Tạp chí Phát triển Khoa học và Công nghệ – Economics-Law and Management 2025, ():1-9

Open Access Full Text Article

Bài nghiên cứu

Xây dựng ngân hàng câu cảm xúc tài chính về hành vi đầu tư của khối ngoại

Phạm Thị Thanh Xuân, Đặng Anh Như, Từ Hà Phúc*

Use your smartphone to scan this QR code and download this article

TÓM TẮT

Nghiên cứu này xây dựng ``Ngân hàng câu cảm xúc tài chính về hành vi của khối ngoại'', một bộ dữ liệu gán nhãn cảm xúc từ tin tức tài chính nhằm mô phỏng phản ứng thị trường trước giao dịch của khối ngoại tại Việt Nam. Thay vì sử dụng mô hình phân tích tự động có sẵn, nghiên cứu phát triển quy trình tùy chỉnh linh hoạt, tối ưu theo đặc thù ngôn ngữ và thị trường Việt Nam, đồng thời có thể mở rộng sang các thị trường khác. Quy trình đảm bảo độ tin cậy nhờ chuyên gia tài chính gán nhãn thay vì từ điển tự động, giúp loại bỏ sai số và nhiễu thường gặp. Nghiên cứu có hai đóng góp chính: Về học thuật, nghiên cứu cung cấp khung phân tích cảm xúc tài chính chuẩn hóa, phát triển bộ dữ liệu cảm xúc tài chính đầu tiên cho Việt Nam, lấp đầy khoảng trống nghiên cứu về tâm lý nhà đầu tư khối ngoại. Về thực tiễn, ngân hàng câu cảm xúc có thể ứng dụng làm đầu vào cho mô hình phân tích tin tức tài chính trên kiến trúc BERT, dự báo thị trường và hỗ trợ quyết định đầu tư. Ngân hàng gồm 5.126 câu gán nhãn cảm xúc, có quy mô lớn hơn Financial PhraseBank (4.840 câu) – tập dữ liệu phổ biến trong phân tích tài chính. Tuy nhiên, khác với Financial PhraseBank vốn dựa trên tiếng Anh cho thị trường quốc tế, bộ dữ liệu này được thiết kế riêng cho thị trường chứng khoán Việt Nam, phản ánh đặc thù về ngôn ngữ, tâm lý nhà đầu tư và hành vi giao dịch của khối ngoại. Điều này giúp bộ dữ liệu trở thành tài nguyên quan trọng để phát triển mô hình phân tích cảm xúc tài chính tiếng Việt, tương tự như cách Financial PhraseBank hỗ trợ sự phát triển của FinBERT trong phân tích tài chính quốc tế.

Từ khoá: Cảm xúc tài chính, Hành vi nhà đầu tư khối ngoại, Phân tích cảm xúc, Học máy, Tài chính hành vi

Trường Đại học Kinh tế - Luật, Đại học Quốc gia TP. HCM, Việt Nam

Liên hệ

Từ Hà Phúc, Trường Đại học Kinh tế - Luật, Đại học Quốc gia TP. HCM, Việt Nam

Email: phucth22406@st.uel.edu.vn

Lịch sử

- Ngày nhận: 07-01-2025
- Ngày sửa đổi: 11-3-2025
- Ngày chấp nhận: 19-3-2025
- Ngày đăng:

DOI :

Check for updates

VNUHCM PRESS

GIỚI THIỆU

Hành vi của khối ngoại có ảnh hưởng lớn không chỉ đến tâm lý nhà đầu tư trong nước nói riêng mà còn tác động đến cả thị trường chứng khoán Việt Nam nói chung [1,2]. Những động thái của khối ngoại, đặc biệt là việc bán ròng, không chỉ phản ánh các tín hiệu rủi ro mà còn tạo ra cơ hội đầu tư trong bối cảnh kinh tế vĩ mô đầy biến động. Mặc dù tỷ trọng giao dịch của khối ngoại đã giảm dần trong những năm gần đây, vai trò của họ trong việc định hình tâm lý nhà đầu tư trong nước vẫn rất đáng kể. Khi khối ngoại thực hiện các giao dịch bán ròng quy mô lớn, tâm lý lo ngại thường gia tăng, tạo áp lực đáng kể đối với nhà đầu tư cá nhân. Điều này không chỉ ảnh hưởng đến xu hướng đầu tư ngắn hạn mà còn làm gia tăng sự bất ổn trên thị trường, vốn dễ bị chi phối bởi tâm lý đám đông. Hành vi giao dịch của khối ngoại, với phần lớn là các tổ chức đầu tư chuyên nghiệp, được xem như một kênh thông tin và chỉ báo quan trọng về rủi ro thị trường. Các quyết định đầu tư của nhóm này thường phản ánh những biến động từ cấp độ vĩ mô đến vi mô, bao gồm bất ổn kinh tế, lạm phát và rủi ro hệ thống. Năm 2024 đã chứng kiến mức bán ròng kỷ lục của khối ngoại, với hơn 83.700 tỷ đồng chỉ tính riêng đến giữa tháng 11 và tổng giá trị bán ròng gần 90.000 tỷ đồng, cao gấp đôi so với năm 2023. Không chỉ các quỹ ETF mà cả các quỹ đầu tư chủ động cũng rút vốn mạnh, trong khi tỷ lệ sở hữu của nhà đầu tư nước ngoài trên VN-Index tiếp tục giảm. Điều này phản ánh những thách thức ngày càng lớn trong việc duy trì sức hút của thị trường chứng khoán Việt Nam đối với dòng vốn quốc tế, đặc biệt trong bối cảnh Việt Nam vẫn chưa được nâng hạng lên thị trường mới nổi. Chính vì vậy, nghiên cứu này được thực hiện với trọng tâm là năm 2024, nhằm phân tích tác động của hành vi khối ngoại đối với thị trường, đặc biệt là dưới góc độ tâm lý hành vi.

Phần lớn các nghiên cứu trước đây tập trung vào việc xác định bằng chứng thực nghiệm về tác động của hành vi khối ngoại đối với thị trường mới nổi, từ đó góp phần luận giải và củng cố các lý thuyết tài chính hành vi, đơn cử như lý thuyết bầy đàn [3]. Tuy nhiên, thiếu vắng nghiên cứu đầu tư vào việc phát triển những tri thức này thành các công cụ có thể ứng dụng trực tiếp trong thực tiễn, chẳng hạn như xây dựng một bộ từ điển cảm xúc tài chính chuẩn

Trích dẫn bài báo này: Xuân P T T, Như D A, Phúc T H. **Xây dựng ngân hàng câu cảm xúc tài chính về hành vi đầu tư của khối ngoại**. *Sci. Tech. Dev. J. - Eco. Law Manag.* 2025; ():1-9.

hóa. Trong khi đó, sự phát triển nhanh chóng của các mô hình phân tích cảm xúc tự động đặt ra nhu cầu cấp thiết về việc chuyển hóa những bằng chứng thực nghiệm này thành cơ sở dữ liệu khoa học có thể phục vụ trực tiếp cho các hệ thống phân tích tài chính thông minh. Nghiên cứu này đi theo hướng đó, không dừng lại ở việc xác nhận rằng hành vi khối ngoại có tác động đến nhà đầu tư và thị trường, mà còn phát triển một quy trình gán nhãn cảm xúc để xác lập mối quan hệ này, từ đó tạo thành một cơ sở dữ liệu chuẩn hóa. Cụ thể, nhóm nghiên cứu áp dụng phương pháp phân tích tin tức tài chính nhằm đánh giá mức độ tích cực hoặc tiêu cực của các thông tin liên quan đến hành vi giao dịch của khối ngoại. Mục tiêu cuối cùng là xây dựng một nguồn dữ liệu có hệ thống, phục vụ cho cả nghiên cứu tài chính hành vi lẫn ứng dụng thực tiễn trong phân tích thị trường. Nghiên cứu đã phân tích 4.065 bản tin và xây dựng sản phẩm "Ngân hàng câu cảm xúc tài chính về hành vi của khối ngoại"—một bộ dữ liệu gồm 5.126 câu đã được gán nhãn cảm xúc tương ứng. Bộ dữ liệu này có giá trị ứng dụng cao, tương tự như Financial PhraseBank—một tập dữ liệu gồm 4.840 câu từ tin tức tài chính, được gán nhãn tích cực hoặc tiêu cực, do Malo và cộng sự khai thác[4]. Financial PhraseBank đã đóng vai trò quan trọng trong sự phát triển của FiinBERT, một mô hình BERT tinh chỉnh chuyên biệt cho phân tích cảm xúc tài chính. Tuy nhiên, trong khi Financial PhraseBank được xây dựng dựa trên ngữ liệu tiếng Anh và phục vụ cho các nghiên cứu trên thị trường tài chính quốc tế, thì bộ 5.126 câu từ nghiên cứu này được thiết kế riêng cho bối cảnh thị trường chứng khoán Việt Nam—một thị trường mới nổi với nhiều đặc thù về ngôn ngữ, tâm lý nhà đầu tư và hành vi giao dịch của khối ngoại. Sự khác biệt này giúp bộ dữ liệu trở thành một công cụ quan trọng để phát triển các mô hình phân tích cảm xúc tài chính tiếng Việt, tương tự như cách Financial PhraseBank đã hỗ trợ sự phát triển của FiinBERT trong phân tích tài chính quốc tế.

Có lẽ đây là một trong những nghiên cứu đầu tiên trong lĩnh vực tài chính tại Việt Nam đi theo hướng phát triển cơ sở dữ liệu cảm xúc tài chính chuẩn hóa, thay vì chỉ dừng lại ở việc xác nhận mối quan hệ giữa hành vi khối ngoại và thị trường. Thay vì chỉ cung cấp bằng chứng thực nghiệm, nghiên cứu này tiến xa hơn bằng cách hệ thống hóa tri thức, xây dựng dữ liệu có thể ứng dụng vào thực tế, góp phần tạo nền tảng cho các nghiên cứu về phân tích cảm xúc tài chính bằng AI tại Việt Nam. Theo đó, nghiên cứu này mang lại ba đóng góp quan trọng. *Thứ nhất*, nghiên cứu phát triển bộ dữ liệu cảm xúc tài chính đầu tiên cho thị trường Việt Nam, giúp lấp đầy khoảng trống trong nghiên cứu về tâm lý nhà đầu tư khối ngoại. *Thứ hai*, nghiên cứu đề xuất một quy trình gán nhãn cảm xúc tài chính chuẩn hóa, có thể mở rộng và áp dụng vào các mô hình phân tích tài chính tự động, góp phần nâng cao độ chính xác của các hệ thống đánh giá tâm lý thị trường. *Thứ ba*, nghiên cứu này đặt nền móng cho sự phát triển của các mô hình AI và NLP chuyên biệt cho tài chính tại Việt Nam, giúp cải thiện khả năng phân tích dữ liệu và hỗ trợ ra quyết định đầu tư dựa trên thông tin từ tin tức tài chính. Nói cách khác, bộ dữ liệu 5.126 câu không chỉ có ý nghĩa như một tập hợp dữ liệu thuần túy, mà còn mở ra cơ hội tối ưu hóa các mô hình xử lý ngôn ngữ tự nhiên trong tài chính, đồng thời cung cấp một công cụ hữu ích cho cả các nhà nghiên cứu, nhà đầu tư và nhà hoạch định chính sách. Với cách tiếp cận này, nghiên cứu không chỉ đóng góp về mặt học thuật mà còn có tiềm năng ứng dụng vào thực tiễn.

Phần còn lại của bài viết được tổ chức như sau: Phần 2 trình bày lý thuyết nền, cung cấp cơ sở học thuật và các nghiên cứu liên quan. Phần 3 mô tả quy trình nghiên cứu được đề xuất, làm rõ sự kế thừa từ các nghiên cứu trước và phân tích những ưu điểm nổi bật của quy trình này. Phần 4 trình bày chi tiết sản phẩm nghiên cứu, đồng thời thảo luận về độ tin cậy của kết quả. Phần 5 đưa ra kết luận và gợi mở các hàm ý nghiên cứu trong tương lai.

LÝ THUYẾT NỀN

Lý thuyết tài chính cổ điển ban đầu không thừa nhận vai trò của thông tin trong việc tác động đến thị trường tài chính. Theo quan điểm này, các nhà đầu tư được cho là hành động hợp lý, và sự cạnh tranh giữa họ sẽ dẫn đến trạng thái cân bằng, trong đó giá cổ phiếu phản ánh đầy đủ giá trị hiện tại của dòng tiền kỳ vọng, còn lợi nhuận kỳ vọng chỉ phụ thuộc vào mức độ rủi ro hệ thống. Tuy nhiên, lý thuyết thị trường hiệu quả sau đó đã điều chỉnh quan điểm này, công nhận rằng thông tin có tác động đến thị trường tài chính, nhưng vẫn dựa trên giả định rằng tất cả nhà đầu tư tiếp cận thông tin như nhau—một giả định thiếu thực tế. Lý thuyết thông tin bất cân xứng đã bác bỏ giả định này, chỉ ra rằng thị trường không chỉ chịu tác động của thông tin, mà còn tồn tại sự chênh lệch trong khả năng tiếp cận và xử lý thông tin giữa các nhà đầu tư. Một số nhà đầu tư có lợi thế khi tiếp cận thông tin sớm hơn hoặc có khả năng phân tích thông tin tốt hơn, trong khi số khác nhận được thông tin muộn hơn hoặc không đủ năng lực xử lý thông tin. Sự bất cân xứng này dẫn đến hiện tượng định giá sai, gây ra những biến động khó lường trên thị trường.

Khi các lý thuyết tài chính dần khẳng định rằng thông tin có tác động đáng kể đến thị trường, mối quan hệ giữa định giá sai và thông tin trở thành chủ đề nghiên

cứu quan trọng, đặt nền móng cho sự phát triển của tài chính hành vi. Trọng tâm của dòng lý thuyết này là giải thích cách nhà đầu tư đưa ra quyết định trong bối cảnh bị ảnh hưởng bởi nhiều yếu tố, trong đó thông tin đóng vai trò then chốt. Các nghiên cứu tài chính hành vi không chỉ xác nhận rằng thông tin ảnh hưởng đến hành vi nhà đầu tư mà còn chỉ ra rằng nhà đầu tư có thể phản ứng quá mức hoặc phản ứng không đủ mức trước thông tin, dẫn đến các biến động đơn lẻ hoặc mang tính hệ thống trên thị trường.

Một nhánh quan trọng của tài chính hành vi tập trung vào cảm xúc từ tin tức tài chính. Các nghiên cứu thực nghiệm đã chứng minh rằng thông tin tích cực hoặc tiêu cực trên các phương tiện truyền thông và mạng xã hội có thể tác động đến hành vi của nhà đầu tư, ảnh hưởng đến quyết định giao dịch và từ đó làm thay đổi giá cả trên thị trường. Hiện tượng này thường được gọi là tác động cảm xúc của tin tức. Tuy nhiên, một câu hỏi mới được đặt ra: Liệu bản thân tin tức có thực sự mang cảm xúc hay không? Trên thực tế, nhiều bản tin không chứa đựng yếu tố cảm xúc, mà chỉ truyền tải thông tin khách quan về thị trường. Tuy nhiên, ngay cả những tin tức trung tính này vẫn có tác động đến hành vi nhà đầu tư thông qua cách mà nhà đầu tư tiếp nhận và diễn giải thông tin. Từ đây, một hướng nghiên cứu mới đã ra đời, đưa ra khái niệm "cảm xúc của nhà đầu tư hình thành từ tin tức".

Điều quan trọng là có sự khác biệt giữa cảm xúc từ tin tức và cảm xúc của nhà đầu tư hình thành từ tin tức. Nhiều nghiên cứu trước đây đã có sự nhầm lẫn giữa hai khái niệm này, dẫn đến việc sử dụng phương pháp đo lường không phù hợp hoặc chọn sai biến đại diện, làm ảnh hưởng đến độ chính xác của kết quả nghiên cứu. Chính vì vậy, sau khi có một loạt các nghiên cứu chứng minh rằng cảm xúc từ tin tức ảnh hưởng đến thị trường, một dòng nghiên cứu khác đã xuất hiện, tập trung vào cảm xúc của nhà đầu tư hình thành từ tin tức, cảm xúc của nhà đầu tư phản ánh từ tin tức chứ không phải cảm xúc trực tiếp trong tin tức. Đây cũng chính là định hướng nghiên cứu của chúng tôi.

Ngoài ra, bên cạnh ảnh hưởng của thông tin, một số nghiên cứu cũng đã xem xét tác động của hành vi khối ngoại đến thị trường chứng khoán nội địa. Gần đây nhất, Dương Ngân Hà đã cung cấp bằng chứng thực nghiệm cho thấy tồn tại mối quan hệ một chiều giữa khối lượng giao dịch ròng của khối ngoại đến khối lượng giao dịch ròng của khối tự doanh trong nước[5]. Điều này cho thấy khối ngoại không chỉ tác động đến giá cổ phiếu hay tâm lý nhà đầu tư cá nhân mà còn ảnh hưởng đến quyết định giao dịch của khối tự doanh trong nước. Đây là một phát hiện quan trọng, củng cố thêm quan điểm rằng hành vi khối ngoại có thể đóng vai trò dẫn dắt trên thị trường chứng khoán Việt Nam, tác động đến cách mà các tổ chức và cá nhân nội địa phản ứng với thông tin.

Dựa trên nền tảng lý thuyết này, chúng tôi hướng thiết kế quy trình nghiên cứu lấy trọng tâm vào việc một bản tin có thể mang nhiều cung bậc cảm xúc khác nhau, tùy thuộc vào bối cảnh xuất hiện, đối tượng tiếp nhận và thị trường mà bản tin đó hướng đến. Các nghiên cứu trước đây thường tự động gán nhãn cảm xúc cho thông tin dựa trên sự xuất hiện của các từ ngữ có tính cảm xúc, chẳng hạn như "sụt giảm" hay "leo thang". Tuy nhiên, cách tiếp cận này có nhiều hạn chế, vì cùng một từ có thể mang ý nghĩa khác nhau trong các ngữ cảnh khác nhau. Nhờ sự phát triển của công nghệ phân tích cảm xúc, các nghiên cứu gần đây đã khắc phục hạn chế này bằng cách lượng hóa mức độ cảm xúc thay vì chỉ gán nhãn định danh. Thay vì chỉ xác định một bản tin là tích cực hay tiêu cực, phương pháp này đo lường mức độ cảm xúc theo thang điểm liên tục, ví dụ, một hệ số cảm xúc gần 0 thể hiện mức độ tiêu cực mạnh, trong khi một hệ số gần 1 biểu thị mức độ tích cực cao. Dựa trên những phát hiện này, nghiên cứu của chúng tôi tiếp tục phát triển một phương pháp gán nhãn cảm xúc phù hợp với thị trường Việt Nam, đồng thời chuyển hóa thông tin tài chính thành một cơ sở dữ liệu có hệ thống, giúp phục vụ các nghiên cứu tài chính hành vi và ứng dụng trong phân tích thị trường.

PHƯƠNG PHÁP NGHIÊN CỨU

Cơ sở dữ liệu

Nghiên cứu này sử dụng một tập dữ liệu lớn gồm 4.065 bản tin được chọn lọc từ tổng số 36.461 bản tin công bố trên các Trang thông tin chính của thị trường chứng khoán Việt Nam trong năm 2024, gồm Vnexpress, Vneconomy. Các bản tin này chứa các thông điệp chính thức về hành vi của nhà đầu tư nước ngoài, phản ánh các giao dịch và tác động của khối ngoại đối với thị trường. Năm 2024 được lựa chọn vì đây năm thời điểm tiêu biểu để phân tích hành vi của nhà đầu tư nước ngoài, khi thị trường chứng khoán Việt Nam chứng kiến nhiều biến động mạnh mẽ do tác động của dòng vốn ngoại.

Quy trình phân tích

Nghiên cứu này xây dựng quy trình 4 bước để trích xuất và phân tích hành vi đầu tư của nhà đầu tư nước ngoài từ tin tức tài chính tại Việt Nam. Kết quả của mỗi bước là đầu vào cho bước tiếp theo, đảm bảo tính hệ thống và độ tin cậy cao.

Bước 1: Trích xuất câu thông điệp

Sử dụng thuật toán Regex, nghiên cứu trích xuất 7.130 câu thông điệp từ 4.065 bản tin, phản ánh 7 hành vi

đầu tư của nhà đầu tư nước ngoài (mua, bán, mua ròng, bán ròng, gom, xả, chốt).

Bước 2: Gán nhãn cảm xúc

Các chuyên gia có hơn 20 năm kinh nghiệm trên thị trường chứng khoán Việt Nam trực tiếp gán nhãn cảm xúc cho 7.130 câu, tạo ra Ngân hàng dữ liệu gán nhãn, phản ánh trạng thái cảm xúc của từng câu.

Bước 3: Mã hóa và chuẩn hóa dữ liệu

Ngân hàng dữ liệu từ Bước 2 được mã hóa và chuẩn hóa bằng mô hình nhúng ngữ nghĩa "text-embedding-3-large" của OpenAI, giúp dữ liệu tương thích với các mô hình học máy và học sâu.

Bước 4: Phân loại cảm xúc và kiểm định độ tin cậy

Dữ liệu đã được mã hóa và chuẩn hóa từ Bước 3 được phân loại cảm xúc bằng Logistic Regression, SVM và Random Forest. Mô hình nhúng "text-embedding-3-large" của OpenAI tiếp tục được sử dụng trong giai đoạn này. Đây là bước kiểm tra độ tin cậy: Nếu kết quả phân loại của các mô hình ổn định và có độ chính xác cao, điều đó chứng minh tính đồng nhất và đáng tin cậy của quá trình mã hóa dữ liệu. Kết quả phân loại được đánh giá bằng các chỉ số: Precision, Recall, F1-score, Accuracy và Confusion Matrix, đảm bảo chất lượng đầu ra.

Cuối cùng, quy trình này tạo ra sản phẩm nghiên cứu chính – "Ngân hàng câu cảm xúc tài chính về hành vi của khối ngoại", cung cấp nền tảng quan trọng cho nghiên cứu tài chính hành vi và hỗ trợ phát triển các mô hình phân tích cảm xúc tài chính như FinBERT. Chi tiết thuật toán được trình bày trong Code sources đính kèm.

Căn cứ thiết kế quy trình phân tích

Quy trình nghiên cứu được thiết kế nhằm tối ưu hóa độ chính xác trong nhận diện cảm xúc thị trường, đặc biệt trong bối cảnh phân tích hành vi nhà đầu tư khối ngoại trên thị trường chứng khoán Việt Nam. Hai yếu tố then chốt quyết định chất lượng nghiên cứu bao gồm: phương pháp trích xuất câu thông điệp chính và quy trình gán nhãn dữ liệu thủ công bởi chuyên gia.

Thứ nhất, việc trích xuất câu thông điệp chính thay vì phân tích toàn bộ bản tin giúp giảm tải dữ liệu, tiết kiệm tài nguyên và tăng tốc độ xử lý. Quá trình này dựa trên bộ từ khóa đặc trưng phản ánh hành vi của nhà đầu tư khối ngoại (mua, bán, mua ròng, bán ròng, gom, xả, chốt), được lựa chọn bởi chuyên gia tài chính để đảm bảo tính chính xác và đại diện cao. Các câu thông điệp sau khi trích xuất được mã hóa bằng mô hình "text-embedding-3-large" của OpenAI, một công cụ có hiệu suất cao trên các bài toán phân tích ngữ nghĩa. Theo OpenAI, mô hình này vượt trội so với các phiên bản trước, với điểm trung bình hiệu suất trên tập dữ liệu thực nghiệm cao hơn từ 23,4% đến 54,9%, giúp tối ưu hóa khả năng nhận diện sắc thái ngữ nghĩa trong tin tức tài chính[6].

Thứ hai, độ chính xác của phân tích cảm xúc phụ thuộc lớn vào chất lượng nhãn dữ liệu. Thay vì sử dụng phương pháp gán nhãn tự động bằng AI hay từ điển cảm xúc, nghiên cứu này áp dụng phương pháp gán nhãn thủ công, do các chuyên gia tài chính có hơn 20 năm kinh nghiệm trên thị trường Việt Nam thực hiện. Cách tiếp cận này giúp phản ánh đúng ngữ cảnh tài chính thực tế, hạn chế sai lệch do cách diễn giải máy móc của mô hình tự động. Nhiều nghiên cứu đã chứng minh rằng gán nhãn thủ công vượt trội hơn các phương pháp tự động. Van Atteveldt và cộng sự[7] so sánh bốn phương pháp gán nhãn (thủ công, gán nhãn đám đông, từ điển tự động, mô hình máy học) và kết luận rằng dữ liệu do chuyên gia gán nhãn đạt hiệu suất cao nhất, trong khi từ điển cảm xúc tự động có độ tin cậy thấp nhất. Nghiên cứu của Chen và cộng sự[8] cũng nhấn mạnh rằng gán nhãn thủ công là tiêu chuẩn bắt buộc trong phân tích cảm xúc tài chính cấp sự kiện, do các nhãn tự động thường chứa nhiều nhiễu. Hayaty & Pratama[9] chứng minh rằng mô hình LSTM đạt 80% độ chính xác khi sử dụng nhãn thủ công, trong khi nhãn tự động từ từ điển như VADER, AFINN chỉ đạt 54–56%, SentiWordNet đạt 49% và từ điển Liu&Hu chỉ 26%. Điều này khẳng định rằng các mô hình học máy chỉ có thể phát huy tối đa hiệu quả khi được huấn luyện trên dữ liệu có chất lượng cao, tức dữ liệu gán nhãn thủ công. Ngoài độ chính xác, tính nhất quán của nhãn dữ liệu cũng quyết định độ tin cậy của phân tích. Chen và cộng sự[8] cảnh báo rằng nhãn tự động nếu không được kiểm định kỹ có thể gây sai lệch toàn bộ kết quả, đặc biệt khi sử dụng từ điển cảm xúc tổng quát trong tài chính. Chẳng hạn, từ "đáo hạn" trong tài chính là trung tính nhưng có thể bị hiểu sai trong các lĩnh vực khác. Do đó, nghiên cứu này ưu tiên phương pháp gán nhãn thủ công, đảm bảo tính chính xác, nhất quán và phản ánh sát thực tế thị trường chứng khoán Việt Nam. Hơn nữa. phương pháp gán nhãn chuyên gia không chỉ đảm bảo độ chính xác mà còn giúp mô hình học được cách đánh giá và suy nghĩ tương tự như chuyên gia, nâng cao khả năng phản ánh thực tiễn thị trường. Nhờ đó, dữ liệu đầu vào không chỉ được chuẩn hóa mà còn mang tính định hướng cao, giúp mô hình có khả năng dự đoán chính xác hơn trong các tình huống thực tế.

KẾT QUẢ VÀ THẢO LUẬN

Kết quả

Bảng 1 trình bày tóm tắt sản phẩm chính của nghiên cứu này là "Ngân hàng câu cảm xúc tài chính về hành

vi của khối ngoại", một hệ thống dữ liệu gồm 5.126 câu thông điệp được trích xuất từ các bản tin tài chính tại Việt Nam. Các thông điệp này đã được phân loại theo cảm xúc tích cực hoặc tiêu cực, tạo nên một công cụ hữu ích để phân tích cảm xúc trong lĩnh vực tài chính. Cấu trúc Ngân hàng dữ liệu được xây dựng với các trường thông tin chính, trong đó nội dung trọng tâm là các câu thông điệp được trích dẫn trực tiếp từ các bản tin tài chính. Những câu trích dẫn này không chỉ đảm bảo tính chính xác trong phân tích mà còn cung cấp một cơ sở dữ liệu có thể tái sử dụng cho các nghiên cứu tiếp theo. Việc bổ sung thông tin nguồn gốc từ các bản tin gốc đảm bảo tính minh bạch của dữ liệu, đồng thời tạo điều kiện cho việc xác thực và mở rộng phạm vi ứng dụng. Ngân hàng câu cảm xúc này không chỉ là một sản phẩm nghiên cứu, mà còn đóng vai trò như một nền tảng quan trọng cho các ứng dụng phân tích cảm xúc và hành vi đầu tư, đặc biệt trong bối cảnh thị trường tài chính Việt Nam.

Bên cạnh sản phẩm "Ngân hàng câu cảm xúc tài chính về hành vi của khối ngoại", nghiên cứu này còn mang lại những phát hiện giá trị. Một trong số đó là việc áp dụng kỹ thuật nhúng câu (Sentence Embedding) ở cấp độ toàn câu, thay vì sử dụng kỹ thuật nhúng và mã hóa truyền thống cho từng từ riêng lẻ (Word Embedding), đã giúp cải thiện đáng kể hiệu quả phân tích. Sathvik đã khẳng định tính ưu việt của kỹ thuật nhúng câu, khi so sánh với phương pháp truyền thống mã hóa từng từ thành token, như minh họa trong Hình 1 tương ứng [10].

Bên cạnh sản phẩm "Ngân hàng câu cảm xúc tài chính về hành vi của khối ngoại", nghiên cứu này còn có những phát hiện có giá trị. Một trong số đó là việc áp dụng kỹ thuật nhúng câu (Sentence Embedding) cấp độ toàn câu thay vì kỹ thuật nhúng, mã hóa truyền thống áp dụng cho từng từ riêng lẻ (Word Embedding) giúp cải thiện đáng kể kết quả phân tích [10]. Sathvik đã khẳng định điều này, cho thấy có sự tối ưu hơn của việc sử dụng kỹ thuật mã hóa nguyên câu so phương pháp truyền thống mã hóa từng từ thành token (Hình 1).

Theo Sathvik, phương pháp truyền thống yêu cầu nhiều bước tiền xử lý phức tạp, bao gồm: chuyển đổi chữ viết thường (lowercasing), loại bỏ các dữ liệu không quan trọng như dấu chấm, dấu phẩy, tokenization, loại bỏ từ không mang nhiều ngữ nghĩa trong bài (stop words), và quá trình chuyển đổi một từ về dạng nguyên gốc (lemmas) của nó, dựa trên ý nghĩa và ngữ cảnh [10]. Ví dụ: "bán", "bán tháo", "bán ròng" sẽ được chuyển về gốc là "bán". Bức này, tiêu tốn nhiều tài nguyên, thời gian và còn làm "giảm đi mức độ cảm xúc của thông tin". Hành vi "bán ròng" hay "bán tháo" có nhiều cảm xúc hơn so với "bán". Ngược lại, khi sử dụng mô hình "GPT embeddings", như phiên bản mới nhất "text-embedding-3-large" (ra mắt đầu năm 2024), các bước tiền xử lý này không còn cần thiết. GPT có khả năng giữ lại nguyên bản cảm xúc và ngữ nghĩa của các cụm từ như "bán tháo" hay "bán ròng", nhờ được huấn luyện trên lượng dữ liệu ngôn ngữ tự nhiên khổng lồ. Theo Sathvik, toàn bộ quá trình tiền xử lý, nếu cần thiết, đều được thực hiện tự động bên trong GPT, giúp tiết kiệm thời gian và công sức mà vẫn đảm bảo độ chính xác cao [10].

Kỹ thuật nhúng câu (sentence embedding) đã chứng minh tính nhất quán và độ tin cậy thông qua kết quả minh họa trong Hình 1. Hình này thể hiện và so sánh kết quả nhúng hai câu có ý nghĩa tương đồng nhưng khác nhau về cách diễn đạt:

- Câu 1: "Nhà đầu tư nước ngoài đã liên tục bán ròng trên sàn HOSE."
- Câu 2: "Nhà đầu tư khối ngoại đã duy trì xu hướng bán ròng trên thị trường HOSE."

Kết quả cho thấy, mặc dù hai câu xuất phát từ hai bài báo khác nhau và có sự khác biệt về từ ngữ, cách diễn đạt, kỹ thuật nhúng câu vẫn trả về hai vector có độ tương đồng rất cao, gần như trùng khớp. Điều này là minh chứng khẳng định khả năng vượt trội của kỹ thuật nhúng cả câu trong việc nắm bắt ngữ nghĩa, đảm bảo rằng biểu diễn vector phản ánh chính xác nội dung ngữ nghĩa, bất kể sự khác biệt về hình thức ngôn ngữ. Nhờ chất lượng nhúng cao như vậy, khi chuyển dữ liệu đã mã hóa vào huấn luyện, mô hình đạt được độ chính xác cao và nhờ đó nhận diện và phân loại cảm xúc với độ tin cậy cao. Theo đó, nghiên cứu này, nhờ cải tiến trong kỹ thuật nhúng và phân cực cảm xúc, đã khắc phục được các hạn chế trước đó, góp phần nâng cao hiệu quả phân tích ngữ nghĩa trong lĩnh vực tài chính.

Thảo luận kết quả nghiên cứu

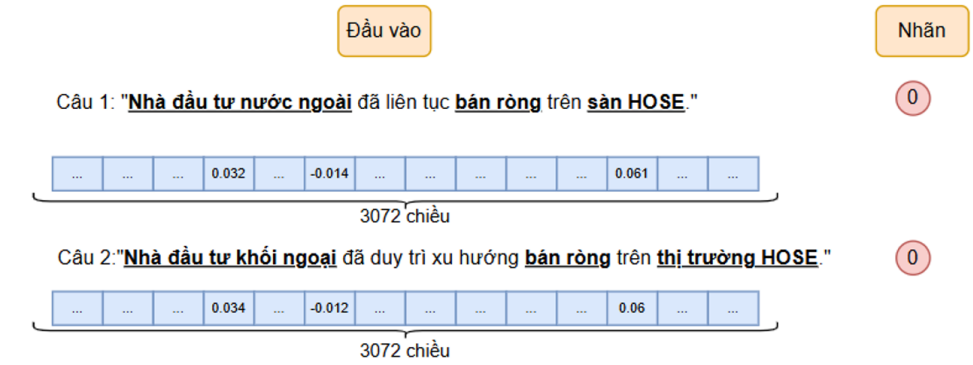
Sản phẩm chính của nghiên cứu, "Ngân hàng câu cảm xúc tài chính về hành vi của khối ngoại, đã được kiểm tra độ tin cậy thông qua việc đánh giá độ chính xác và sự ổn định của các mô hình trong quá trình nhận diện, phân loại cảm xúc. Kết quả cho thấy các mô hình không chỉ đạt hiệu suất cao mà còn thể hiện sự ổn định vượt trội, là minh chứng cho tính vững chắc của phương pháp nghiên cứu mà chúng tôi đề xuất.

Bảng 2 cung cấp kết qua của cả ba mô hình, trong đó, Logistic Regression, Random Forest và SVM đều cho kết quả ổn định với độ chính xác dao động từ 87% đến 89%. Trong đó, SVM đạt hiệu suất cao nhất với độ chính xác 89%, khẳng định ưu thế vượt trội trong việc nắm bắt ngữ nghĩa và phân loại chính xác cảm xúc từ

Bảng 1: Ngân hàng câu cảm xúc tài chính về hành vi của khối ngoại

Câu	SVM	Random Forest	Logistic Regression	Nhãn đúng	Trạng thái
ACV hôm nay bị khối ngoại bán ròng mạnh khoảng 70 tỷ đồng; ngoài ra họ cũng bán ròng tại VTP, BSR, TIE, ...	0	0	0	0	Tiêu cực
VEA hôm nay bị khối ngoại bán ròng khoảng 1 tỷ đồng; ngoài ra họ cũng bán ròng tại VTP, UDC, SKV, ...	0	0	0	0	Tiêu cực
Ngược chiều, QNS hôm nay bị khối ngoại bán ròng khoảng 5 tỷ đồng; ngoài ra họ cũng bán ròng tại VTP, VEA, BSR, ...	0	0	0	0	Tiêu cực
QNS hôm nay bị khối ngoại bán ròng mạnh khoảng 15 tỷ đồng; ngoài ra họ cũng bán ròng tại MCH, BSR, NAB, ...	0	0	0	0	Tiêu cực
Sau quãng xã mạnh cuối năm ngoái, dòng vốn ngoại trở lại với cổ phiếu bán lẻ này, thị giá tang tốt lên vùng đỉnh 2 năm	1	1	1	1	Tích cực
Ngoài ra, khối ngoại cũng chi vài vào tỷ đồng để gom ròng HUT, VFS, MST	1	1	1	1	Tích cực
Bên sàn UPCoM, khối ngoại mua ròng 2,5 tỷ đồng.	1	1	1	1	Tích cực
Khối ngoại nối dài nhịp bán ròng, hôm nay ghi nhận 320 tỷ đồng, tập trung vào chứng chỉ quỹ FUEVFVND, DGC, PVS	1	1	1	0	Tiêu cực
Khối ngoại tiếp tục mua ròng gần 60 tỷ đồng, tập trung vào HDB, FPT, MWG …	1	1	1	1	Tích cực
Khối ngoại cũng mua ròng mạnh tại cổ phiếu HPG và PDR với 192 tỷ và 176 tỷ	1	1	1	1	Tích cực
Trên sàn HNX, nhà đầu tư nước ngoài đã mua ròng 5 phiên liên tiếp với tổng khối lượng 2,39 triệu đơn vị, tổng giá trị mua ròng hơn 71 tỷ đồng …	1	1	1	1	Tích cực
Tính từ đầu tháng 7 đến nay, khối ngoại đã mua ròng khoảng 33 triệu cổ phiếu Vinamilk, tương ứng giá trị gần 2.400 tỷ đồng	1	1	1	1	Tích cực

Nguồn: Nhóm tác giả.



Hình 1: Minh chứng sự nhất quán và tin cậy của kỹ thuật nhúng câu[a]

[a] Nguồn: Nhóm tác giả.

văn bản. Những kết quả này không chỉ minh chứng tính ổn định giữa các mô hình mà còn nhấn mạnh giá trị của phương pháp nhúng câu tiên tiến dựa trên "text-embedding-3-large" của OpenAI. SVM, với độ chính xác 89%, là mô hình hiệu quả nhất, vượt trội cả về khả năng phân loại cảm xúc và độ chính xác trong các trạng thái cảm xúc tích cực và tiêu cực. Từ ma trận nhầm lẫn ở Bảng 3, SVM cho thấy tỷ lệ nhận diện đúng lần lượt là 94,71% (cảm xúc tiêu cực) và 84,27% (cảm xúc tích cực).

Kết quả đồng bộ giữa các mô hình với độ chính xác từ 87% đến 89% phản ánh tính vững chắc và tin cậy của phương pháp luận và sản phẩm nghiên cứu. Phương pháp tách câu để phân tích, kết hợp với gán nhãn từ chuyên gia, đã đảm bảo tính nhất quán và chất lượng cao trong từng bước nghiên cứu. Kết quả nghiên cứu đến đây đã khẳng định tính vững chắc và giá trị khoa học của phương pháp. Hiệu suất cao và sự ổn định của các mô hình, đặc biệt là SVM, cho thấy rằng "Ngân hàng câu cảm xúc tài chính về hành vi của khối ngoại là một nền tảng đáng tin cậy, cung cấp dữ liệu chính xác cho phân tích cảm xúc tài chính, đồng thời tạo nên bước tiến quan trọng trong lĩnh vực nghiên cứu và ứng dụng thực tiễn.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết luận

Nghiên cứu này đã thành công trong việc xây dựng "Ngân hàng câu cảm xúc tài chính về hành vi của khối ngoại", một tập dữ liệu gồm 5.126 câu thông điệp được gán nhãn cảm xúc với độ chính xác và tính ổn định cao. Ngân hàng này không chỉ mở ra các tiềm năng ứng dụng trong lĩnh vực phân tích tài chính mà còn khẳng định độ tin cậy thông qua phương pháp nghiên cứu, quy trình gán nhãn và hiệu quả của các mô hình phân tích cảm xúc được áp dụng. Đây là cơ sở quan trọng để phát triển các công cụ phân tích ngôn ngữ tự nhiên dựa trên kiến trúc BERT, hỗ trợ phân tích báo cáo tài chính, tin tức thị trường và tài liệu đầu tư một cách hiệu quả. Độ tin cậy của Ngân hàng cảm xúc được đảm bảo thông qua nhiều yếu tố. Các mô hình phân tích cảm xúc, bao gồm Logistic Regression, Random Forest và SVM, cho thấy độ chính xác ổn định trong khoảng từ 87% đến 89%. Đặc biệt, mô hình SVM đạt hiệu suất cao nhất với độ chính xác 89%, cho khả năng phân loại cảm xúc tốt, với tỷ lệ nhận diện đúng là 94,71% đối với cảm xúc tiêu cực và 84,27% đối với cảm xúc tích cực (Bảng 3). Sự đồng bộ giữa các mô hình, cùng với hiệu quả vượt trội của SVM, nhấn mạnh tính vững chắc của phương pháp nghiên cứu và khả năng áp dụng vào thực tế. Quy trình gán nhãn và nghiên cứu được thiết kế tỉ mỉ để đảm bảo tính chính xác và phù hợp với bối cảnh tài chính. Thay vì phân tích toàn bộ bản tin, nghiên cứu sử dụng phương pháp tách câu để giảm khối lượng dữ liệu, tăng hiệu quả phân tích. Các câu được trích xuất từ bản tin tài chính và được gán nhãn cảm xúc bởi các chuyên gia có kinh nghiệm trên thị trường chứng khoán Việt Nam. Việc gán nhãn thủ công bởi chuyên gia tài chính không chỉ đảm bảo tính chính xác cao mà còn giúp các mô hình học được cách đánh giá và suy nghĩ như chuyên gia tài chính, tăng khả năng phản ánh thực tiễn. Một điểm nổi bật khác là việc sử dụng kỹ thuật nhúng câu (sentence embedding) thay vì nhúng từ (word embedding), giúp giữ lại nguyên vẹn cảm xúc và ngữ nghĩa của thông tin. Công nghệ nhúng "text-embedding-3-large" của OpenAI được áp dụng để chuyển đổi các câu thành vector, đảm bảo ngữ nghĩa và bối cảnh được giữ nguyên, đồng thời đạt hiệu suất cao hơn so với các phương pháp trước đó. Điều này khắc phục hạn

Bảng 2: Kết quả trên tập kiểm tra của tập dữ liệu Ngân hàng câu cảm xúc tài chính về hành vi của khối ngoại

Mô hình	Phân lớp	Precision	Recall	F1-score	Accuracy
Logistic Regression	0 (Negative)	0,91	0,83	0,87	0,87
	1 (Positive)	0,83	0,91	0,87	
Random Forest	0 (Negative)	0,88	0,87	0,87	0,87
	1 (Positive)	0,86	0,88	0,87	
Support Vector Machine	0 (Negative)	0,95	0,84	0,89	0,89
	1 (Positive)	0,85	0,95	0,89	

Nguồn: Nhóm tác giả

Bảng 3: Ma trận nhầm lẫn

Mô hình	True Positive	True Negative	False Positive	False Negative
Logistic Regression	445/534	89/534	45/492	447/492
Random Forest	462/534	72/534	61/492	431/492
SVM	450/534	84/534	26/492	466/492

Nguồn: Nhóm tác giả

chế của việc phân tích cảm xúc hoàn toàn tự động và tăng cường độ chính xác thông qua việc kết hợp ý kiến chuyên gia tài chính. Nghiên cứu đã tạo ra một quy trình có hệ thống, đảm bảo tính kế thừa, linh hoạt và khả năng mở rộng, đóng góp vào việc chuẩn hóa các phương pháp phân tích cảm xúc trong lĩnh vực tài chính. Ngân hàng cảm xúc không chỉ là nền tảng đáng tin cậy để phát triển các mô hình phân tích ngôn ngữ tự nhiên trên kiến trúc BERT bằng tiếng Việt mà còn mở rộng khả năng ứng dụng sang các tác vụ khác như tóm tắt tin tức, dự đoán xu hướng thị trường và đánh giá tác động của thông tin tài chính. Ngoài ra, nghiên cứu còn cung cấp một khuôn khổ phân tích toàn diện để đánh giá động thái của khối ngoại trên thị trường Việt Nam, giúp nhà đầu tư hiểu rõ hơn về động lực thị trường.

Hướng phát triển nghiên cứu

"Ngân hàng câu cảm xúc tài chính về hành vi của khối ngoại" là một nền tảng đáng tin cậy, được xây dựng dựa trên phương pháp nghiên cứu chặt chẽ, quy trình gán nhãn chuyên nghiệp và công nghệ tiên tiến trong xử lý ngôn ngữ tự nhiên. Sự ổn định và chính xác của mô hình không chỉ đóng góp quan trọng trong lĩnh vực phân tích cảm xúc tài chính mà còn đặt nền móng vững chắc cho các nghiên cứu và ứng dụng tương lai, đặc biệt trong bối cảnh các thị trường tài chính mới nổi như Việt Nam. Những đóng góp này không chỉ có ý nghĩa khoa học mà còn mang lại giá trị thực tiễn nhất Mặc dù đạt được nhiều thành tựu quan trọng, nghiên cứu vẫn tồn tại một số hạn chế. Đầu tiên, việc phụ thuộc vào dữ liệu chính thống khiến nghiên cứu chưa bao quát hết các thông tin phi chính thức và hành vi đầu tư dài hạn, điều này có thể làm giảm tính toàn diện của kết quả phân tích. Thêm vào đó, phạm vi nghiên cứu hiện chỉ giới hạn tại Việt Nam và chưa được kiểm chứng ở các thị trường mới nổi khác, dẫn đến việc chưa thể khái quát hóa hoàn toàn các phát hiện. Để khắc phục, các nghiên cứu trong tương lai cần mở rộng phạm vi địa lý, tích hợp thêm nguồn dữ liệu đa dạng và xem xét nhiều hành vi giao dịch khác nhằm tăng tính toàn diện và khả năng ứng dụng. Đồng thời, việc nghiên cứu sâu hơn về ước tính hệ số cảm xúc, thay vì chỉ dừng lại ở gán nhãn cảm xúc, sẽ giúp cải thiện độ chính xác và giá trị thực tiễn của các mô hình phân tích.

DANH MỤC CÁC TỪ VIẾT TẮT

ACV: Tổng công ty Cảng hàng không Việt Nam
API: Application Programming Interface
BERT: Bidirectional Encoder Representations from Transformers
BSR: Công ty Lọc hóa dầu Bình Sơn
DGC: Công ty Cổ phần Tập đoàn Hóa chất Đức Giang
ETF: Exchange-Traded Fund
FinBERT: Financial Bidirectional Encoder Representations from Transformers
FLAIR: Framework for Linguistic Analysis and Information Retrieval
F1-score: F1 Measurement (Chỉ số F1)
FPT: Công ty cổ phần FPT
FUEVFVND: Chứng chỉ quỹ ETF FUEVFVND

GPT: Generative Pre-trained Transformer

HDB: Ngân hàng TMCP Phát triển Thành phố Hồ Chí Minh (HDBank)

HPG: Công ty cổ phần Tập đoàn Hòa Phát

HUT: Công ty cổ phần Tasco

IPO: Initial Public Offering (Phát hành cổ phiếu lần đầu)

LSTM: Long Short-Term Memory (Mạng hồi quy dài-ngắn hạn)

MIRACL: Multilingual Information Retrieval Across a Continuum of Languages

MLP: Multi-Layer Perceptron

MSMARCO: Microsoft MAchine Reading Comprehension Dataset

MTEB: Massive Text Embedding Benchmark

MWG: Công ty cổ phần Đầu tư Thế Giới Di Động

NLP: Natural Language Processing (Xử lý Ngôn ngữ Tự nhiên)

PCA: Principal Component Analysis (Phân tích Thành phần Chính)

PDR: Công ty cổ phần Phát triển Bất động sản Phát Đạt

PVS: Công ty cổ phần Dịch vụ Kỹ thuật Dầu khí Việt Nam

SVM: Support Vector Machine

TextBlob: Thư viện Python để phân tích cảm xúc

UPCoM: Thị trường giao dịch cổ phiếu chưa niêm yết

VADER: Valence Aware Dictionary and sEntiment Reasoner

VEA: Tổng Công ty Máy động lực và máy nông nghiệp Việt Nam

VFS: Công ty Cổ phần Chứng khoán Nhất Việt

VN-Index: Vietnam Index

VTP: Tổng công ty Bưu chính Viettel

XUNG ĐỘT LỢI ÍCH

Tác giả xin cam đoan rằng không có bất kỳ xung đột lợi ích nào trong công bố bài báo này.

ĐÓNG GÓP CỦA TÁC GIẢ

Phạm Thị Thanh Xuân: Chịu trách nhiệm kiểm soát nội dung: Ý tưởng bài nghiên cứu, Phương pháp nghiên cứu, Phân tích kết quả và tổng hợp.

Đặng Anh Như: Chịu trách nhiệm kiểm soát nội dung: Ý tưởng bài nghiên cứu, Phương pháp nghiên cứu, Phân tích kết quả và tổng hợp.

Từ Hà Phúc: Chịu trách nhiệm kiểm soát nội dung: Ý tưởng bài nghiên cứu, Phương pháp nghiên cứu, Phân tích kết quả và tổng hợp.

TÀI LIỆU THAM KHẢO

1. Phan TNT, Bertrand P, Phan HH, Vo XV. The role of investor behavior in emerging stock markets: Evidence from Vietnam. The Quarterly Review of Economics and Finance 2023;87:367–76;Available from: https://doi.org/https://doi.org/10.1016/j.qref.2021.07.001.
2. Nguyen Ngoc Xuan My P, Luu Duc Toan H, Ngoc Xuan My Huynh N, Duc Toan Nguyen L, Kim Cuong T. Empirical evaluation of overconfidence hypothesis among investors – The evidence in Vietnam stock market". Vietnam Economist Annual Meeting (VEAM) 2016, VEAM; 2016;.
3. Nguyen MH, Nguyen HN, Nguyen ND Le. Foreign Investor Trading and Herding Behavior in Vietnam Stock Market. Journal of International Economics and Management 2016:84–95. [4]Malo P, Sinha A, Takala P, Korhonen PJ, Wallenius J. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. CoRR 2013;abs/1307.5336;.
4. Malo P, Sinha A, Takala P, Korhonen PJ, Wallenius J. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. CoRR 2013;abs/1307.5336;.
5. Hà DN. Kiểm định mối quan hệ nhân quả giữa hoạt động giao dịch của công ty chứng khoán, nhà đầu tư nước ngoài và biến động chỉ số VN-Index. Tạp Chí Khoa Học & Đào Tạo Ngân Hàng 2021;133;.
6. OpenAI. New Embedding Models and API Updates 2025;Available from: https://openai.com/index/new-embedding-models-and-api-updates.
7. van Atteveldt W, van der Velden MACG, Boukes M. The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. Commun Methods Meas 2021;15:121–40;Available from: https://doi.org/10.1080/19312458.2020.1869198.
8. Chen T, Zhang Y, Yu G, Zhang D, Zeng L, He Q, et al. EFSA: Towards Event-Level Financial Sentiment Analysis. ArXiv Preprint ArXiv:240408681 2024;.
9. Hayati M, Pratama A. Performance of Lexical Resource and Manual Labeling on Long Short-Term Memory Model for Text Classification. Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika 2023;9:74–84;Available from: https://doi.org/10.26555/jiteki.v9i1.25375.
10. SATHVIK M. Enhancing Machine Learning Algorithms using GPT Embeddings for Binary Classification. IEEE Trans Emerg Top Comput Intell 2023;Available from: https://doi.org/10.36227/techrxiv.22331053.v1.

Science & Technology Development Journal – Economics - Law and Management 2025, ():1-1

Open Access Full Text Article

Research Article

Building a financial sentiment bank on the investment behavior of foreign investors

Xuan T.T. Pham, Nhu A. Dang, Phuc H. Tu*

Use your smartphone to scan this QR code and download this article

ABSTRACT

This study focuses on building a Financial Sentiment PhraseBank on Foreign Investors' Trading Behavior in the Vietnamese Stock Market. 4,065 financial news articles containing information about eight typical trading behaviors, including ``buying, selling, net buying, net selling, accumulating, offloading, and locking in profits,'' were analyzed. The study successfully extracted a database comprising 7,130 key sentences from these articles and assigned sentiment labels to the content. The final product, the Financial Sentiment PhraseBank on Foreign Investors' Trading Behavior, compiles 5,126 sentences with corresponding sentiment labels. This PhraseBank holds significant value for advancing future research, particularly as a resource for developing models and tools for financial news sentiment analysis based on the BERT architecture, similar to the FinBERT tool. Furthermore, the study provides empirical evidence that applying Sentence Embedding techniques at the sentence level, instead of traditional Word Embedding methods for individual words, greatly enhances analytical efficiency and opens new directions for deeper exploration in the future. This research makes a substantial contribution to the body of knowledge on foreign investors' trading behavior and offers valuable tools and datasets to support both scientific research and practical investment activities.

Key words: Financial sentiment, Foreign trading, Foreign investor transactions, Sentiment scoring, Behavioral theory

University of Economics and Law, Ho Chi Minh City, Vietnam National University, Ho Chi Minh City, Viet Nam

Correspondence

Phuc H. Tu, University of Economics and Law, Ho Chi Minh City, Vietnam National University, Ho Chi Minh City, Viet Nam

Email: phucth22406@st.uel.edu.vn

History

- Received: 07-01-2025
- Revised: 11-3-2025
- Accepted: 19-3-2025
- Published Online:

DOI :

Check for updates

VNUHCM PRESS

Cite this article : Pham X T, Dang N A, Tu P H. **Building a financial sentiment bank on the investment behavior of foreign investors** . *Sci. Tech. Dev. J. - Eco. Law Manag.* 2025; ():1-1.