SO SÁNH CÁC KỸ THUẬT XỬ LÝ NGÔN NGỮ TỰ NHIÊN KHÁC NHAU TRONG VIỆC PHÁT HIỆN VĂN BẢN DO AI TẠO RA VÀ DIỄN GIẢI LẠI

Lê Trường Long Hưng - 230101007

Tóm tắt

- Lớp: CS2205.MAR2024
- Link Github: https://github.com/HungLTL/CS2205.MAR2024
- Link YouTube video: https://www.youtube.com/watch?v=_brG2jb7OP8
- Họ và Tên: Lê Trường Long Hưng

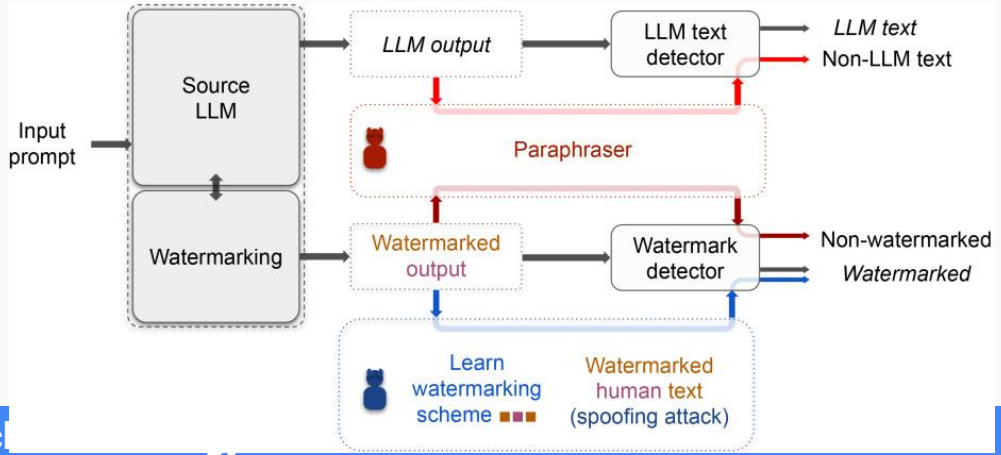
Giới thiệu

- AI: ngày càng là chuyên ngành trọng điểm
- Chatbot Q&A (e.g. Gemini, ChatGPT) có khả năng tạo nội dung theo yêu cầu
- Chưa có quy định/luật lệ rõ ràng -> dễ bị lợi dụng vào những mục đích xấu

Giới thiệu

- Nhiều công cụ, mô hình xác thực nội dung với nhiều phương pháp, nền tảng khác nhau
- Một số lỗ hỏng vẫn tồn, e.g. diễn giải lại

=> Nghiên cứu so sánh, tìm phương pháp hiệu quả nhất, chỉnh sửa để khắc phục khuyết điểm, định hướng phát triển công cụ xác thực



Mục tiêu

- Xây dựng, huấn luyện mô hình xác thực văn bản dựa trên phương pháp sẵn có, phối hợp cải thiện để khắc phục khiếm khuyết
- So sánh các phương pháp thông qua thực nghiệm trên các mô hình, đặc biệt chú ý khả năng chống lại các lỗ hỏng đã biết
- Xác định các phương pháp, thuật toán hữu hiệu nhất để định hướng phát triển công cụ phát hiện AI

Nội dung

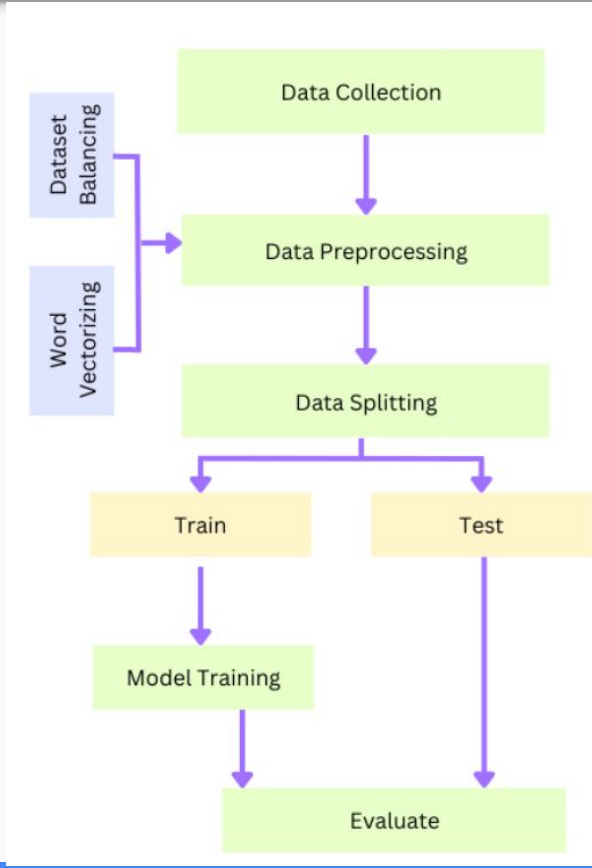
- Tìm hiểu nghiên cứu một số phương pháp, thuật toán NLP được vận dụng phổ biến (e.g. zero-shot, chữ ký điện tử, học đối kháng, etc.)
- Điều tra phân tích nguyên nhân các lỗ hỏng đã biết ảnh hưởng độ chính xác (đặc biệt chú ý: tấn công diễn giải lại)
- Xây dựng các mô hình nguyên mẫu, chú ý cải thiện để khắc phục các lỗ hỏng
- Thử nghiệm trên các bộ dữ liệu, so sánh kết quả của từng mô hình

Phương pháp

- Chuẩn bị các bộ dữ liệu:
  - Dataset có sẵn: XSum, WritingPrompts, etc.
  - Tạo dataset riêng: crawl dữ liệu từ diễn đàn hỏi đáp (e.g. r/explainlikeimfive trên Reddit, StackOverflow)
  - Thử nghiệm lỗ hỏng tấn công diễn giải đệ quy: chạy các bộ dữ liệu qua một mô hình diễn giải lại văn bản

Phương pháp

- Nghiên cứu các phương pháp của những công cụ hiện tại
- Xây dựng lại mô hình mới dựa trên tiến trình chung (hình)
- Phân tích nguyên nhân các lỗ hỏng giảm độ chính xác -> chỉnh sửa các mô hình
- Huấn luyện, đánh giá và thống kê kết quả



Kết quả dự kiến

- Cải thiện các phương pháp hiện có, đặc biệt trước các lỗ hỏng thường gặp
- Các mô hình có độ chính xác tốt sau thử nghiệm
- Xác định những phương pháp, thuật toán nên được tập trung nghiên cứu, phát triển thêm

Tài liệu tham khảo

[1]. Elkhatat, A.M., Elsaid, K. & Almeer, S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educ Integr 19, 17 (2023). https://doi.org/10.1007/s40979-023-00140-5
[2]. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P. & Waddington, L. Testing of detection tools for AI-generated text. Int J Educ Integr 19, 26 (2023). arXiv:2306.15666
[3]. Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W. & Feizi, S. Can AI-Generated Text be Reliably Detected? (2023) arXiv:2303.11156
[4] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., Finn, C. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. ICML 2023 arXiv:2301.11305
[5] Zhao, X., Ananth, P., Li, L., Wang, Y. X. Provable Robust Watermarking for AI-Generated Text (2023) arXiv:2306.17439
[6] Krishna, K., Song, Y., Karpinska, M., Wieting, J., Iyyer, M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. NeurIPS 2023 arXiv:2303.13408
[7] Hu, X., Chen, P. Y., Ho, T. Y. RADAR: Robust AI-Text Detection via Adversarial Learning. NeurIPS 2023 arXiv:2307.03838
[8] Narayan, S., Cohen, S. B., Lapata, M., Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. EMNLP 2018 arXiv:1808.08745
[9] Fan, A., Lewis, M., Dauphin, Y. Hierarchical Neural Story Generation. Facebook AI Research (2018) arXiv:1805.04833
[10] Islam, N., Sutradhar, D., Noor, H., Raya, J. T., Maisha, M. T., Farid, D. M. Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning (2023) arXiv:2306.01761