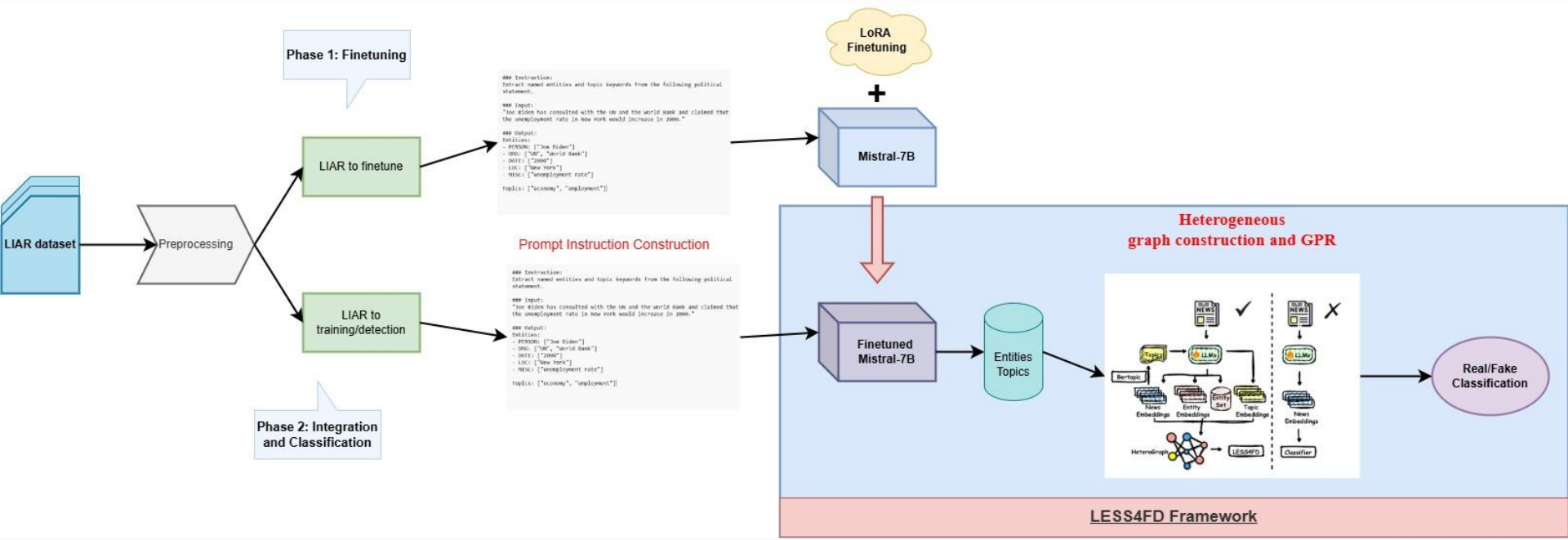
- **Phase 2: Phân loại với framework LESS4FD**

❑ Phần 1: Trích xuất thực thể và topic với mô hình LLM đã được tinh chỉnh

❑ Phần 2: Xây dựng đồ thị (heterogeneous graph) và lan truyền đặc trưng (feature propagation)

❑ Phần 3: Đánh giá công việc phát hiện tin giả

Nội dung và Phương pháp



Kết quả dự kiến

- Mô hình LLM được tinh chỉnh được kì vọng sẽ tạo ra các trích xuất thực thể và chủ đề chính xác hơn
- Việc tích hợp các đặc trưng ngữ nghĩa được cải thiện này vào framework LESS4FD sẽ dẫn đến hiệu suất phát hiện tin giả được cải thiện,
- Chứng minh hiệu quả của việc tinh chỉnh cho lĩnh vực trích xuất ngữ nghĩa và phát hiện tin giả

Tài liệu tham khảo

[1] Ma, Xiaoxiao, et al. "On Fake News Detection with LLM Enhanced Semantics Mining." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

[2] https://mistral.ai/news/announcing-mistral-7b

[3] https://paperswithcode.com/dataset/liar

[4] https://spacy.io/

[5] Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint arXiv:2203.05794 (2022).

[6] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3.