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

Trường ĐH Công nghệ Thông tin
HCMC, Vietnam

Làm gì?

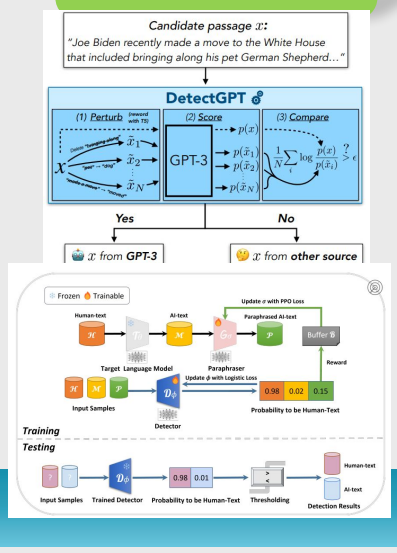
- Xây dựng, huấn luyện mô hình xác thực văn bản dựa trên phương pháp sẵn có, phối hợp cải thiện để khắc phục khiếm khuyết
- So sánh các phương pháp thông qua thực nghiệm trên các mô hình
- Xác định các phương pháp, thuật toán hữu hiệu nhất để định hướng phát triển công cụ phát hiện AI

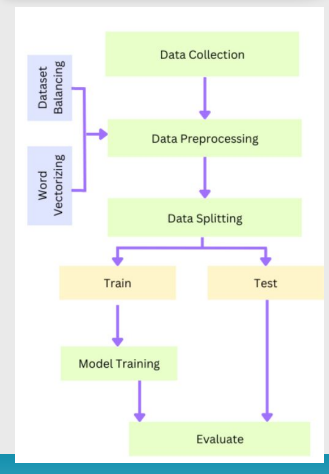
Tại sao?

- AI ngày càng phát triển NHƯNG thiếu quy định/luật lệ -> gây bất an nếu bị kẻ xấu sử dụng
- Nhiều công cụ, mô hình trên thị trường NHƯNG vẫn tồn một số khuyết điểm, lỗ hỏng
- Cần định hướng xem nên tiếp tục nghiên cứu, phát triển các phương pháp nào để tạo các công cụ phát hiện AI mạnh hơn, chính xác hơn

Tổng quan

Nghiên cứu phương pháp → Xây dựng bộ dữ liệu → Xây dựng, chỉnh sửa mô hình → Thực nghiệm, thống kê



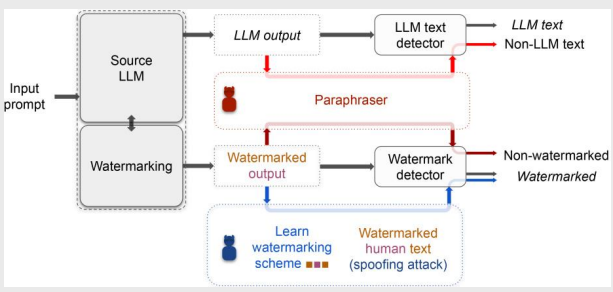


Mô tả

1. Nghiên cứu phương pháp

- Nghiên cứu cấu trúc và cách hoạt động của một số phương pháp, thuật toán xử lý ngôn ngữ tự nhiên khác nhau đã và đang được vận dụng (e.g. zero-shot của DetectGPT, học đối kháng của RADAR chữ ký điện tử, etc.)
- Tìm hiểu và phân tích một số lỗ hỏng, khiếm khuyết đã biết khiến cho các phương pháp trên bị giảm độ chính xác (đặc biệt chú ý: tấn công diễn giải đệ quy)

2. Xây dựng bộ dữ liệu



- Dùng XSun & WritingPrompts là dataset có sẵn
- Crawl dữ liệu từ diễn đàn Q&A (e.g. Reddit, StackOverflow, Quora) làm dataset riêng
- Chạy các dataset qua mô hình diễn giải lại để mô phỏng tấn công diễn giải đệ quy

3. Xây dựng, chỉnh sửa mô hình

- Xây dựng các prototype, sử dụng các phương pháp và thuật toán đã nghiên cứu kết hợp với quy trình chung tiêu chuẩn
- Dựa vào các khiếm khuyết, lỗ hỏng đã phân tích, thay đổi, cập nhật, chỉnh sửa các mô hình sao cho phù hợp
- Huấn luyện mô hình sử dụng các bộ dữ liệu đã chuẩn bị từ trước

4. Thực nghiệm, thống kê

- Thực hiện thử nghiệm trên các mô hình, sử dụng các bộ dữ liệu đã tạo + dự liệu thực tế bên ngoài (vừa người tạo, vừa AI tạo)
- Nhận định đây là bài toán phân loại nhị phân - kết quả khi đưa vào một văn bản là phân loại rằng văn bản đó là người hay AI tạo
- So sánh kết quả của các mô hình; tập trung vào hiệu suất làm việc và độ chính xác

NII
Lê Trường Long Hưng - Trường ĐH Công nghệ Thông tin
TEL : 0961 808 455 Email : hungltl.18@grad.uit.edu.vn