Hiệu suất và hiệu của của những mô hình ngôn ngữ lớn LLMs trong gán nhãn dữ liệu kinh tế

Nguyễn Duy Hoàng - 23020368
Nguyễn Trung Hiếu - 23020366
Dương Lý Khánh Hạ - 23020362

Ngày 6 tháng 5 năm 2025

Mục lục

Tóm tắt nội dung

Nghiên cứu về hiệu suất và hiệu quả của những mô hình ngôn ngữ lớn LMMs trong gán nhãn dữ liệu kinh tế

1 Giới thiệu

1.1 Bối cảnh nghiên cứu

Các mô hình ngôn ngữ lớn LLMs đã không còn quá xa lạ với mọi người trong thời đại 4.0, chúng ta biết rằng những mô hình này đã đạt được nhiều thành tựu to lớn trong nhiều lĩnh vực khác nhau. Tuy nhiên, trong lĩnh vực kinh tế, việc áp dụng những mô hình này để gán nhãn dữ liệu vẫn còn khá mới mẻ và chưa được nghiên cứu nhiều. Điều đó làm cho mọi người quan ngại về hiệu suất của những mô hình này trong công việc gán nhán dữ liệu kinh tế.

1.2 Mục tiêu nghiên cứu

- Đánh giá hiệu quả của những mô hình ngôn ngữ lớn trong việc gán nhãn trên những tập dữ liệu tổng quát và chuyên biệt.
- So sánh độ chính xác, tốc độ và chi phí giữa phương pháp truyền thống và phương pháp sử dụng LLMs.
- Phân tích chi tiết kết quả của các phương pháp, mô hình được đưa vào sử dụng
- Tổng hợp kết quả để đưa ra những nhận xét và đề xuất cho việc sử dụng mô hình ngôn ngữ lớn LLMs trong việc gán nhãn dữ liệu kinh tế.
- Đề xuất hướng phát triển cho việc nghiên cứu trong tương lai.

1.3 Phạm vi nghiên cứu

• Những tập dữ liệu được đưa vào nghiên cứu:
  - REFinD: Tập dữ liệu chuyên biệt về kinh tế
  - .....
• Các mô hình ngôn ngữ lớn được sử dụng trong nghiên cứu :
  - chatGP, GPT-3, BERT, RoBERTa, T5, ...
• Các tiêu chí đánh giá hiệu suất

2 Tổng quan về LMMs và gán nhãn dữ liệu kinh tế

2.1 Mô hình ngôn ngữ lớn (LMMs) là gì?

Mô hình ngôn ngữ lớn (Large Language Models - LLMs) là các mô hình học sâu (deep learning) được huấn luyện trên tập dữ liệu văn bản khổng lồ nhằm hiểu, tạo, và thao tác với ngôn ngữ tự nhiên. Chúng được xây dựng dựa trên kiến trúc Transformer, nổi bật với cơ chế self-attention giúp mô hình hiểu được ngữ cảnh của từ trong câu và sinh văn bản có ý nghĩa.

Các mô hình LLMs phổ biến: GPT-3, BERT, RoBERTa, T5, chatGP, ...

Nguyên lý hoạt động: Mô hình học máy hoạt động theo quy trình cơ bản như sau

1.**tiền huấn luyện**: Mô hình được huấn luyện trên tập dữ liệu lớn để hiểu ngôn ngữ tự nhiên

2.**tinh chỉnh**: Mô hình được tinh chỉnh trên tập dữ liệu nhỏ để thực hiện các nhiệm vụ cụ thể

3.**dự đoán**: Mô hình được sử dụng để dự đoán, phân loại, sinh văn bản, ...

2.2 Gán nhãn dữ liệu kinh tế

2.2.1 Gán nhãn dữ liệu là gì?

Gán nhãn dữ liệu (Data Labeling) là quá trình gán các nhãn hoặc thông tin mô tả cho dữ liệu để giúp mô hình học máy (Machine Learning - ML) học cách nhận diện các mẫu (patterns). Dữ liệu đã được gán nhãn được gọi là dữ liệu có giám sát (labeled data) và thường được sử dụng trong các bài toán học có giám sát (Supervised Learning).

Ví dụ:

- Trong xử lý ngôn ngữ tự nhiên (NLP), gán nhãn dữ liệu có thể là việc xác định câu nào mang ý nghĩa tích cực hay tiêu cực (phân tích cảm xúc).

- Trong nhận dạng hình ảnh, gán nhãn có thể là đánh dấu hình ảnh có chứa một con chó, mèo hoặc xe hơi.

2.2.2 Vai trò của gán nhãn dữ liệu trong kinh tế

1.Cải thiện độ chính xác trong phân tích kinh tế

2.Hỗ trợ mô hình hóa rủi ro và dự báo thị trường

3.Cải thiện hiệu suất dự báo kinh tế bằng AI

4.Hỗ trợ hoạch định chính sách kinh tế

5.Cá nhân hóa dịch vụ tài chính

2.2.3 Các phương pháp gán nhãn dữ liệu kinh tế truyền thống

1.Gán nhãn thủ công(manutual labeling)

Các chuyên gia hoặc nhân viên nhập liệu trực tiếp xem xét dữ liệu và gán nhãn tương ứng.

Ưu điểm: Chính xác, đáng tin cậy

Nhược điểm: Tốn kém, tốn thời gian, không thể áp dụng cho dữ liệu lớn

2.Gán nhãn bằng crowdsourcing (Crowdsourced Labeling)

Nhiều người tham gia (crowd workers) cùng thực hiện gán nhãn thông qua các nền tảng như Amazon Mechanical Turk, Appen, hoặc Labelbox.

Ưu điểm: Chi phí thấp, nhanh chóng

Nhược điểm: Chất lượng không đảm bảo, cần kiểm soát chất lượng

3. Học bám sát (Semi-supervised Learning - SSL)

Kết hợp một lượng nhỏ dữ liệu có gán nhãn với một lượng lớn dữ liệu chưa gán nhãn. Sử dụng mô hình AI để dự đoán nhãn cho dữ liệu chưa gán nhãn, sau đó xác minh và tinh chỉnh nhãn.

Ưu điểm : Tiết kiệm chi phí, có thể sử dụng lượng lớn dữ liệu chưa gán nhãn để cải thiện độ chính xác

Nhược điểm : Mô hình ban đầu có thể tạo ra nhãn sai nếu không được tinh chỉnh đúng cách, cần một lượng dữ liệu có nhãn chất lượng cao để huấn luyện mô hình

2.2.4 Phương pháp gán nhãn dữ liệu kinh tế sử dụng LLMs

1. Gán nhãn bán tự động (Human-in-the-loop)

- LLMs gợi ý nhãn và con người xác nhận hoặc chỉnh sửa.
- Giúp tiết kiệm thời gian so với gán nhãn thủ công.

2. Gán nhãn tự động (Auto-labeling)

- LLMs tự động gán nhãn dựa trên ngữ cảnh và dữ liệu huấn luyện.
- Phù hợp với tập dữ liệu lớn nhưng cần kiểm tra chất lượng.

3. Học có giám sát kết hợp LLMs (Supervised Learning with LLMs)

- Dùng LLMs để gán nhãn dữ liệu chưa có nhãn, sau đó sử dụng tập dữ liệu này để huấn luyện mô hình học có giám sát.
- Giúp cải thiện độ chính xác của mô hình AI dự báo kinh tế

4. Gán nhãn dữ liệu bằng phương pháp few-shot learning

- Cung cấp một số ví dụ mẫu, sau đó LLMs suy luận và gán nhãn dữ liệu mới.
- Giảm phụ thuộc vào tập dữ liệu có nhãn lớn.

5. Gán nhãn dữ liệu bằng phương pháp zero shot learning

- LLMs gán nhãn mà không cần dữ liệu huấn luyện trước, chỉ dựa vào kiến thức sẵn có.
- Tiết kiệm chi phí nhưng có thể không chính xác bằng các phương pháp khác.

3 Phương pháp nghiên cứu

3.1 Dữ liệu

Tập dữ liệu kinh tế REFinD

- Nguồn gốc và lý do chọn REFINDREFIND: Tập dữ liệu này được trích xuất từ các báo cáo hàng quý và hàng năm của các công ty giao dịch công khai. REFinD là tập dữ liệu lớn nhất hiện có dành cho nhiệm vụ trích xuất quan hệ trong lĩnh vực tài chính. Đây cũng là tập dữ liệu tài chính duy nhất có thể tiếp cận với các gán nhãn từ cả chuyên gia và lao động đám đông.
- Kích thước và các loại thực thể: REFinD bao gồm 28.676 mẫu dữ liệu, với 22 loại quan hệ trên 8 cặp thực thể khác nhau. Một tập dữ liệu tài chính khác là FinRED (Sharma et al. 2022) có quy mô nhỏ hơn đáng kể (6.767 mẫu và 29 loại quan hệ), nhưng không công khai dữ liệu gán nhãn từ từng lao động đám đông riêng lẻ. Thí nghiệm này sử dụng 3.598 mẫu từ tập kiểm tra (test set) của REFinD, do chi phí sử dụng LLM cao.

3.2 Models

Thí nghiệm này đã sử dụng ba mô hình ngôn ngữ lớn (LLM) gồm GPT-4, PaLM 2, và MPT Instruct. Các mô hình này được chọn dựa trên hiệu suất xuất sắc của chúng trong các bảng xếp hạng chuẩn hóa (benchmark leaderboards), khả năng truy cập, khả dụng của API và giấy phép sử dụng linh hoạt.
Ba mô hình với kích thước khác nhau:

- Chat GPT-4 có khoảng 1700 tỉ tham số
- PaLM có khoảng có 340 tỉ tham số
- MPT Instruct với khoảng 7 tỉ tham số

Những thiết lập nhỏ để thí nghiệm các mô hình:

- với mỗi mô hình ta sẽ thực hiện thí nghiệm với hai mức nhiệt độ khác nhau là 0.2 và 0.7 nhằm kiểm kiểm tra tác động của mức độ ngâu nhiên lên hiệu suất mô hình
- Mô hình không đặt được random seed cho nên sẽ cho ra những kết quả khác nhau

3.3 Prompts và các thước đo đánh giá hiệu suất của mô hình

3.3.1 Prompts

Chất lượng của lời nhắc hướng dẫn mô hình ngôn ngữ lớn (LLM) ảnh hưởng đáng kể đến hiệu suất của chúng, tương tự như cách soạn hướng dẫn cho những người lao động đám đông. Mỗi lời nhắc đầu vào bao gồm:

- Mô tả bằng văn bản nhiệm vụ.
- Một câu có các thực thể được đánh dấu.
- Danh sách các tùy chọn quan hệ (nhãn) được đánh số, cụ thể cho từng cặp thực thể.

Thí nghiệm này đã sử dụng 6 loại lời nhắc khác nhau và chia thành 3 nhóm:

- Nhóm 1: Zero-shot:
  - Lời nhắc đơn giản – Mô tả nhiệm vụ ngắn gọn bằng tiếng Anh đơn giản.
  - Lời nhắc hướng dẫn đầy đủ – Phiên bản mở rộng với mô tả chi tiết hơn, dựa trên hướng dẫn gán nhãn của tập dữ liệu REFinD được cung cấp cho người lao động trên nền tảng Amazon Mechanical Turk (MTurk).
- Nhóm 2: Few-shot:
  - 1-shot – Dựa trên lời nhắc hướng dẫn đầy đủ và thêm một ví dụ về nhiệm vụ, phù hợp với loại cặp thực thể cụ thể.
  - 5-shot – Giống như 1-shot, nhưng bao gồm năm ví dụ thay vì một.
- Nhóm 3: Few-shot Chain-of-Thought (CoT):
  - 1-shot CoT – Kết hợp giữa mô tả nhiệm vụ, ví dụ và cả suy luận giải thích lý do cho từng quyết định.
  - 5-shot CoT – Phiên bản mở rộng của 1-shot CoT, sử dụng năm ví dụ thay vì một.

3.3.2 Thước đo đánh giá hiệu suất của mô hình

1.Mô hình sẽ được đánh giá bằng cách so sánh với các nhãn được gán bởi chuyên gia bằng 2 thước đo chính:

- Độ chính xác(acurrancy): là tỉ lệ giữa số lượng dự đoán đúng và tổng số dự đoán.
- F_1- score: là một thước đo tổng hợp của độ chính xác và độ phủ của mô hình, được tính bằng công thức:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

2.Chỉ số tin cậy(Reliability Index - LLM-Rellndex):
Đây là chỉ số để đánh gía mức độ tin cậy của mô hình.Để tổng hợp nhãn cuối cùng cho mỗi mẫu dữ liệu từ nhiều mô hình gán nhãn khác nhau, cách tiếp cận đơn giản nhất là đếm số phiếu bầu cho mỗi nhãn từ K người gán nhãn. Tuy nhiên, phương pháp này có nhược điểm: nếu mỗi người gán nhãn chọn một nhãn khác nhau, thì việc chọn nhãn cuối cùng sẽ mang tính ngẫu nhiên. Do đó, chúng tôi cải tiến cách tính điểm số phiếu bầu bằng cách tính đến mức độ tương đồng giữa các nhãn theo đánh giá của chuyên gia bằng chỉ số Rellndex.

Chỉ số này giúp xác định nhãn nào là đáng tin cậy nhất trong mỗi trường hợp và cũng có thể dùng để lọc ra những trường hợp cần sự can thiệp của chuyên gia.

Cách tính của Rellndex:

- Đối với mỗi nhãn dán l, chúng tôi tính điểm phiếu bầu có trọng số như sau

$$\text{vote}(i, j) = \text{sim}(a_i, l)$$

- Sau đó tính điểm tin cậy của phiếu bầu bằng cách lấy trung bình điểm phiếu bầu

$$\text{confid}(l) = \frac{1}{K} * \sum_{1}^{K} \text{vote}(i, l)$$

- Cuối cùng chỉ số Rellndex được xác định là độ tin cậy cao nhất trong số các nhãn dán được đề xuất

$$\text{Rellndex} = \max_{l} \text{confid}(l)$$

3.Chi phí:

Đối với các mô hình được truy cập qua API, chi phí trên mỗi mẫu phụ thuộc vào số lượng tokens (đối với GPT-4) hoặc ký tự (đối với PaLM 2) trong cả lời nhắc đầu vào và đầu ra được tạo ra. Do đó, chi phí gán nhãn được tính bằng cách nhân số lượng trung bình của tokens/ký tự trong lời nhắc và đầu ra với số lượng mẫu, sau đó nhân với giá mỗi token/ký tự.

$$\text{Chi phí} = \frac{\text{Số token/ký tự đầu vào, đầu ra}}{\text{Giá mỗi token/ký tự}} \times \text{Số mẫu}$$

4 Kết quả và thảo luận

4.1 Kết quả

1. Độ chính xác của các mô hình với các prompt và chỉ số khác nhau

Micro-Averaged F1 Score/ Accuracy(%)								
Annotator	**Type**	**Temperature Setting**	**Zero-Shot Prompt**		**Few-Shot Prompt**		**Few-Shot CoT Prompt**	
			simple prompt	**full instruction**	**1-shot**	**5-shot**	**1-shot CoT**	**5-shot CoT**
LLM	GPT-4	0.2	67.4/63.4	68.5/64.6	65.0/60.1	67.6/63.8	64.5/58.4	68.4/**65.4**
	GPT-4	0.7	**67.6/63.6**	68.4/64.6	65.0/60.0	67.7/63.9	64.6/**58.4**	68.4/**65.4**
	PaLM 2	0.2	62.3/53.9	62.2/53.8	66.4/60.1	66.0/59.2	64.7/55.9	65.6/57.2
	PaLM 2	0.7	64.5/56.0	64.4/56.0	**67.3/60.9**	**68.7**/63.8	64.9/57.4	65.9/59.2
	MPT Instruct	0.2	20.0/21.9	31.1/27.6	18.6/18.0	42.5/36.7	20.1/18.5	45.2/36.1
	MPT Instruct	0.7	20.8/24.7	24.8/27.3	22.7/24.2	30.5/31.1	22.2/23.2	33.9/30.8
	Ensemble (All LLMs)	0.2	65.2/60.1	66.0/60.7	63.9/58.1	68.1/63.3	63.3/56.4	**68.8**/63.8
	Ensemble (GPT-4 w Palm 2)	0.2	67.2/63.2	**68.6/64.7**	65.0/60.1	67.8/**64.0**	64.3/58.1	68.2/65.2
	Ensemble (GPT-4 w MPT Instruct)	0.2	67.2/63.2	**68.6/64.7**	65.0/60.1	67.8/**64.0**	64.3/58.1	68.2/65.2
	Ensemble (Palm 2 w MPT Instruct)	0.2	62.6/54.3	61.9/53.6	66.7/60.5	66.1/59.4	64.5/55.7	65.4/56.9
Human	Mturk Annotators	-	-	38.6/40.7	-	-	-	-

Table 1: Annotator performance in terms of micro-averaged F1-Score and accuracy against expert assigned labels.

- Mô hình chat GPT-4 nhìn chung hoạt động khá ổn trên hầu hết các tham số và prompt được cung cấp
- Mức độ ngẫu nhiên (temperature) có ảnh hưởng rất nhỏ đến khả năng dự đoán của GPT-4
- Mô hình GPT-4 hoạt động cân bằng và chính xác với prompt 5-shot CoT, với chỉ số cân bằng / độ chính xác = 68.4/65.4
- Mô hình MPT Instruct có hiệu suất thấp hơn đáng kể so với GPT-4 và PaLM 2, đặc biệt là với những prompt đơn giản mô hình gần như không thể dự đoán chính xác được nhãn nào cho các mẫu dữ liệu
- MPT Instruct có độ chính xác cao hơn với các prompt 5-shot CoT và 1-shot CoT, nhưng vẫn thấp hơn đáng kể so với GPT-4 và PaLM 2
- Mô hình PaLM 2 hoạt động tốt với các prompt đơn giản, nhưng không thể đạt được độ chính xác cao như GPT-4 với các prompt phức tạp hơn.