THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): https://www.youtube.com/watch?v=_brG2jb7OP8
- Link slides (dạng .pdf đặt trên Github): *https://github.com/HungLTL/CS2205.MAR2024/blob/main/H%C6%B0ng%20L%C3%AA%20Tr%C6%B0%E1%BB%9Dng%20Long%20-%20xCS2205.DeCuong.FinalReport.Template.Slide.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

• Họ và Tên: Lê Trường Long Hưng • MSSV: 230101007 	• Lớp: CS2205.MAR2024 • Tự đánh giá (điểm tổng kết môn): 7/10 • Số buổi vắng: 1 • Số câu hỏi QT cá nhân: 0 • Link Github:

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

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

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

COMPARISON OF NATURAL LANGUAGE PROCESSING TECHNIQUES IN AI-GENERATED AND PARAPHRASED TEXT DETECTION

TÓM TẮT

Sự phát triển các mô hình ngôn ngữ tự nhiên và tích hợp trí tuệ nhân tạo, chẳng hạn như các chatbot ChatGPT hay Google Gemini trong những năm gần đây đã tạo nên một bước tiến lớn trong lĩnh vực công nghệ. Tuy nhiên, đi chung với những thay đổi mới này là sự e ngại về những ảnh hưởng xấu phát sinh từ việc phát triển trí tuệ nhân tạo một cách thiếu kiểm soát, với việc sử dụng các chatbot này cho những hoạt động phi pháp, thiếu trung thực là một nỗi lo âu thường xuyên được nhắc đến. Nhiều phương pháp, công cụ khác nhau đã được đề xuất làm giải pháp để phát hiện nội dung do máy tạo ra [1][2], song một số nghiên cứu đã làm rõ các lỗ hỏng những phương pháp này hay mắc phải [3]. Đề tài này sẽ đi sâu vào tìm hiểu những phương pháp đó, vận dụng vào xây dựng và huấn luyện một loạt mô hình để tiến hành thực nghiệm phát hiện một số văn bản khác nhau do người hay máy tạo ra hoặc diễn giải lại, từ đó xác định những phương pháp, cấu trúc nào hiệu quả nhất trong việc bảo đảm sự minh bạch trong việc sử dụng trí tuệ nhân tạo, để định hướng phát triển những công cụ kiểm tra chính xác hơn với hiệu suất tốt hơn.

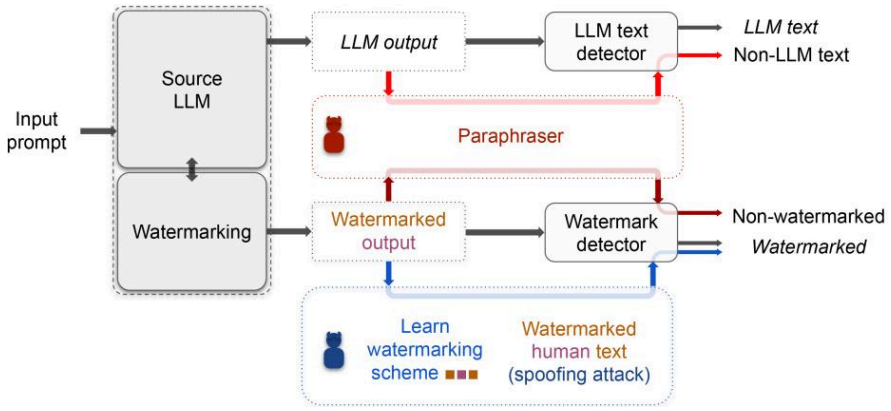
GIỚI THIỆU

Trong những năm gần đây, AI (trí tuệ nhân tạo), cụ thể là xử lý ngôn ngữ tự nhiên, đã dần phát triển thành một trong những chuyên ngành đáng quan tâm trong lĩnh vực công nghệ thông tin. Với sự phát triển của các AI vấn đáp có khả năng tạo nội dung như ChatGPT hay Google Gemini, tiềm năng phát triển của lĩnh vực này trong nhiều

mục đích khác nhau là một thực tế rất khó phủ định được.

Tuy nhiên, việc hiện vẫn chưa có luật lệ hay quy định rõ ràng trong phát triển và sử dụng AI đã tạo nhiều e ngại về mặt đạo đức và tính hợp pháp của AI. Khi được sử dụng vào những mục đích thiếu trung thực, chẳng hạn như tạo bình luận hay tin tức giả, hay văn bản, nghiên cứu gian dối, góp phần vào tệ nạn gian lận học thuật, AI có thể gây thiệt hại uy tín cho những đối tượng, tổ chức bị ảnh hưởng.

Ngay trước cuộc cách mạng AI, trên thị trường đã tồn tại một số công cụ giúp xác thực nội dung [1][2], cụ thể là văn bản, và sự bùng nổ của lĩnh vực AI đã thúc đẩy nhiều công trình nghiên cứu đề xuất nhiều cách khác nhau để phát hiện văn bản không phải do người tạo ra, từ zero-shot [4] đến chữ ký điện tử [5]. Song, một số nghiên cứu gần đây đã vạch ra một số lỗ hổng có thể bị lợi dụng bởi kẻ xấu để lừa cả những công cụ phát hiện nội dung AI tân tiến nhất [3].



Sơ đồ minh họa một số lỗ hổng có thể bị kẻ xấu sử dụng để khiến các công cụ phát hiện nội dung AI trả về kết quả sai, từ đó gây mất uy tín cho người dùng [3].

Trong hàng loạt nhiều biện pháp khác nhau làm nền tảng xây dựng công cụ phát hiện nội dung AI tạo, nghiên cứu này sẽ thực hiện so sánh một số phương pháp xử lý ngôn ngữ tự nhiên thông dụng nhất trong việc phát triển các công cụ này. Thông qua tìm hiểu về cấu trúc và cách vận hành các mô hình vận dụng các phương pháp đó, nghiên cứu mong đợi sẽ tìm ra được phương pháp tốt nhất về mặt chính xác, hiệu suất, và đặc biệt có khả năng cải thiện để khắc phục được các lỗ hổng đã biết để từ đó định hướng phát triển các công cụ xác thực nội dung tốt hơn.

Input: Các văn bản cần được xác thực

Output: Kết quả xác thực văn bản: hoặc người tạo, hoặc AI tạo. Văn bản sẽ được xem là AI tạo nếu AI được sử dụng trong bất cứ giai đoạn nào trong quá trình tạo văn bản, bất kể đó là tạo hay diễn giả lại nội dung.

MỤC TIÊU

- Xây dựng và huấn luyện một số mô hình phát hiện văn bản AI, dựa trên những phương pháp xử lý ngôn ngữ tự nhiên đã biết và có qua chỉnh sửa, cải thiện để khắc phục những khuyết điểm còn sót lại
- So sánh các phương pháp đã và đang được vận dụng vào các công cụ phát hiện AI thông qua kết quả thực nghiệm các mô hình nêu trên; đặc biệt chú ý khả năng phòng vệ trước các lỗ hổng đã biết
- Xác định được các phương pháp và thuật toán nào hữu hiệu nhất và có khả năng khắc phục được khuyết điểm tốt nhất

NỘI DUNG VÀ PHƯƠNG PHÁP

<u>Nội dung:</u>

- Nghiên cứu một số phương pháp xử lý ngôn ngữ tự nhiên khác nhau được áp dụng vào công nghệ phát hiện nội dung, cụ thể là văn bản do AI tạo ra. Một số phương pháp được chú ý bao gồm: zero-shot [4], chữ ký điện tử [5], tham khảo [6], học đối kháng [7], etc.
- Tìm hiểu nguyên nhân gây giảm hiệu suất từ những lỗ hổng thường gặp trong các công cụ xác thực AI, từ đó cải thiện các phương pháp sẵn có, tạo ra một số mô hình nguyên mẫu tích hợp những cải thiện mới.
- Thực hiện thử nghiệm các mô hình trên một số bộ dữ liệu khác nhau. Một số bộ dữ liệu sẽ được tiền xử lý thêm thông qua diễn giải lại để kiểm tra khả năng của mô hình nhận biết AI trong các thay đổi mới.
- So sánh độ chính xác và hiệu suất của các mô hình, từ đó định hướng các phương pháp, thuật toán tốt nhất để phát triển công cụ phát hiện AI.

<u>Phương pháp:</u>

- Chuẩn bị bộ dữ liệu dùng để thử nghiệm trên các mô hình

    - Dataset có sẵn: XSum [8] (tổng hợp hơn 200 nghìn bài viết báo BBC), WritingPrompts [9] (tổng hợp hơn 300 nghìn truyện ngắn viết theo yêu cầu cho trước lấy từ một diễn đàn online)
    - Tạo dataset riêng: tham khảo công trình của Krishna & khác [6], crawl dữ liệu từ các diễn đàn hỏi đáp như r/explainlikeimfive trên Reddit, Quora, Stack Overflow, etc.
    - Bộ dữ liệu thử nghiệm tấn công diễn giải đệ quy [3]: chạy các bộ dữ liệu nêu trên qua một mô hình diễn giải lại liên tục để tạo một bộ dữ liệu mới
- Nghiên cứu kiến trúc và phương thức hoạt động của các mô hình nêu trên, tham khảo theo nghiên cứu của Islam & khác [10]
- Nghiên cứu nguyên nhân diễn giải lại gây giảm độ chính xác, tham khảo theo nghiên cứu của Sadisivan & khác [3]; vận dụng để chỉnh sửa cấu trúc các mô hình nhằm khắc phục các lỗ hổng, tạo các mô hình nguyên mẫu mới
- Huấn luyện các nguyên mẫu mới theo các bộ dữ liệu đã nêu trên, đánh giá và thống kê kết quả dựa trên nhận định rằng đây là một bài toán phân lớp nhị phân (văn bản do người hay AI tạo?)

KẾT QUẢ MONG ĐỢI

- Cải thiện được các phương pháp sẵn có, đặc biệt trước các lỗ hổng thường gặp như diễn giải lại
- Các mô hình tạo ra được cho kết quả có độ chính xác tốt trên các thử nghiệm; tối thiểu có độ chính xác 70% để được xem là phương pháp, thuật toán khả dụng nên được tập trung phát triển và nghiên cứu thêm
- Vạch ra được những phương pháp, thuật toán nào hiệu quả nhất trong phát hiện nội dung AI

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[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