

ChatGPT: Bắt đầu Kết thúc Chú thích Dữ liệu Ngôn ngữ Thủ công?

Trường hợp sử dụng "Nhận dạng thể loại tự động"



GVHD: Mr. Nguyễn Hồng Bửu Long
Mr. Lương An Vinh

** CÁC THÀNH VIÊN TRONG NHÓM:

1. LÊ ĐỨC ĐẠT - 20127674
2. BÙI TUẤN DŨNG - 20127141

1. Tóm tắt sơ lược bài báo:

Tác giả nói về việc sử dụng ChatGPT, một mô hình ngôn ngữ tự động tạo văn bản, cho việc phân loại văn bản theo thể loại (genre) mà không cần đào tạo (zero-shot).

- **Phân loại văn bản theo thể loại (genre)** là quá trình gán nhãn hoặc xác định thể loại của một văn bản, dựa trên nội dung, cấu trúc, ngôn ngữ và đặc điểm khác của văn bản đó. Thể loại có thể bao gồm tiểu thuyết, truyện ngắn, kịch, thơ, báo cáo khoa học, bài báo, văn bản luật pháp và nhiều thể loại khác.

Phân loại văn bản theo thể loại thường được sử dụng trong các ứng dụng như phân loại sách, phân tích dữ liệu văn bản, hệ thống khuyến nghị nội dung, và cải thiện trải nghiệm người dùng trên các nền tảng đọc sách điện tử.

từ hoặc tìm kiếm thông tin.

Các phương pháp phân loại văn bản theo thể loại thường dựa trên các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và học máy, bao gồm cả học có giám sát và học không giám sát. Các thuật toán phổ biến cho việc phân loại văn bản bao gồm Naive Bayes, hỗ trợ máy vectơ (SVM), mạng nơ-ron nhân tạo, và các mô hình dựa trên biểu diễn ngôn ngữ như BERT, GPT và các mô hình tương tự.

- **Zero-shot text classification (phân loại văn bản không cần dữ liệu huấn luyện)** là một phương pháp phân loại văn bản mà không cần dữ liệu huấn luyện có nhãn cho từng thể loại cụ thể. Thay vào đó, mô hình sử dụng kiến thức đã học từ dữ liệu huấn luyện tổng quát (đã được tiền huấn luyện) để phân loại văn bản dựa trên ngữ cảnh và đặc điểm ngôn ngữ của chúng.

Các mô hình zero-shot thường dựa trên các mô hình học sâu như BERT, GPT, RoBERTa, và các kiến trúc tương tự. Các mô hình này được tiền huấn luyện trên một lượng lớn dữ liệu văn bản không nhãn từ nhiều nguồn khác nhau, giúp chúng học được các đặc trưng ngôn ngữ tổng quát và ngữ nghĩa của văn bản.

Khi áp dụng cho bài toán phân loại văn bản, mô hình zero-shot có thể nhận diện và phân loại các thể loại mà nó chưa từng gặp trong quá trình huấn luyện. Điều này làm cho zero-shot text classification đặc biệt hữu ích trong các tình huống có ít hoặc không có dữ liệu huấn luyện có nhãn, hoặc khi cần phân loại văn bản theo các thể loại mới mà mô hình chưa từng gặp.

Tuy nhiên, zero-shot text classification có thể không hoạt động tốt như các phương pháp học có giám sát đối với các bài toán phân loại văn bản có dữ liệu huấn luyện đầy đủ và cụ thể.

Bài báo so sánh ChatGPT với một mô hình ngôn ngữ khác đã được đào tạo, ở đây là 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 để phân loại văn bản theo thể loại trên các tập dữ liệu tiếng Anh và Slovenian. 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, cho thấy giới hạn của ChatGPT trong việc sử dụng trên các ngôn ngữ nhỏ hơn. 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, chẳng hạn như tiếng Slovenian.

- **Nói thêm về mô hình ngôn ngữ XLM-RoBERTa, (Cross-lingual Language Model - RoBERTa)** là một mô hình học sâu dựa trên kiến trúc Transformer, được thiết kế để xử lý các tác vụ xử lý ngôn ngữ tự nhiên (NLP) đa ngôn ngữ. Mô hình này được phát triển bởi Facebook AI và kết hợp hai phương pháp tiên tiến trong lĩnh vực NLP: RoBERTa và XLM.

+ **RoBERTa (Robustly optimized BERT pretraining approach)** là một phiên bản tối ưu hóa của mô hình BERT, được phát triển bởi nhóm nghiên cứu của Facebook AI. RoBERTa tập trung vào việc cải thiện hiệu suất của BERT thông qua việc điều chỉnh quy trình tiền huấn luyện, bao gồm việc tăng kích thước dữ liệu huấn luyện, tăng tốc độ học và loại bỏ phần "Next Sentence Prediction" trong kiến trúc BERT.

+ **XLM (Cross-lingual Language Model)** là một mô hình dựa trên Transformer được thiết kế để xử lý các tác vụ NLP đa ngôn ngữ, phát triển bởi Facebook AI. XLM sử dụng kỹ thuật "unsupervised machine translation" để học biểu diễn đa ngôn ngữ thông qua việc tiền huấn luyện trên dữ liệu song ngữ.

+ **XLM-RoBERTa** kết hợp ưu điểm của cả RoBERTa và XLM để tạo ra một mô hình đa ngôn ngữ hiệu quả. XLM-RoBERTa được tiền huấn luyện trên một bộ dữ liệu đa ngôn ngữ lớn (100 ngôn ngữ) được gọi là CommonCrawl. Kết quả là một mô hình có khả năng xử lý và hiểu nhiều ngôn ngữ khác nhau một cách hiệu quả, đồng thời đạt được hiệu suất cao trong nhiều tác vụ NLP đa ngôn ngữ như phân loại văn bản, dịch máy, và hỏi đáp.

Tóm lại, XLM-RoBERTa là một mô hình NLP đa ngôn ngữ mạnh mẽ, kế thừa các ưu điểm của RoBERTa và XLM, và được huấn luyện trên một lượng lớn dữ liệu đa ngôn ngữ để đạt được hiệu suất cao trong các tác vụ NLP đa ngôn ngữ.

2. Giới thiệu:

ChatGPT đã cho thấy khả năng mạnh mẽ trong hệ thống đối thoại, cung cấp câu trả lời rõ ràng và hữu ích hơn so với con người. Tuy nhiên, hiệu suất của nó trong các tác vụ phân loại văn bản, cụ thể là xác định thể loại tự động, chưa được so sánh với các mô hình ngôn ngữ lớn như XLM-RoBERTa. Mặc dù ChatGPT mới được công bố cách đây vài tháng, nhưng đã có nhiều nghiên cứu phân tích tiềm năng của mô hình này trong các tác vụ NLP. Kết quả cho thấy ChatGPT thường kém hơn so với các mô hình ngôn ngữ lớn tinh chỉnh.

Trong một số trường hợp, ChatGPT đạt kết quả tốt hơn như trong tác vụ phát hiện quan điểm. ChatGPT cũng đạt hiệu suất cạnh tranh trong dịch máy cho các ngôn ngữ có nguồn lực cao, nhưng kém hơn đối với các ngôn ngữ có nguồn lực thấp. Một số nghiên cứu cũng so sánh hiệu suất của mô hình với chú thích thủ công, cho thấy tiềm năng của ChatGPT trong việc phát hiện và giải thích ngôn ngữ thù địch tiềm ẩn.

Bài báo này nghiên cứu hiệu suất của ChatGPT trong xác định thể loại tự động, một tác vụ phân loại văn bản khó cho con người. Nghiên cứu 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.

Bài báo đặt ra câu hỏi liệu các chiến dịch chú thích lớn có trở nên dư thừa hay không, vì hiệu suất của ChatGPT cho thấy nó có thể được sử dụng để chú thích dữ liệu cho mục đích nghiên cứu. Cụ thể, trong phần 2, giới thiệu nhiệm vụ xác định thể loại tự động. Trình bày bộ dữ liệu được chú thích theo thể loại trong Phần 3, trong đó thử nghiệm mô hình ChatGPT và LLM tinh chỉnh, được trình bày trong Phần 4. Thảo luận về các dự đoán của ChatGPT trong Phần 5 và so sánh nó với mô

hình tinh chỉnh trong Phần 6. Cuối cùng, trong Phần 7, kết thúc bài báo thảo luận về những phát hiện chính, đề xuất dự án tiếp theo.

- Trong phần này, có một số thứ ta cần lưu ý:

Thứ nhất, GLUE (General Language Understanding Evaluation) Benchmark là một bộ chuẩn đo lường được thiết kế để đánh giá và so sánh hiệu suất của các mô hình xử lý ngôn ngữ tự nhiên (NLP) trên một loạt các tác vụ hiểu ngôn ngữ chung. GLUE được giới thiệu vào năm 2018 bởi một nhóm các nhà nghiên cứu từ các tổ chức khác nhau như NYU, Facebook AI, và Google AI.

GLUE bao gồm 9 tác vụ NLP phổ biến, dựa trên các bộ dữ liệu khác nhau:

1. CoLA (Corpus of Linguistic Acceptability): Đánh giá mức độ ngữ pháp của các câu.
2. SST-2 (Stanford Sentiment Treebank): Phân loại cảm xúc dựa trên đoạn văn bản.
3. MRPC (Microsoft Research Paraphrase Corpus): Xác định liệu hai câu có phải là đồng nghĩa hay không.
4. STS-B (Semantic Textual Similarity Benchmark): Tính điểm tương đồng ngữ nghĩa giữa hai câu.
5. QQP (Quora Question Pairs): Phát hiện câu hỏi trùng lặp trong Quora.
6. MNLI (Multi-Genre Natural Language Inference): Xác định mối quan hệ logic giữa hai câu (kết luận, mâu thuẫn, trung lập).
7. QNLI (Question-answering Natural Language Inference): Tác vụ hỏi đáp dựa trên tập dữ liệu SQuAD.
8. RTE (Recognizing Textual Entailment): Xác định liệu một câu có kết luận từ câu kia hay không.
9. WNLI (Winograd Natural Language Inference): Đọc hiểu dựa trên các câu Winograd Schema.

Để đánh giá hiệu suất, các mô hình được đưa vào GLUE leaderboard, nơi chúng được xếp hạng dựa trên điểm số trung bình trên tất cả các tác vụ. Điểm số này được gọi là GLUE Score.

GLUE Benchmark đã giúp thúc đẩy sự phát triển và cạnh tranh trong lĩnh vực NLP, khiến các nhà nghiên cứu không ngừng tìm kiếm các mô hình và kỹ thuật mới để cải thiện hiệu suất. Tuy nhiên, do sự tiến bộ nhanh chóng trong NLP, GLUE đã được thay thế bởi SuperGLUE, một bộ chuẩn đo lường mới và khó khăn hơn, để tiếp tục thúc đẩy sự phát triển trong lĩnh vực này.

Ở đây, tác giả đã lấy GLUE benchmark so sánh ChatGPT-với các mô hình ngôn ngữ tinh chỉnh BERT và RoBERTa

Thứ hai, thuật ngữ "SOTA" là viết tắt của "State Of The Art", được sử dụng để chỉ các mô hình hoặc phương pháp hàng đầu trong một lĩnh vực nghiên cứu hoặc công nghệ cụ thể. Trong ngữ cảnh của học máy và xử lý ngôn ngữ tự nhiên (NLP), các mô hình SOTA là những mô hình đạt được hiệu suất cao nhất hoặc xuất sắc nhất trên một tác vụ hoặc bộ chuẩn đo lường cụ thể.

Khi nói đến các mô hình SOTA, chúng ta thường nghĩ đến các mô hình đã được huấn luyện và tinh chỉnh để đạt được kết quả tốt nhất trong một lĩnh vực nghiên cứu hoặc tác vụ cụ thể. Các mô hình này thường được công nhận rộng rãi trong cộng đồng nghiên cứu và phát triển công nghệ, đồng thời được sử dụng làm điểm chuẩn để so sánh với các mô hình và phương pháp mới.

Thứ ba, WMT22 và GINCO datasets là hai bộ dữ liệu được sử dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) cho các tác vụ khác nhau.

1. **WMT22:** WMT là viết tắt của "Workshop on Machine Translation", một hội thảo hàng năm tập trung vào các vấn đề liên quan đến dịch máy. WMT22 đề cập đến kỳ hội thảo năm 2022. Trong khuôn khổ của hội thảo này, các nhà nghiên cứu tham gia các cuộc thi dịch máy để đánh giá và so sánh hiệu suất của các mô hình dịch máy khác nhau. WMT22 dataset là bộ dữ liệu được sử dụng trong cuộc thi dịch máy của hội thảo WMT năm 2022, bao gồm các cặp câu nguồn và mục tiêu trong nhiều ngôn ngữ khác nhau.

2. **GINCO dataset:** 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. GINCO dataset bao gồm các văn bản trong nhiều ngôn ngữ khác nhau, được gán nhãn với các thể loại văn bản tương ứng. Bộ dữ liệu này hỗ trợ việc huấn luyện và kiểm tra các mô hình phân loại văn bản đa ngôn ngữ để xác định thể loại của các văn bản mới.

Cả hai bộ dữ liệu này đều quan trọng và hữu ích trong việc nâng cao hiệu suất của các mô hình NLP và giúp cải thiện các ứng dụng thực tế của công nghệ này.

3. Nhận dạng thể loại tự động:

Automatic Genre Identification (AGI) là gì?

Đó 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. Thể loại văn bản là một phân loại dựa trên nội dung, chức năng, cấu trúc và/hoặc phong cách của văn bản. Ví dụ, một số thể loại phổ biến bao gồm tin tức, tiểu luận, thơ, tiểu thuyết, hướng dẫn và báo cáo khoa học.

Tác vụ AGI đòi hỏi phát triển các mô hình và thuật toán có khả năng phân tích văn bản và xác định thể loại của chúng một cách chính xác. Điều này đặc biệt quan trọng trong các ứng dụng như lọc thông tin, gợi ý nội dung, tổ chức và phân loại văn bản lớn.

Các kỹ thuật phổ biến để giải quyết tác vụ AGI bao gồm các phương pháp học máy truyền thống như Naive Bayes, học máy vector hỗ trợ (SVM) và cây quyết định, cũng như 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, chẳng hạn như bộ dữ liệu GINCO.

Khác biệt so với phát hiện chủ đề, thể loại không thể phân loại chỉ dựa trên thông tin từ vệt. Các mô hình cần tìm ra mẫu mã cao hơn trong văn bản, thường dựa trên đặc trưng văn bản hoặc cú pháp không liên quan đến chủ đề. Phân loại thể loại văn bản là một tác vụ khó, vì văn bản có thể là ví dụ đặc trưng hơn hoặc ít đặc trưng hơn cho thể loại, hiển thị dấu hiệu của nhiều lớp hoặc thiếu dấu hiệu của bất kỳ thể loại nào.

Các phương pháp phi mạng nơ-ron 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 nơ-ron sâu đã mang lại bước đột phá. Các mô hình giống BERT được tiền huấn luyện trên lượng lớn văn bản và tinh chỉnh cho tác vụ xác định thể loại, cho thấy chúng có thể xác định thể loại trên các tập dữ liệu và ngôn ngữ chưa từng thấy, thậm chí chỉ cần huấn luyện trên

1.000 văn bản. Tuy nhiên, các mô hình vẫn cần văn bản được gán nhãn thủ công, một công việc tốn thời gian và chi phí.

Labels	EN-GINCO	GINCO	X-GENRE training
Information/Explanation	25%	24%	17%
Promotion	22%	17%	16%
Opinion/Argumentation	18%	11%	14%
News	18%	29%	19%
Other	6%	7%	4%
Forum	6%	5%	8%
Instruction	5%	5%	12%
Legal	0%	1%	4%
Prose/Lyrical	0%	1%	6%

Bảng 1: Phân phối nhãn trong các tập dữ liệu kiểm tra tiếng Anh (EN-GINCO), tập dữ liệu kiểm tra tiếng Slovenia (GINCO) và tập dữ liệu được sử dụng để huấn luyện bộ phân loại X-GENRE. Những con số này giúp hiểu rõ hơn về cấu trúc và đặc điểm của từng tập dữ liệu và có thể được sử dụng để đánh giá hiệu suất của mô hình trên các tác vụ phân loại thể loại.

4. Tập dữ liệu được gán nhãn thể loại:

Genre-annotated datasets (tập dữ liệu được gán nhãn thể loại) là những tập dữ liệu văn bản mà mỗi văn bản đều được gán nhãn với một hoặc nhiều thể loại văn bản. Thể loại văn bản thường được xác định dựa trên mục đích của tác giả, chức năng và hình thức thông thường của văn bản, chẳng hạn như tin tức, pháp lý, quảng cáo, ý kiến, hướng dẫn, văn chương, và nhiều hơn nữa.

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. Việc gán nhãn thể loại giúp các mô hình học cách nhận diện các đặc trưng và mẫu mã riêng biệt của từng thể loại văn bản, từ đó cải thiện hiệu suất trong các tác vụ phân loại và ứng dụng thực tế.

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. Các mô hình được kiểm tra trên 100 mẫu từ mỗi tập dữ liệu. Vì trình phân loại X-GENRE được huấn luyện trên phần tập huấn luyện của GINCO, các mẫu kiểm tra được lấy từ phần tập kiểm tra. Trong khi đó, tập dữ liệu EN-GINCO không phải là một phần của dữ liệu huấn luyện để tinh chỉnh mô hình X-GENRE.

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. Tập dữ liệu EN-GINCO chưa được công bố trực tuyến, vì vậy, không thể ChatGPT đã thấy nó trước đây. Nó đã được công bố vào năm 2021.

Các mẫu được gán nhãn thủ công với 24 thể loại và hai người gán nhãn có nền tảng ngôn ngữ học tham gia. Độ đồng nhất giữa các người gán nhãn đạt chỉ số Krippendorff's alpha là 0,71, cho thấy độ tin cậy tương đối của dữ liệu được gán nhãn và khó khăn trong việc gán nhãn thủ công cho tác vụ này.

Ở đây, ta sẽ nói thêm về **slWaC 2.0**, **MaCoCu-sl 1.0** và **enTenTen202**. Nó bao gồm hàng triệu từ và được sử dụng trong nghiên cứu và phát triển ứng dụng xử lý ngôn ngữ tự nhiên (NLP) cho tiếng Slovenia và tiếng Anh.

a. 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 các nghiên cứu và ứng dụng NLP, genre schema thường được sử dụng để gán nhãn thể loại cho các văn bản trong các bộ dữ liệu, giúp huấn luyện và đánh giá các mô hình phân loại văn bản tự động theo thể loại. Genre schema có thể bao gồm nhiều thể loại khác nhau như tin tức, pháp lý, quảng cáo, ý kiến, hướng dẫn và nhiều thể loại khác.

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. Tác giả đã ánh xạ các nhãn gốc trong cả hai tập dữ liệu kiểm tra, được gán nhãn với một lược đồ có độ chi tiết hơn, sang lược đồ X-GENRE. Lược đồ cuối cùng bao gồm 9 nhãn như hình 1 ở trên.

Labels	EN-GINCO	GINCO	X-GENRE training
Information/Explanation	25%	24%	17%
Promotion	22%	17%	16%
Opinion/Argumentation	18%	11%	14%
News	18%	29%	19%
Other	6%	7%	4%
Forum	6%	5%	8%
Instruction	5%	5%	12%
Legal	0%	1%	4%

5. Các mô hình:

A. 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.

B. 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.

6. Lời nhắc ChatGPT và phỏng đoán các thể loại:

Để phân loại văn bản theo thể loại với ChatGPT, 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.

Một ví dụ về lời nhắc tiếng Anh và câu trả lời của ChatGPT gồm nhãn và giải thích. Trong hầu hết trường hợp, câu trả lời của ChatGPT tuân theo dạng của ví dụ, bao gồm nhãn trong một dòng và giải thích trong đoạn văn tiếp theo. Tuy nhiên, trong một số trường hợp, nhãn là một phần của giải thích. Lời nhắc tiếng Slovenia là bản dịch của lời nhắc tiếng Anh, trong đó cả tên của các lớp cũng được dịch sang tiếng Slovenia. Khi sử dụng lời nhắc tiếng Slovenia, câu trả lời của ChatGPT cũng bằng tiếng Slovenia.


```
Opinion/Argumentation, Instruction, Information/Explanation, Prose/Lyrical, Forum, Other.
The text to classify: Shower pods install in no time. . .
<p> 1. Prepare the floor with the waste and the water supply pipes.
<p> 2. Attach shower equipment to the shower pod shell running flexible tails (H&C or just C) down back.
<p> 3. Move unit into position connecting water supplies on the way and the waste outlet trap.
<p> 4. Having secured the shower pod shell to the building structure doors may now be fitted.
Example of the ChatGPT's answer:
Genre: Instruction
Explanation: The text provides step-by-step instructions for installing shower pods.
It is structured in a sequential manner, detailing the process of preparing the floor,
attaching shower equipment, moving the unit into position, and securing the shower pod shell.
The language used is imperative, giving clear and concise commands to the reader.
There is no attempt to persuade, promote, or argue for any particular point of view.
Instead, the purpose of the text is to instruct the reader on how to install shower pods.
Therefore, the text can be classified as an instruction manual or guide.
```

Ví dụ của tác giả

Trong một số trường hợp, đầu ra của ChatGPT gồm hai lớp thay vì một. Nếu một trong số các lớp phù hợp với nhãn thực tế, tác giả sử dụng lớp này để đánh giá. Ngoài ra, trong một số trường hợp, ChatGPT trả lời bằng nhãn mới không thuộc tập hợp lớp thể loại, chẳng hạn như Đánh giá và Chia buồn. tác giả đã thử nghiệm yêu cầu nó sửa đổi quyết định và chọn nhãn từ tập hợp đóng.

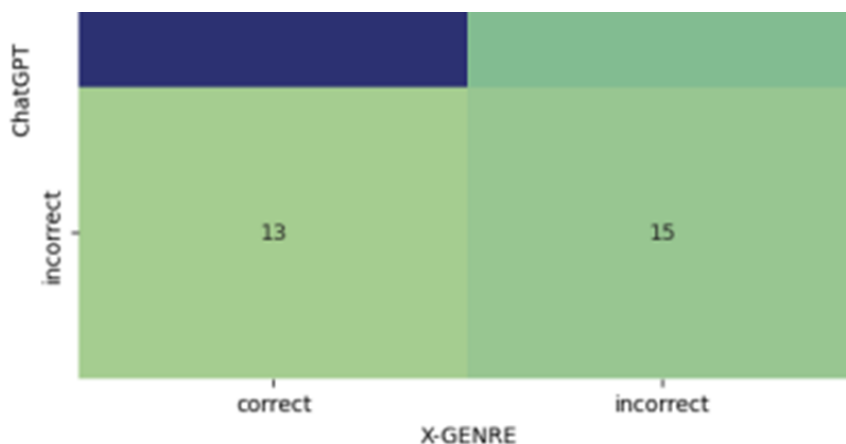
7. So sánh với mô hình tinh chỉnh <fine-tuned model>:

Tác giả so sánh hai mô hình trong ba kịch bản: Trên bộ kiểm tra tiếng Anh (EN-GINCO) 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.

Test set	Prompt	Model	Micro F1	Macro F1	Accuracy
EN-GINCO	EN	ChatGPT	0.74	0.66	0.72
		X-GENRE	0.67	0.61	0.67
GINCO	EN	ChatGPT	0.75	0.64	0.75
		X-GENRE	0.91	0.91	0.91
GINCO	SL	ChatGPT	0.68	0.56	0.68
		X-GENRE	0.91	0.91	0.91

Bảng 2: So sánh ChatGPT và mô hình X-GENRE tinh chỉnh trên hai bộ thử nghiệm và với lời nhắc trong hai bộ ngôn ngữ.





Mô hình 1: So sánh sự khác biệt về dự đoán đúng và sai giữa

ChatGPT và X-GENRE.

8. Kết luận - Hướng phát triển mới:

Đến nay, học máy có giám sát là phương pháp đạt kết quả cao nhất trong hầu hết các tác vụ Xử lý ngôn ngữ tự nhiên (NLP). Tuy nhiên, việc gán nhãn thủ công tốn nhiều thời gian, công sức và tiền bạc, và dẫn đến sự không đồng nhất giữa người gán nhãn. Kết quả từ một số nghiên cứu trước đó và thí nghiệm của Tác giả cho thấy ChatGPT có thể thay thế các chiến dịch gán nhãn thủ công lớn, thay đổi quy trình làm việc của các nhà nghiên cứu NLP. Trong bài báo này, Tác giả phân tích hiệu suất của ChatGPT trong phát hiện thể loại tự động mà không cần học có giám sát, và so sánh nó với bộ phân loại X-GENRE dựa trên hơn 1.700 văn bản đã được gán nhãn thủ công.

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 kết quả khả quan trên tập dữ liệu tiếng Anh, Tác giả dự kiến ChatGPT sẽ không hoạt động tốt trên các ngôn ngữ khác, đặc biệt là tiếng Slovenia. Tuy nhiên, kết quả cho thấy hiệu suất ổn định bất kể ngôn ngữ.

Mặc dù có hạn chế, kết quả này không ảnh hưởng nhiều đến việc sử dụng ChatGPT trong cộng đồng nghiên cứu. Tuy nhiên, sự khác biệt trong kết quả do lời nhắc cho thấy cần nghiên cứu chi tiết hơn về lời nhắc trong tác vụ này. Các hướng nghiên cứu tiếp theo bao gồm thử nghiệm các kỹ thuật nhắc nâng cao hơn, so sánh hiệu suất của ChatGPT dựa trên các lược đồ thể loại khác nhau, mở rộng so sánh với các tập dữ liệu và ngôn ngữ lớn hơn, và khám phá khả năng của ChatGPT cho phân loại đa nhãn...

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. Trong tương lai, Tác giả dự định

nghiên cứu ảnh hưởng của kích thước tập dữ liệu gán nhãn thủ công đối với hiệu suất của các mô hình được tinh chỉnh, so sánh với hiệu suất zero-shot của ChatGPT. Bên cạnh đó, do việc sử dụng ChatGPT và các mô hình tương tự để gán nhãn các bộ sưu tập văn bản lớn đòi hỏi nhiều chi phí tính toán, Tác giả có thể thử nghiệm việc sử dụng các mô hình ngôn ngữ lớn để gán nhãn dữ liệu huấn luyện thay vì gán nhãn thủ công, và sau đó tinh chỉnh các mô hình cơ sở để gán nhãn lượng dữ liệu lớn.

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. Điều này mở ra hướng nghiên cứu mới về việc tối ưu hóa lời nhắc, sử dụng các kỹ thuật nhắc nâng cao hơn, mở rộng so sánh với các tập dữ liệu và ngôn ngữ khác, và đánh giá liệu việc gán nhãn thủ công lớn vẫn còn cần thiết hay không trong tương lai.