JDI
JOURNAL OF DEVELOPMENT AND INTEGRATION

TẠP CHÍ PHÁT TRIỂN VÀ HỘI NHẬP
Số 88 (2026) 56-63 | jdi.uef.edu.vn

Phát hiện sự phàn nàn trên website thương mại điện tử với mô hình ngôn ngữ lớn

Thái Kim Phụng[1], Trần Sơn Nam[2,*], Đỗ Thị Ngọc Thuỳ[3]
[1]Đại học Kinh tế Thành phố Hồ Chí Minh, Việt Nam
[2]Đại học Kinh tế Thành phố Hồ Chí Minh - Phân hiệu Vĩnh Long, Việt Nam
[3]Trung tâm Thông tin tín dụng Quốc gia Việt Nam

TỪ KHÓA

Bình luận trực tuyến, mô hình ngôn ngữ lớn, phát hiện phàn nàn, thương mại điện tử, xây dựng câu lệnh.

TÓM TẮT

Trong bối cảnh thương mại điện tử phát triển mạnh mẽ, các bình luận của khách hàng trên nền tảng trực tuyến, đặc biệt là những lời phàn nàn, mang giá trị quan trọng đối với doanh nghiệp. Hướng tiếp cận học máy truyền thống mặc dù phù phổ biến trong phân loại văn bản nhưng tồn tại một số vấn đề, gây khó khăn khi ứng dụng trong các bài toán bao gồm phát hiện phàn nàn. Do đó, nghiên cứu này được thực hiện nhằm đề xuất một hướng tiếp cận mới, ứng dụng mô hình Ngôn ngữ Lớn (Large Language Model - LLM) vào nhiệm vụ phát hiện phàn nàn thương mại điện tử mà không thông qua tinh chỉnh phức tạp. Kết quả cho thấy các LLM đạt hiệu suất cao, với Accuracy và F1-score trung bình vượt mốc 0,90 trong hầu hết các trường hợp. Nghiên cứu góp phần mở rộng phạm vi ứng dụng của LLM trong xử lý ngôn ngữ tiếng Việt và cung cấp công cụ hữu ích cho doanh nghiệp nhằm tự động phát hiện vấn đề từ bình luận khách hàng.

1. Giới thiệu

Sự bùng nổ của thương mại điện tử và xu hướng tiêu dùng trực tuyến vẫn đang không ngừng phát triển. Năm 2024, tổng GMV (Gross Merchandise Value) của thị trường thương mại điện tử Việt Nam đạt mức 13,82 tỷ USD, tăng tương đương 40% so với năm 2023 (YouNetECI, 2024). Trong hành trình mua sắm của mình, khách hàng thường để lại những bình luận và chúng trở thành một nguồn tham khảo đối với những khách hàng tương lai và là nguồn dữ liệu quý giá đối với doanh nghiệp. Đây không chỉ là yếu tố ảnh hưởng đến quyết định mua hàng mà còn mang lại thông tin hỗ trợ định hướng, xây dựng các kế hoạch. Đặc biệt là những phàn nàn, đóng vai trò quan trọng trong cải thiện chất lượng sản phẩm, dịch vụ, nâng cao trải nghiệm và tăng trưởng doanh thu. Phân loại văn bản là một ý tưởng giải pháp có thể hỗ trợ phát hiện những phàn nàn. Với sự phổ biến của hướng tiếp cận học máy, các thuật toán truyền thống như logistic regression, support vector machine, decision tree… thường được sử dụng cho phân loại văn bản nhưng vẫn gặp một vài hạn chế. Yêu cầu về lượng lớn dữ liệu gán nhãn để huấn luyện mô hình và khả năng mở rộng phạm vi ứng dụng bị hạn chế là những vấn đề chủ yếu (Wang & cộng sự, 2023). Đồng thời, quá trình tiền xử lý dữ liệu để nâng cao hiệu suất mô hình làm tiêu tốn nhiều thời gian cũng là một vấn đề đáng quan tâm.

*Tác giả liên hệ. Email: *namts@ueh.edu.vn*
https://doi.org/10.61602/jdi.2026.88.07
Ngày nộp bài: 18/8/2025; Ngày chỉnh sửa: 25/9/2025; Ngày duyệt đăng: 06/10/2025; Ngày online: 02/02/2026
ISSN (print): 1859-428X, ISSN (online): 2815-6234

Trong kỷ nguyên phát triển của AI tạo sinh, các Mô hình Ngôn ngữ Lớn (Large Language Model - LLM) là những đại diện tiêu biểu, có khả năng ảnh hưởng đến hoạt động nghiên cứu và ứng dụng. Một trong những lĩnh vực bị ảnh hưởng đáng kể bởi LLM là xử lý ngôn ngữ tự nhiên với những thay đổi căn bản (Zhao & cộng sự, 2023). Nhờ vào ưu điểm về quy mô dữ liệu lớn đã được sử dụng cho huấn luyện, LLM mang lại tiềm năng to lớn ở hàng loạt các nhiệm vụ khác nhau, bao gồm phân loại văn bản (Chae & Davidson, 2025; Marvin & cộng sự, 2023). Tuy nhiên, số lượng các nghiên cứu trên ngôn ngữ tiếng Việt còn hạn chế và phạm vi ứng dụng chưa được mở rộng, đặc biệt là ở lĩnh vực thương mại điện tử. Mức độ quan tâm dành cho bài toán phát hiện phàn nàn vẫn còn thấp hơn giá trị mà nó có thể mang lại.

Do đó, nghiên cứu này được tiến hành với mục tiêu chính là ứng dụng LLM để đề xuất một hướng tiếp cận mới cho nhiệm vụ phát hiện phàn nàn, áp dụng trên các bình luận thương mại điện tử tại Việt Nam. Bằng việc tận dụng ưu điểm LLM, nghiên cứu kỳ vọng sẽ giải quyết hiệu quả một số hạn chế tồn tại trong hướng tiếp cận học máy truyền thống và mang lại một hiệu suất cao. Từ đó giúp nâng cao hiệu suất vận hành doanh nghiệp thông qua việc nhanh chóng triển khai và phát hiện kịp thời các vấn đề về sản phẩm, dịch vụ, hỗ trợ doanh nghiệp hoạch định kế hoạch kinh doanh tại Việt Nam. Ngoài ra, điều này còn góp phần gia tăng sự đa dạng của các nghiên cứu LLM trong một miền ngôn ngữ có những đặc trưng riêng biệt như tiếng Việt và mở rộng phạm vi ứng dụng trong các bài toán kinh doanh.

2. Các nghiên cứu liên quan

Sở hữu khả năng ấn tượng, các LLM có thể thực hiện hàng loạt nhiệm vụ liên quan đến ngôn ngữ khác nhau, từ tóm tắt văn bản đến sáng tạo nội dung. Việc tương tác và điều khiển hoạt động của LLM được thực hiện thông qua các chỉ dẫn đầu vào hay câu lệnh (prompt). Đây là thành phần có vai trò quan trọng, ảnh hưởng đến hiệu suất của các mô hình. Do đó, Marvin và cộng sự (2023) đã đưa ra những khuyến nghị trong việc xây dựng câu lệnh phù hợp, bao gồm việc xác định rõ mục tiêu câu lệnh, hiểu khả năng mô hình, lựa chọn định dạng câu lệnh phù hợp, cung cấp ngữ cảnh, kiểm tra và tinh chỉnh. Ngoài ra, nhóm tác giả cũng thực hiện khảo sát và cung cấp tổng quan về các chiến lược hay kỹ thuật xây dựng câu lệnh. Trong đó, few-shot cho phép mô hình học trong ngữ cảnh, các ví dụ minh họa được đưa trực tiếp vào câu lệnh nhằm định hướng mô hình đạt hiệu suất tốt hơn (Ahmed & cộng sự, 2023). Ngược lại, zero-shot sẽ không bao gồm các ví dụ trong câu lệnh mà chỉ có hướng dẫn thực hiện nhiệm vụ cho mô hình (Kojima & cộng sự, 2022).

Với những tiến bộ của LLM, Chae và Davidson (2025) đã nghiên cứu ứng dụng LLM vào lĩnh vực xử lý ngôn ngữ tự nhiên tại nhiệm vụ phân loại văn bản. Nhóm nghiên cứu so sánh mười mô hình có quy mô tham số khác nhau, áp dụng zero-shot, few-shot, tinh chỉnh với dữ liệu huấn luyện bổ sung và tinh chỉnh theo hướng dẫn. Phân tích các quan điểm được chọn là nhiệm vụ đánh giá với các bộ dữ liệu thu thập từ mạng xã hội. Kết quả đã cho thấy các mô hình lớn (như GPT-4o) nhìn chung cho hiệu suất dự đoán tốt nhất ngay cả có hoặc không có ví dụ. Ngoài ra, để khám phá và xác thực khả năng của LLM trong phân loại văn bản, Wang và cộng sự (2023) đề xuất phương pháp gợi ý chuỗi tư duy và zero-shot. Các mô hình có thể sử dụng các gợi ý suy luận từng bước, thay vì các định dạng hỏi-đáp thông thường. Llama2, GPT-3.5 và GPT-4 là các mô hình được sử dụng và đánh giá toàn diện trên bốn bộ dữ liệu khác nhau, bao gồm phân tích cảm xúc (tweet về COVID-19 và văn bản kinh tế), phân loại bốn lớp (văn bản thương mại điện tử) và phát hiện thư rác (SMS). Kết quả thực nghiệm khẳng định hiệu quả của LLM trong phân loại văn bản, với hiệu suất tốt trên hầu hết các bộ dữ liệu.

Ứng dụng khả năng phân loại của LLM, một số nghiên cứu trên thế giới đã thực hiện trong các bài toán kinh doanh cụ thể. Roumeliotis và cộng sự (2025) đã chứng minh hiệu quả phân loại khiếu nại của người tiêu dùng với LLM không tinh chỉnh. Trong bối cảnh tài chính, hiệu quả zero-shot của 14 mô hình, có sử dụng các mô hình suy luận (một dạng mở rộng của LLM) được xem xét thông qua câu lệnh giao nhiệm vụ phân loại khiếu nại. Nghiên cứu cho thấy các mô hình suy luận thể hiện năng lực vượt trội và đạt độ chính xác phân loại cao hơn, đánh dấu năng lực đáng kể trong tự động hóa xử lý khiếu nại. Nhằm ứng dụng phân tích cảm xúc để mang lại những hiểu biết toàn diện, Wangsa và cộng sự (2025) đề xuất một khung phân tích cảm xúc theo chủ đề. Trong đó, nghiên cứu ứng dụng BERT và phân cụm để nhóm các tài liệu theo các chủ đề khác nhau. Sau đó, sử dụng GPT-3.5-Turbo để phân loại cảm xúc của nội dung liên quan đến từng chủ đề thông qua các câu lệnh zero-shot. Kết quả cho thấy tính hiệu quả phân loại đạt Accuracy 0,87.

Tại Việt Nam, các nghiên cứu ứng dụng LLM trong các bài toán phân loại đã nhận được sự quan tâm trong những năm gần đây. Một số có thể kể đến như nghiên cứu điều tra hiệu quả của kỹ thuật thiết kế câu lệnh với các LLM khác nhau cho phân tích cảm xúc tiếng Việt của Dang Van Thin và cộng sự (2024). Nghiên cứu xem xét ba mẫu câu lệnh, kết hợp với chiến lược zero-shot và few-shot, trên các mô hình thuộc họ GPT và các mô hình mã nguồn mở, sử dụng sáu bộ dữ liệu chuẩn. Mô hình GPT-4, mẫu nhập vai mang lại hiệu suất cao nhất trên hầu hết các bộ dữ liệu. Đồng thời, few-shot đã nâng cao hiệu suất tổng thể cho phân tích cảm xúc, bất kể LLM là nguồn mở hay nguồn đóng. Nhằm đánh giá hiệu quả giữa PhoBERT và GPT-3.5-Turbo, Nguyen Ngoc Long và cộng sự (2023) đã có một thực nghiệm so sánh giữa hai mô hình này. Trong đó, nghiên cứu có thực hiện tiền xử lý dữ liệu để chuẩn bị đưa vào các mô

hình so sánh. Áp dụng đối với phân tích cảm xúc với ba cực trong lĩnh vực giáo dục, GPT-3.5-Turbo sử dụng one-shot và few-shot đã đạt độ chính xác lên đến 0,85.

Nhìn chung, các nghiên cứu trước đây đã chứng minh tiềm năng to lớn của LLM trong phân loại văn bản tại các miền khác nhau. Vai trò quan trọng của xây dựng câu lệnh đối với LLM được thiết lập rõ ràng, zero-shot và few-shot là những chiến lược phổ biến, định hướng mô hình đạt hiệu suất tốt. Ngoài ra, các mô hình GPT của OpenAI như GPT-3.5, GPT-4, GPT-4o hay các mô hình suy luận đã thể hiện năng lực trong các thực nghiệm trước đây. Một trong số ít các nghiên cứu tại Việt Nam cũng đã xác định mẫu nhập vai là mẫu câu lệnh mang lại hiệu suất cao nhất trong nhiều bộ dữ liệu tiếng Việt khác nhau. Những phát hiện này sẽ là nền tảng vững chắc cho nghiên cứu của chúng tôi, tập trung vào việc phát hiện phàn nàn trong các bình luận trên sàn thương mại điện tử Việt Nam.

3. Phương pháp nghiên cứu

3.1. Quy trình nghiên cứu

Trong nghiên cứu này, phương pháp nghiên cứu được thiết kế nhằm xây dựng đề xuất và đánh giá hiệu quả của các LLM cho nhiệm vụ phát hiện phàn nàn mà không cần tinh chỉnh. Tổng quan phương pháp nghiên cứu được trình bày theo một quy trình bốn giai đoạn chính như hình 1, giai đoạn đầu là chuẩn bị bộ dữ liệu nhằm tổng hợp và phân tách dữ liệu thành các bộ dữ liệu cần thiết cho quá trình phân tích với các LLM và đánh giá hiệu suất. Với mục tiêu khai thác sức mạnh của LLM mà không cần tinh chỉnh mô hình, giai đoạn hai sẽ tiến hành thiết kế các câu lệnh để hướng dẫn mô hình thực hiện nhiệm vụ phát hiện phàn nàn theo các chiến lược như zero-shot và few-shot. Sau đó, các LLM nổi bật sẽ nhận các câu lệnh và dữ liệu cần phân tích thông qua API để thực hiện nhiệm vụ. Cuối cùng, hiệu suất của hoạt động phát hiện phàn nàn được đánh giá dựa trên các chỉ số phổ biến trong bài toán phân loại như Accuracy và F1-score, xác định hiệu quả của các chiến lược và mô hình tối ưu trên các bình luận.

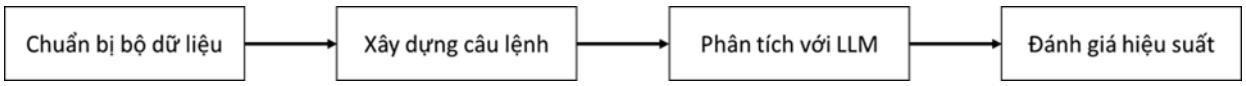
3.2. Chuẩn bị bộ dữ liệu

Để khám phá đề xuất, nghiên cứu sử dụng bộ dữ liệu UIT-ViOCD (Nguyen Thi-Hong Nhung & cộng sự, 2021) cho giai đoạn phân tích với LLM và giai đoạn đánh giá hiệu suất. Đây là bộ dữ liệu về bình luận khách hàng, được thu thập từ các website thương mại điện tử. Theo đó, mỗi bình luận sẽ thuộc về một trong bốn miền, gồm có "Di động", "Thời trang", "Mỹ phẩm" và "Ứng dụng". Đồng thời, mỗi bình luận cũng được gán một trong hai nhãn thể hiện sự tồn tại của phàn nàn là "1" và "0". Cụ thể, nhãn "1" dành cho các bình luận thường thể hiện sự không hài lòng của khách hàng giữa thực tế và kỳ vọng, chiếm khoảng 52,0%. Ngược lại, nhãn "0" được gán cho các bình luận không phàn nàn, lời khen, sự hài lòng, chiếm khoảng 48,0%. Đánh giá về độ sạch, đây là bộ dữ liệu có chứa nhiều vấn đề liên quan đến dữ liệu văn bản, như viết tắt, teen code, sai chính tả, … thường yêu cầu quá trình làm sạch trong các hướng tiếp cận học máy truyền thống.

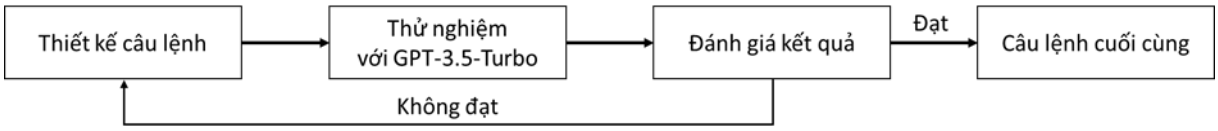
Bởi bình luận ở từng miền có những đặc điểm khác biệt, nghiên cứu tiến hành tách bộ dữ liệu UIT-ViOCD thành bốn bộ dữ liệu miền. Sau đó, mỗi bộ dữ liệu miền sẽ được tách thành bộ dữ liệu huấn luyện và kiểm tra theo tỷ lệ 9:1 để phục vụ cho phân tích và đánh giá.

3.3. Xây dựng câu lệnh

Tham khảo những khuyến nghị để xây dựng câu lệnh phù hợp của Marvin và cộng sự (2023), nghiên cứu đặt mục tiêu tạo ra phản hồi phân loại phàn nàn là "1" hoặc "0" từ LLM. Dựa trên kết quả thực nghiệm của Dang Van Thin và cộng sự (2024) và hướng dẫn tạo câu lệnh của OpenAI (2025b), mẫu câu lệnh nhập vai được áp dụng để thiết kế câu lệnh, chỉ định vai trò cho LLM và yêu cầu thực hiện nhiệm vụ. Theo đó, LLM nhập vai như một chuyên viên phân loại văn bản với nhiệm vụ gán nhãn các bình luận thương mại điện tử có chứa yếu tố phàn nàn. Đồng thời, định dạng Markdown và XML cũng được sử dụng trong câu lệnh để giúp mô hình hiểu rõ hơn về yêu cầu. Sử dụng GPT-3.5-Turbo, một LLM



Hình 1. Quy trình nghiên cứu



Hình 2. Quy trình thiết kế câu lệnh

phiên bản cũ của OpenAI, nghiên cứu thử nghiệm khám phá khả năng LLM đối với việc đáp ứng mục tiêu đặt ra.

Kết quả thử nghiệm được đánh giá dựa trên mức độ thực hiện đúng yêu cầu, xem xét

định dạng và sự đa dạng kết quả của phản hồi. Trường hợp kết quả không đạt, câu lệnh được thiết kế lại và ngược lại thì sẽ được chọn làm câu lệnh cuối cùng. Hình 2 trình bày quy trình để thiết kế câu lệnh trong nghiên cứu này. Kết quả xây dựng, câu lệnh cuối cùng có dạng như sau:

Identity

Bạn là chuyên gia phân loại văn bản, hãy gán nhãn các bình luận thương mại điện tử có sự phàn nàn là '1', ngược lại là '0'.

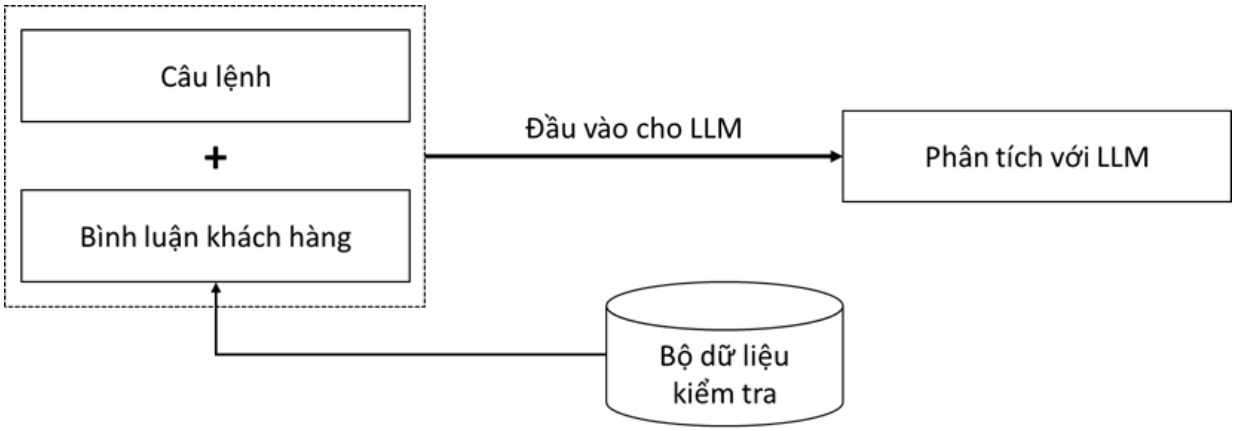
Instructions

Câu trả lời có một từ duy nhất, là '1' hoặc '0' và không có bình luận bổ sung.

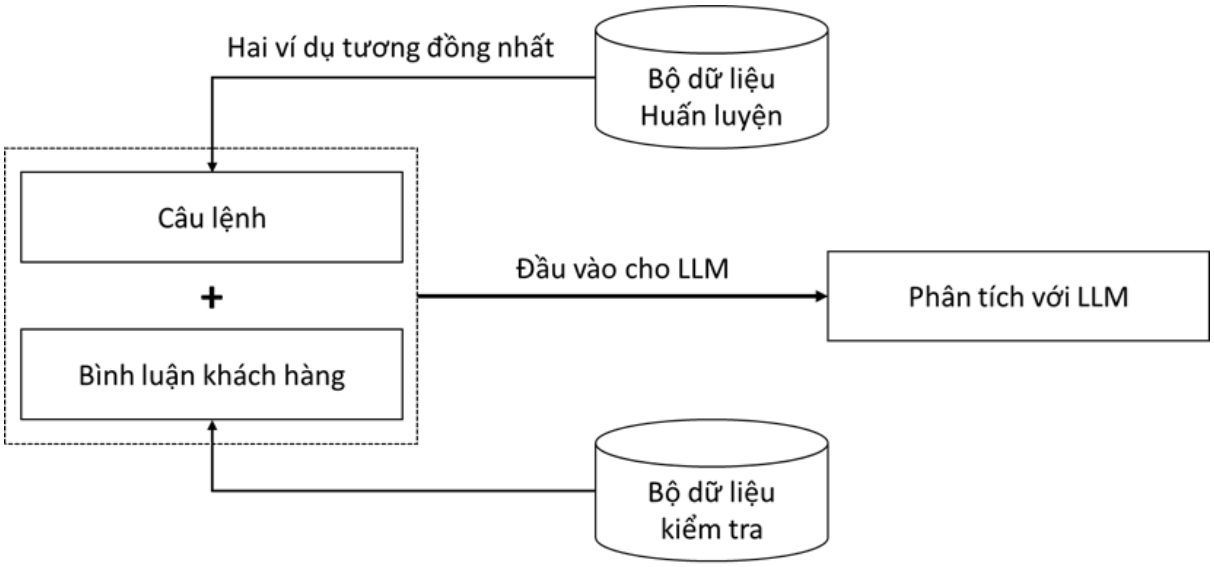
Với câu lệnh cuối cùng, nghiên cứu áp dụng song song chiến lược zero-shot và few-shot để tìm kiếm sự tối ưu. Hình 3 cho thấy cách thức thiết lập chiến lược zero-shot với câu lệnh kết hợp lần lượt với các bình luận từ bộ dữ liệu kiểm tra để làm đầu vào cho giai đoạn phân tích với LLM.

Trong khi đó, chiến lược few-shot trong nghiên cứu này sẽ áp dụng 2-shot, với 2 ví dụ cho mỗi bình luận cần được phân tích, mỗi ví dụ đại diện cho một nhóm, gồm nhóm có phàn nàn và không phàn nàn. Đáng chú ý, nghiên cứu lựa chọn các ví dụ cho mỗi nhóm dựa trên mức độ tương đồng so với bình luận được phân tích theo Cosine similarity. Điều này đặt ra kỳ vọng sẽ mang đến những ví dụ sát với trường hợp đang phân tích hơn là lựa chọn ví dụ mang tính ngẫu nhiên. Bộ dữ liệu huấn luyện được chọn là nguồn cung cấp các ví dụ cho chiến lược few-shot. Hình 4 mô tả quy trình chiến lược few-shot.

Các câu lệnh cuối cùng cho chiến lược zero-shot và few-shot được trình bày cụ thể trong bảng 1.



Hình 3. Quy trình triển khai zero-shot



Hình 4. Quy trình triển khai few-shot

Bảng 1. Các câu lệnh cho chiến lược zero-shot và few-shot

	Câu Lệnh
Zero-shot	# Identity Bạn là chuyên gia phân loại văn bản, hãy gán nhãn các bình luận thương mại điện tử có sự phàn nàn là '1', ngược lại là '0'. # Instructions Câu trả lời có một từ duy nhất, là '1' hoặc '0' và không có bình luận bổ sung.
Few-shot	# Identity Bạn là chuyên gia phân loại văn bản, hãy gán nhãn các bình luận thương mại điện tử có sự phàn nàn là '1', ngược lại là '0'. # Instructions Câu trả lời có một từ duy nhất, là '1' hoặc '0' và không có bình luận bổ sung. # Examples <review id='example-1'> *[bình luận không phàn nàn giống với bình luận cần phân tích nhất]* </review> <response id='example-1'> 0 </response> <review id='example-2'> *[bình luận phàn nàn giống với bình luận cần phân tích nhất]* </review> <response id='example-2'> 1 </response>

3.4. Phân tích với LLM và đánh giá hiệu suất

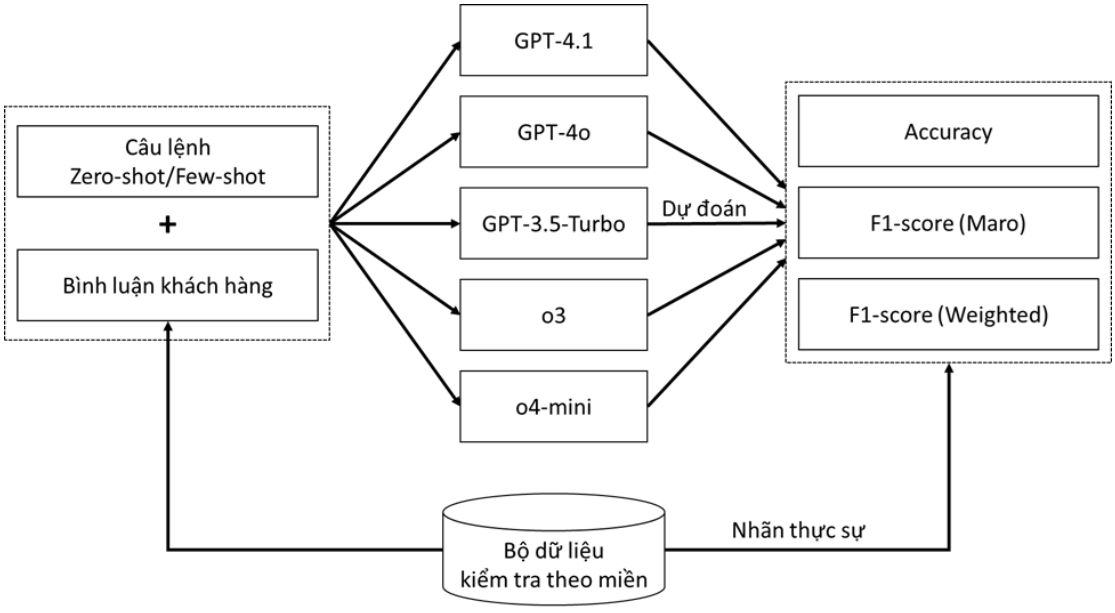
Là thành phần chính thực hiện vai trò phân tích trong đề xuất, nghiên cứu này tận dụng các LLM của OpenAI thông qua API, bao gồm GPT-4.1, GPT-4o, GPT-3.5-Turbo, o3 và o4-mini (OpenAI, 2025a). Đặc điểm của các mô hình như sau:

- GPT-4.1, mô hình GPT được đánh giá là sở hữu mức độ thông minh rất cao và tốc độ trung bình, chuyên dành cho các tác vụ phức tạp trong nhiều lĩnh vực.
- GPT-4o, mô hình GPT có mức độ thông minh cao và tốc độ trung bình, được xem là một hình linh hoạt, chấp nhận đa dạng các loại đầu vào.
- GPT-3.5-Turbo, là một mô hình GPT thuộc nhóm phiên bản cũ nên có mức độ thông minh thấp và tốc độ chậm. Đây cũng là một mô hình có chi phí rẻ, có thể hiểu, tạo ngôn ngữ tự nhiên và hoạt động tốt cho các tác vụ không phải trò chuyện.
- o3, mô hình lý luận mạnh mẽ nhất của OpenAI trên nhiều lĩnh vực nhưng có tốc độ rất chậm. Mô hình này vượt trội trong việc làm theo hướng dẫn, phù hợp phân tích văn bản.
- o4-mini, mô hình lý luận được đánh giá thấp hơn o3 nhưng được cải thiện tốc độ nhanh hơn đáng kể và có chi phí thấp hơn.

Với bộ dữ liệu đã chuẩn bị, nghiên cứu thực hiện phân tích trên từng bộ dữ liệu miền và trên từng LLM tách biệt. Dữ liệu bình luận từ bộ kiểm tra sẽ là nguồn để các mô hình phân tích. Kết quả sẽ kết hợp với danh sách nhãn thực sự của bộ dữ liệu kiểm tra để đánh giá thông qua các chỉ số phổ biến là Accuracy, F1-score (macro) và F1-score (weighted). Hình 5 tổng quan về quy trình phân tích với LLM và đánh giá hiệu suất.

4. Kết quả và thảo luận

Bảng 2 trình bày hiệu suất trung bình của năm LLM khi áp dụng vào nhiệm vụ phát hiện phàn nàn trong bốn miền bình luận. Mỗi miền được đánh giá trong hai thiết



Hình 5. Quy trình phân tích với LLM và đánh giá hiệu suất

lập chiến lược: zero-shot (Zs) và few-shot (Fs), với ba chỉ số đo lường hiệu suất: Accuracy; F1-score (macro): F1 (M); F1-score (weighted): F1 (W). Nhìn chung, các mô hình thể hiện hiệu suất vượt trội trên cả bốn miền, hầu hết chỉ số đều trên 0,90. Điều này khẳng định tiềm năng và hiệu quả cao trong việc ứng dụng LLM cho bài toán phát hiện phàn nàn của khách hàng mà không cần huấn luyện lại toàn bộ mô hình trên một tập dữ liệu lớn.

Đánh giá giữa các miền, hiệu suất trung bình của năm LLM tại zero-shot có sự chênh lệch nhỏ, phản ánh đặc điểm ngôn ngữ khác biệt. "Thời trang" và "Mỹ phẩm" là hai miền mà các mô hình hoạt động tốt nhất, các chỉ số lên tới 0,94 và 0,95. Nguyên nhân của hiệu suất cao này có thể đến từ việc ngôn ngữ phàn nàn trong các miền này thường mang tính phổ thông, xoay quanh các chủ đề quen thuộc. Những mẫu câu này xuất hiện phổ biến trong dữ liệu huấn luyện khổng lồ của LLM, giúp mô hình dễ dàng nhận diện. Hai miền còn lại là "Di động" và "Ứng dụng" có hiệu suất thấp hơn, sự sụt giảm này có thể được lý giải bởi tính phức tạp của ngôn ngữ phàn nàn. Các bình luận thường chứa thuật ngữ chuyên biệt (chip, seal, fix, …) hoặc đề cập đến các vấn đề chi tiết. Những bình luận này không chỉ đa dạng mà còn đòi hỏi mô hình phải có một mức độ hiểu biết kỹ thuật nhất định để phân loại chính xác, dẫn đến việc các LLM gặp nhiều khó khăn.

Phân tích sâu hơn về hai chiến lược, cho thấy một vài điểm đáng chú ý về sự cải thiện của few-shot so với zero-shot. Tại miền "Thời trang", sự cải thiện là rõ rệt nhất với mức tăng 0,03 ở tất cả các chỉ số (đều tăng từ 0,94 lên 0,97). Điều này thể hiện rằng ngôn ngữ phàn nàn các sản phẩm thời trang cũng có một số đặc thù nhất định, khả năng tự phân tích của các mô hình (zero-shot) chưa thể tối ưu nhất, bằng việc hỗ trợ một vài ví dụ sẽ giúp nhận diện phàn nàn chính xác hơn. Tại miền "Di động" và "Mỹ phẩm", few-shot cũng mang lại sự cải thiện, mặc dù không lớn bằng miền "Thời trang". Một trường hợp đặc biệt là miền "Ứng dụng", nơi hiệu suất của zero-shot và few-shot là hoàn toàn tương đồng (Accuracy 0,91, F1-score (macro) 0,88, F1-score (weighted) 0,91). Vấn đề này có thể được giải thích rằng khả năng phân tích ban đầu của các mô hình (zero-shot) đã đạt đến mức tối ưu trên miền này với phương pháp đề xuất, đòi hỏi các phương pháp can thiệp chuyên sâu hơn. Theo cách giải thích khác, các ví dụ trong few-shot không đủ đa dạng và mới mẻ để bao quát hết các loại vấn đề kỹ thuật, do đó không mang lại thông tin hữu ích giúp mô hình cải thiện tốt hơn.

Bảng 2. Hiệu suất trung bình của các LLM trên các miền

Đánh giá	Di động		Thời trang		Mỹ phẩm		Ứng dụng	
	Zs	Fs	Zs	Fs	Zs	Fs	Zs	Fs
Accuracy	0,91	0,92	0,94	0,97	0,95	0,95	0,91	0,91
F1 (M)	0,90	0,91	0,94	0,97	0,94	0,95	0,88	0,88
F1 (W)	0,91	0,92	0,94	0,97	0,95	0,95	0,91	0,91

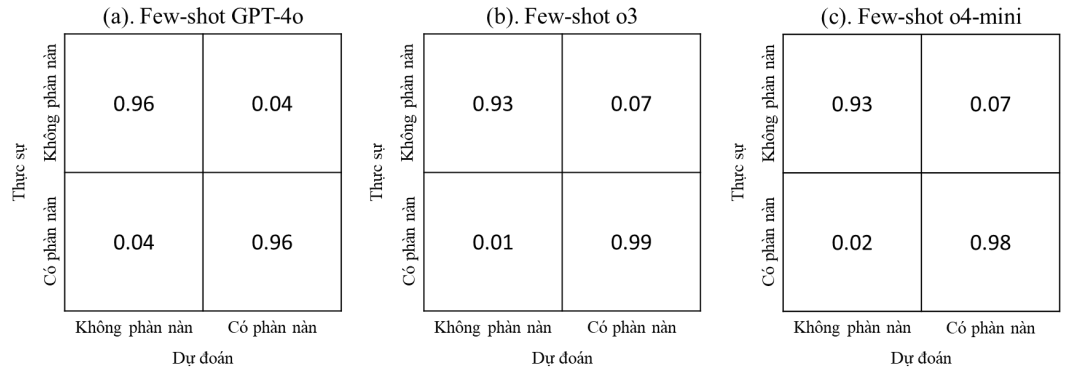
Bảng 3 trình bày kết quả đánh giá, so sánh hiệu suất của năm LLM khác nhau trên bộ dữ liệu UIT-ViOCD. Kết quả thể hiện một sự phân cấp rõ rệt về hiệu suất giữa các mô hình được sử dụng. Nhóm mô hình GPT-4o, o3 và o4-mini nổi bật với hiệu suất cao nhất và gần như tương đương nhau. Ở zero-shot, cả ba mô hình này đều đạt điểm số đồng nhất là 0,95 trên tất cả các chỉ số. Khi chuyển sang few-shot, hiệu suất cả ba mô hình cùng tăng lên 0,96. Điều này cho thấy các mô hình có khả năng phân tích mạnh mẽ và học dựa trên ví dụ tốt. Sự vượt trội này phản ánh sức mạnh của các kiến trúc mô hình và quy mô dữ liệu huấn luyện, cho phép chúng có khả năng suy luận và phân tích mạnh mẽ ngay cả trong chiến lược zero-shot. Mô hình GPT-4.1 cũng chứng minh hiệu suất cao, chênh lệch không đáng kể so với nhóm dẫn đầu. Cụ thể, GPT-4.1 đạt 0,94 ở zero-shot và tăng lên 0,95 ở few-shot. Mặc dù có sự chênh lệch nhỏ, đây vẫn là một kết quả tốt, khẳng định năng lực của mô hình. Cuối cùng, GPT-3.5-Turbo là mô hình có hiệu suất thấp nhất và cách biệt đáng kể so với các mô hình còn lại, tất cả chỉ số đều dưới 0,9.

Một xu hướng nhất quán trên tất cả năm LLM là few-shot luôn mang lại hiệu suất cao hơn so với zero-shot. Đối với các mô hình hiệu suất cao, việc cung cấp một vài ví dụ giúp cải thiện 0,01 trên tất cả các chỉ số. Mức tăng này tuy nhỏ nhưng cho thấy khả năng học theo ngữ cảnh của các mô hình này. Đối với mô hình GPT-3.5-Turbo, tác động của few-shot là rõ rệt nhất, với mức tăng từ 0,03 đến 0,04. Điều này thể hiện mô hình có hiệu suất cơ bản thấp sẽ được hưởng lợi nhiều hơn từ việc được cung cấp các ví dụ minh họa, giúp chúng hiểu rõ hơn về yêu cầu của nhiệm vụ.

Nhìn nhận sâu sắc hơn về hiệu suất của các LLM hàng đầu, hình 6 trình bày ba ma trận nhầm lẫn của ba mô hình GPT-4o, o3, o4-mini tại few-shot. Cả ba đều đạt hiệu suất cao nhưng chúng có những thiên hướng khác biệt. Theo đó, GPT-4o là mô hình cân bằng và toàn diện, độ tin cậy cao

Bảng 3. Hiệu suất của năm LLM trên bộ dữ liệu UIT-ViOCD

Đánh giá	GPT-4.1		GPT-4o		GPT-3.5-Turbo		o3		o4-mini		Trung Bình	
	Zs	Fs	Zs	Fs	Zs	Fs	Zs	Fs	Zs	Fs	Zs	Fs
Accuracy	0,94	0,95	0,95	0,96	0,85	0,88	0,95	0,96	0,95	0,96	0,93	0,94
F1 (M)	0,94	0,95	0,95	0,96	0,84	0,88	0,95	0,96	0,95	0,96	0,93	0,94



Hình 6. Ma trận nhầm lẫn của GPT-4o, o3 và o4-mini tại few-shot

trên cả hai lớp, không chênh lệch và tỷ lệ lỗi false positive và false negative tương đương nhau, phù hợp cho các trường hợp yêu cầu về sự cân bằng. Ngược lại, o3 và o4-mini tối ưu cho việc không bỏ sót phàn nàn, với tỷ lệ true positive cực kỳ cao, o3 là mô hình nhạy nhất trong việc phát hiện phàn nàn. Hai mô hình này phù hợp cho các trường hợp mà việc giải quyết mọi vấn đề của khách hàng là ưu tiên, chấp nhận việc có thể phải xem xét một vài cảnh báo ảo.

Với những kết quả đạt được, LLM ứng dụng vào nhiệm vụ phát hiện phàn nàn đã mang lại nhiều ưu điểm nổi bật. Các mô hình không cần tinh chỉnh phức tạp nhưng vẫn đạt hiệu suất cao trên nhiều miền và có khả năng chấp nhận dữ liệu không thông qua bước tiền xử lý. Tất cả cho thấy sự vượt trội ở một số khía cạnh so với hướng tiếp cận học máy truyền thống. Mặc dù có tiềm năng lớn, việc ứng dụng LLM trong nghiên cứu này vẫn tồn tại một số thách thức nhất định.

Vấn đề đầu tiên là chi phí tính toán, việc sử dụng dịch vụ API từ OpenAI khiến cho chi phí vận hành có thể tăng cao theo quy mô dữ liệu, trở thành rào cản đối với các doanh nghiệp vừa và nhỏ. Đây không chỉ là một vấn đề ngân sách mà còn là một rào cản về khả năng mở rộng. Đối với các doanh nghiệp có hàng triệu tương tác khách hàng mỗi ngày, chi phí có thể trở nên quá lớn, buộc họ phải lựa chọn giữa việc phân tích toàn bộ dữ liệu bằng một mô hình chi phí thấp hơn, kém chính xác hơn hoặc chỉ phân tích một phần dữ liệu với mô hình tốt nhất. Cả hai lựa chọn đều dẫn đến việc bỏ lỡ những thông tin kinh doanh quý giá. Một thách thức quan trọng khác là vấn đề thiên kiến và tính công bằng của mô hình. Do học từ nguồn dữ liệu rộng lớn, các LLM có nguy cơ phản ánh các định kiến xã hội khi sử dụng trực tiếp, điều này có thể ảnh hưởng tiêu cực đến đầu ra của mô hình. Ngoài ra, các LLM cũng không thể đáp ứng hiệu suất tốt đồng đều trên tất cả các miền bình luận và cũng không thể cải thiện được hiệu suất với few-shot tại một số miền, đặt ra các yêu cầu cải thiện nâng cao. Bên cạnh đó, hai mô hình lý luận là o3 và o4-mini thuộc nhóm các mô hình hàng đầu, có khả năng hạn chế bỏ sót phàn nàn rất tốt nhưng gặp vấn đề lớn về tốc độ phản hồi khiến cho hai mô hình này khó lòng đáp ứng trong các trường hợp bình luận sinh ra với tốc độ cao. Trong thực tế, đây là một hạn chế rất đang quan tâm khi các ứng dụng như giám sát mạng xã hội hay chatbot hỗ trợ trực tiếp đòi hỏi phải phát hiện và phản ứng với phàn nàn nhanh chóng. Một mô hình chậm trễ có thể khiến doanh nghiệp chỉ phát hiện một cuộc khủng hoảng khi nó đã lan rộng, làm mất đi thời điểm để can thiệp kịp thời.

5. Kết luận

Nghiên cứu này đã chứng minh tiềm năng của đề xuất ứng dụng LLM trong nhiệm vụ phát hiện phàn nàn từ các bình luận trên website thương mại điện tử tại Việt Nam. Thông qua hai chiến lược câu lệnh là zero-shot và few-shot, các mô hình của OpenAI đều cho thấy hiệu suất cao. Trong đó, các mô hình hiện đại như GPT-4o, o3 và o4-mini đạt độ chính xác gần như tuyệt đối ở nhiều miền dữ liệu. Chiến lược few-shot giúp cải thiện hiệu suất trong hầu hết các trường hợp, đặc biệt với các mô hình có năng lực cơ bản thấp. Nghiên cứu cũng đã mở ra hướng đi mới cho các doanh nghiệp thương mại điện tử tại Việt Nam để tự động hóa và tối ưu hóa một phần trong quy trình chăm sóc khách hàng. Việc ứng dụng LLM giúp doanh nghiệp nhanh chóng triển khai và phát hiện kịp thời các phàn nàn, từ đó đưa ra các phản hồi giúp cải thiện trải nghiệm khách hàng và nâng cao uy tín thương hiệu. Các doanh nghiệp cũng có thể triển khai thí điểm LLM để phân tích xu hướng phàn nàn, làm cơ sở cải thiện chất lượng sản phẩm và dịch vụ. Mặc dù có nhiều tiềm năng, việc ứng dụng LLM vào thực tiễn vẫn còn một số thách thức như chi phí vận hành, tốc độ xử lý hay hiệu suất chênh lệch giữa các miền. Từ đó, hướng đi tiếp theo có thể là tinh chỉnh mô hình cho các miền cụ thể, tối ưu hóa chi phí, cũng như mở rộng ứng dụng cho các bài toán kinh doanh khác trong bối cảnh ngôn ngữ tiếng Việt.

TÀI LIỆU THAM KHẢO

Ahmed, T., Pai, K. S., Devanbu, P., & Barr, E. T. (2023). *Improving few-shot prompts with relevant static analysis products*. arXiv. DOI: https://doi.org/10.48550/arXiv:2304.06815.

Chae, Y., & Davidson, T. (2025). Large language models for text classification: from zero-shot learning to instruction-tuning. *Sociological Methods Research*, 55(2), 501-567. DOI: https://doi.org/10.1177/00491241251325243.

Dang Van Thin, Duong Ngoc Hao & Ngan Luu-Thuy Nguyen (2024). Prompt engineering with large language models for Vietnamese sentiment classification. *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, Tokyo, Japan. Access at https://aclanthology.org/2024.paclic-1.17/.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). *Large language models are zero-shot reasoners*. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, LA, USA. Access at https://dl.acm.org/doi/10.5555/3600270.3601883

Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2023). Prompt engineering in large language models. *International Conference on Data Intelligence and Cognitive Informatics*, Tirunelveli, India. DOI: https://doi.org/10.1007/978-981-99-7962-2_30

Nguyen Ngoc Long, Ngo Doan Kien, Nguyen Thi Hong Hanh, Nguyen Thi Kieu Nhung, Nguyen Son Tung, & Tuan Nguyen. (2023). *Unveiling sentiments in Vietnamese education texts: Could large language model GPT-3.5-turbo beat PhoBERT?* International Conference on Computational Data and Social Networks, Hanoi, Vietnam. DOI: https://doi.org/10.1007/978-981-97-0669-3_12

Nguyen Thi-Hong Nhung, Ha Phan-Dieu Phuong, Nguyen Thanh Luan, Nguyen Van Kiet, & Nguyen Luu-Thuy Ngan (2021). Vietnamese complaint detection on e-commerce websites. *Proceedings of the 20th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques*, Cancun, Mexico. https://doi.org/10.3233/FAIA210058

OpenAI (2025a). *Models*. Access at https://platform.openai.com/docs/models. Access on 31/07/2025

OpenAI (2025b). *Prompt engineering - Enhance results with prompt engineering strategies.* Access at https://platform.openai.com/docs/guides/prompt-engineering. Access on 31/07/2025

Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Think Before You Classify: The Rise of Reasoning Large Language Models for Consumer Complaint Detection and Classification. *Electronics*, 14(6), 1070. DOI: https://doi.org/10.3390/electronics14061070

Wang, Z., Pang, Y., & Lin, Y. (2023). *Large language models are zero-shot text classifiers*. DOI: https://doi.org/10.48550/arXiv.2312.01044.

Wangsa, J. I. P., Agung, Y. J., Rahmi, S. R., Murfi, H., Hariadi, N., Nurrohmah, S., Satria, Y., & Za'in, C. (2025). Large Language Model-Based Topic-Level Sentiment Analysis for E-Grocery Consumer Reviews. *Big Data and Cognitive Computing*, 9(8), 194. DOI: https://doi.org/10.3390/bdcc9080194

YouNetECI (2024). *Báo cáo doanh thu các sàn TMĐT quý IV/2024*. Truy cập tại https://youneteci.com/bao-cao-doanh-thu-cac-san-tmdt-quy-4-2024/. Ngày truy cập: 31/07/2025

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., & Dong, Z. (2023). *A survey of large language models*. arXiv. DOI: https://doi.org/10.48550/arXiv.2303.18223.

Detecting Complaints on E-Commerce Websites Using Large Language Models

Thai Kim Phung[1], Tran Son Nam[2], Do Thi Ngoc Thuy[3]
[1]University of Economics Ho Chi Minh City, Vietnam
[2]University of Economics Ho Chi Minh City - Vinh Long Campus, Vietnam
[3]National Credit Information Centre of Vietnam

Abstract

In the context of the rapid proliferation of e-commerce, customer comment on online platforms, particularly complaints, holds significant value for businesses. Although traditional machine learning approaches are prevalent in text classification, they present certain limitations that hinder their application in tasks such as complaint detection. Therefore, this study proposes a novel approach that utilizes Large Language Model (LLM) for the task of e-commerce complaint detection without requiring complex fine-tuning. The results indicate that LLM achieve high performance, with average Accuracy and F1-scores exceeding 0.90 in most cases. This study contributes to expanding the application scope of LLM in Vietnamese language processing and provides a valuable tool for enterprises to automatically identify issues from customer comments.

Keywords: Detecting complaint, e-commerce, large language model, online comment, prompt engineering.