BÁO CÁO ĐỒ ÁN
MÔN: XỬ LÍ NLP - ỨNG DỤNG HƯỚNG NGHIÊN CỨU
HK2-2023

MAY 24

GVHD: Mr. Nguyen Hong Buu Long
Mr. Luong An Vinh
Authored by: 20127674 – Lê Đức Đạt
20127141 – Bùi Tuấn Dũng

TRUONG DAI HOC KHOA HOC TU NHIEN
KHTN
TP. HO CHI MINH

Mục lục

1. **Danh sách thành viên nhóm:**

STT	Họ và tên	MSSV
1	Lê Đức Đạt	20127674
2	Bùi Tuấn Dũng	20127141

2. Tóm tắt nội dung bài báo:

a. **Tiêu đề bài báo: "ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification"** được viết và báo cáo bởi **Taja Kuzman, Igor Mozetič và Nikola Ljubešić**.

** **Về ý nghĩa nhan đề:** có thể xem them trong file **Comment.pdf** trong phần nộp bài.

b. **Tóm tắt:** Tác giả nói về việc sử dụng ChatGPT cho việc phân loại văn bản theo thể loại mà không cần đào tạo. Bài báo so sánh ChatGPT với một mô hình ngôn ngữ XLM-RoBERTa đa ngôn ngữ đã được tinh chỉnh trên bộ dữ liệu được chú thích theo cách thủ công với các thể loại. Kết quả cho thấy ChatGPT hoạt động tốt hơn so với mô hình đã được đào tạo trên tập dữ liệu tiếng Anh và đạt kết quả tương đương trên tập dữ liệu Slovenian. Tuy nhiên, khi ChatGPT được sử dụng trên tiếng Slovenian mà không có dữ liệu đào tạo, kết quả không tốt bằng khi được sử dụng trên tiếng Anh. Bài báo cho rằng kết quả nghiên cứu này mở ra triển vọng cho việc không cần phải tốn công sức để đào tạo mô hình phân loại thể loại văn bản trên các ngôn ngữ nhỏ hơn.

c. **Giới thiệu:** Tác giả nói về việc sử dụng ChatGPT cho việc phân loại văn bản theo thể loại mà không cần đào tạo. Bài báo so sánh hiệu suất zero-

shot của ChatGPT với bộ phân loại X-GENRE dựa trên mô hình XLM-RoBERTa. Kết quả cho thấy ChatGPT có hiệu suất ấn tượng, vượt trội hơn mô hình LLM tinh chỉnh trên bộ dữ liệu tiếng Anh. Đáng chú ý, mặc dù tiếng Slovenia là ngôn ngữ có nguồn lực thấp, hiệu suất của ChatGPT không kém hơn tiếng Anh nếu như việc hỏi đề xuất bằng tiếng Anh thay vì tiếng Slovenia. Điều này đặt ra câu hỏi liệu các chiến dịch gán nhãn lớn đã trở nên thừa và liệu có thể sử dụng ChatGPT để gán nhãn dữ liệu cho mục đích nghiên cứu hay không.

d. **Nhận dạng thể loại tự động:** là một tác vụ trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) liên quan đến việc tự động xác định thể loại của một văn bản. Các kỹ thuật phổ biến để giải quyết tác vụ AGI bao gồm các phương pháp học máy truyền thống và các mô hình Transformer như BERT, GPT và RoBERTa. Để huấn luyện và đánh giá hiệu suất của các mô hình trong tác vụ AGI, các nhà nghiên cứu sử dụng các bộ dữ liệu đã được gán nhãn thể loại văn bản. Phân loại thể loại văn bản là một tác vụ khó, vì văn bản có thể hiển thị dấu hiệu của nhiều lớp hoặc thiếu dấu hiệu của bất kỳ thể loại nào. Các phương pháp phi mạng nơron đã chứng minh là quá phụ thuộc vào tập dữ liệu và không thể tổng quát hóa cho các tập dữ liệu chưa từng thấy. Tuy nhiên, sự xuất hiện của các mô hình Transformer dựa trên mạng nơron sâu đã mang lại bước đột phá. GINCO (Genre Identification and Clustering of Corpora) là một bộ dữ liệu được sử dụng để nghiên cứu và đánh giá hiệu suất của các mô hình trong tác vụ xác định thể loại văn bản tự động.

e. **Tập dữ liệu được gán nhãn thể loại (labels):** (Genre-annotated datasets) là những tập dữ liệu văn bản mà mỗi văn bản đều được gán nhãn với một hoặc nhiều thể loại văn bản. Các tập dữ liệu này thường được sử dụng trong các tác vụ liên quan đến phân loại thể loại tự động

(Automatic Genre Identification) để huấn luyện, tinh chỉnh và đánh giá các mô hình phân loại văn bản theo thể loại. Để đánh giá hiệu suất của các mô hình trên văn bản tiếng Anh và tiếng Slovenia, nghiên cứu sử dụng mẫu ngẫu nhiên từ hai tập dữ liệu được gán nhãn thủ công: EN-GINCO và GINCO. Riêng tập dữ liệu GINCO bao gồm văn bản web tiếng Slovenia từ hai bộ sưu tập văn bản web tiếng Slovenia, slWaC 2.0 và MaCoCu-sl 1.0. Tập dữ liệu EN-GINCO là mẫu văn bản tiếng Anh từ bộ sưu tập văn bản web tiếng Anh enTenTen202.

- **Genre Schema (Lược đồ thể loại):** là một hệ thống phân loại các thể loại văn bản dựa trên các đặc điểm chung, mục đích của tác giả, chức năng và hình thức thông thường của văn bản. Genre schema giúp nhận dạng và phân loại các văn bản theo thể loại, cho phép các công cụ tìm kiếm thông tin và ứng dụng xử lý ngôn ngữ tự nhiên (NLP) có kết quả tìm kiếm chính xác hơn và hiệu quả hơn trong việc xử lý văn bản. Trong thử nghiệm, tác giả bài báo sử dụng lược đồ thể loại X-GENRE, một sự tổng hợp của nhiều lược đồ được áp dụng cho các tập dữ liệu khác nhau như CORE, FTD và GINCO. Mục đích của lược đồ này là thân thiện với người dùng hơn so với các lược đồ cụ thể khác và cho phép kết hợp dữ liệu huấn luyện từ các tập dữ liệu khác nhau, tạo ra mô hình ổn định hơn.

f. Các mô hình:

- **ChatGPT:** ChatGPT là một mô hình ngôn ngữ lớn do OpenAI cung cấp, được tinh chỉnh trên mô hình GPT-3.5 (OpenAI, 2023) và được tối ưu hóa cho đối thoại bằng phương pháp học tăng cường với phản hồi từ con người. Nói cách khác, mô hình được đào tạo để tạo ra các câu trả lời tốt nhất dựa trên đánh giá từ con người. Các tác giả đã sử dụng phiên bản ChatGPT 13/2 và thực hiện các thí nghiệm trong khoảng thời gian từ ngày 24/2 đến 2/3/2023.

- **X-GENRE:** Ta so sánh mô hình ChatGPT với một mô hình dựa trên XLM-RoBERTa được tinh chỉnh trên các tập dữ liệu được chú thích theo thể loại (genre). Mô hình X-GENRE được tinh chỉnh trên khoảng 1.700 trường hợp từ ba bộ dữ liệu với nhãn thể loại được chú thích bằng tay: CORE (Egbert et al., 2015), FTD (Sharoff, 2018) và GINCO (Kuzman et al., 2022b). Mô hình này đạt điểm F1 micro và macro giữa 0,79 và 0,80 trong kịch bản kiểm tra trong tập dữ liệu. Khi so sánh với các mô hình chỉ được đào tạo trên một bộ dữ liệu, kết quả cho thấy mô hình X-GENRE vượt trội hơn. Mô hình này có sẵn miễn phí trên kho lưu trữ Hugging Face.

g. **Lời nhắc ChatGPT và phỏng đoán các thể loại:** Nhóm tác giả sử dụng các lời nhắc trên nền tảng OpenAI và trích xuất thủ công các danh mục và giải thích từ các câu trả lời của nó. Trong lời nhắc, tác giả xác định các tiêu chí chính để xác định thể loại và danh mục mà mô hình có thể lựa chọn. Tác giả cũng yêu cầu mô hình cung cấp lý do giải thích lựa chọn của mình và cung cấp văn bản cần phân loại. Lời nhắc được lặp lại với mỗi văn bản. Tác giả sử dụng các lớp thể loại được sử dụng bởi bộ phân loại X-GENRE để có thể so sánh hai mô hình.

h. **So sánh với mô hình tinh chỉnh (fine-tuned model):** Tác giả so sánh hai mô hình trong ba kịch bản: Trên bộ kiểm tra tiếng Anh (ENGINCO) với lời nhắc bằng tiếng Anh, trên bộ kiểm tra tiếng Slovenia (GINCO) với lời nhắc bằng tiếng Anh và trên bộ kiểm tra tiếng Slovenia (GINCO) với dấu nhắc tiếng Slovenia. Trong hai trường hợp sau, chỉ có ngôn ngữ của lời nhắc là khác, trong khi các trường hợp văn bản để phân loại là giống nhau. Các kết quả được thể hiện trong Bảng 2.

i. **Kết luận – Hướng phát triển mới:** Kết quả cho thấy khi so sánh trên tập dữ liệu mà cả hai mô hình đều chưa được huấn luyện, ChatGPT vượt trội hơn X-GENRE. Điều này gợi ý rằng trong tương lai, chỉ cần ít

gán nhãn thủ công cho các tập dữ liệu kiểm tra. Mặc dù có hạn chế, kết quả này không ảnh hưởng nhiều đến việc sử dụng ChatGPT trong cộng đồng nghiên cứu. Những kết quả này đã đặt ra câu hỏi liệu việc gán nhãn thủ công lớn vẫn còn cần thiết cho các tác vụ phân loại văn bản hay không. Nói tóm lại, bài báo đã đưa ra những kết quả khả quan cho việc sử dụng ChatGPT trong phát hiện thể loại tự động, thậm chí còn vượt trội hơn so với mô hình được huấn luyện trên dữ liệu gán nhãn thủ công.

3. **Chi tiết bài báo và kiến thức liên quan:** Xem trong file **Script.pdf** trong folder nộp bài.
4. **Bố cục – comment bài báo khoa học:** Xem kết hợp giữa hai files: **Script.pdf** và **Comment.pdf** trong folder nộp bài.
5. **Slide thuyết trình bài báo khoa học:** Xem trong file **Slide.pdf** hoặc **Slide.pptx** trong folder nộp bài.