DỰ BÁO CẢM XÚC NHÀ ĐẦU TƯ VIỆT NAM BẰNG PHƯƠNG PHÁP PHÂN TÍCH CẢM XÚC

ThS. Vũ Văn Đức, Trần Thị Minh Ngọc, Nguyễn Tuấn Dương, Nguyễn Khánh Huy, Lưu Ngân Hà, Hoàng Thái Vân Trang
Trường Đại học Ngoại Thương
Tác giả liên hệ: duc.vuvan@ftu.edu.vn

Ngày nhận: 04/5/2024
Ngày nhận bản sửa: 27/5/2024
Ngày duyệt đăng: 19/6/2024

Tóm tắt

Cảm xúc thị trường từ lâu đã là một chủ đề được các nhà nghiên cứu về tài chính hành vi và các nhà đầu tư trên thị trường quan tâm. Trong nghiên cứu này, chúng tôi đề xuất sử dụng mô hình tích hợp CNN-Bi-GRU-LSTM để phân tích cảm xúc của nhà đầu tư trên thị trường chứng khoán Việt Nam thông qua dữ liệu tin tức tiếng Việt. Mô hình CNN-Bi-GRU-LSTM được huấn luyện trên 15.000 bài báo tài chính để phân loại văn bản theo ba nhóm cảm xúc và đạt độ chính xác 74,5%. Nghiên cứu kết luận rằng phương pháp phân tích cảm xúc có khả năng xác định tâm lý thị trường dựa trên dữ liệu tin tức bằng tiếng Việt.

Từ khóa: *Phân tích cảm xúc, xử lý ngôn ngữ tự nhiên, mạng lưới thần kinh.*

Predicting Investor Sentiment in Vietnam Using Sentiment Analysis

MA. Vu Van Duc, Tran Thi Minh Ngoc, Nguyen Tuan Duong, Nguyen Khanh Huy, Luu Ngan Ha, Hoang Thai Van Trang
Foreign Trade University
Corresponding Authors: duc.vuvan@ftu.edu.vn

Abstract

Market sentiment has long been a subject of interest for researchers in behavioral finance and market participants. This study introduces a CNN-Bi-GRU-LSTM integrated model to analyze investor sentiment in the Vietnamese stock market using Vietnamese news data. The CNN-Bi-GRU-LSTM model is trained on a dataset of 15,000 financial news articles to classify text into three sentiment categories. The model achieved an accuracy of 74.5%, demonstrating its effectiveness in extracting sentiment from text. The findings of this study suggest that sentiment analysis methods hold promise in identifying market sentiment based on Vietnamese news data.

Keywords: *Sentiment analysis, natural language processing, neural network.*

Đặt vấn đề

Trong những năm gần đây, phân tích cảm xúc, một lĩnh vực thuộc xử lý ngôn ngữ tự nhiên (Natural language processing - NLP) đang nhận được nhiều sự quan tâm trong giới học thuật. Ở Việt Nam, phân tích cảm xúc đã được ứng dụng để giải thích hành vi của nhà đầu tư, nhưng nghiên cứu về chủ đề này chủ yếu tập trung vào việc xây dựng chỉ số tâm lý toàn diện (Trúc, Phan và nnk., 2021). Mặc dù phổ biến nhưng chỉ số tâm lý không tính đến phản ứng cảm xúc của nhà đầu tư đối với thông tin đại chúng liên quan đến thị trường chứng khoán. Việc sử dụng phân tích văn bản để đánh giá cảm xúc của nhà đầu tư thông qua tin tức đã được nhiều học giả như Nguyễn và nnk. (2015), Renault (2017), Huang và nnk. (2020), Petropoulos & Siakoulis (2021), Liu và nnk. (2023). Tuy nhiên, ở Việt Nam, còn có khoảng trống nghiên cứu về chỉ số cảm xúc tin tức do thiếu mô hình phân tích văn bản được đào tạo bài bản bằng tiếng Việt. Phân tích cảm xúc của nhà đầu tư dựa trên tin tức còn có khả năng dự đoán lợi nhuận của các cổ phiếu cụ thể cũng như chỉ số thị trường (Li, 2020). Ngày nay, với sự bùng nổ của khoa học máy tính, có nhiều kỹ thuật có thể phân tích cảm xúc và đạt được chỉ số tin cậy cao. Một trong những phương pháp tiếp cận phổ biến và tối ưu hiện nay là CNN (Mạng nơ-ron tích chập) và LSTM (Mạng trí nhớ dài hạn định hướng ngắn hạn). Bên cạnh đó, theo Tran (2019), mô hình tích hợp BiGRU (Mạng nơ-ron truy hồi có cổng hai chiều) có thể đạt hiệu quả cao hơn trong phân tích cảm xúc. Mô hình kết hợp CNN-Bi-GRU-LSTM thể hiện khả năng của nó chủ yếu trong việc đánh giá và phân tích cảm xúc. Ở Việt Nam, phương pháp này chưa được sử dụng rộng rãi để nghiên cứu cảm xúc trên thị trường chứng khoán. Các công nghệ phân tích cảm xúc trên thị trường chứng khoán hiện có chưa nhiều, phần lớn sử dụng mô hình CNN hoặc LSTM. Do đó, mục đích của nghiên cứu này là tạo ra một mô hình linh hoạt, có khả năng phân tích cảm xúc của thị trường thông qua việc sử dụng mô hình tích hợp CNN-Bi-GRU-LSTM.

1. Tổng quan nghiên cứu

Trong thị trường tài chính, tâm lý của nhà đầu tư đóng một vai trò quan trọng trong quá trình ra quyết định và thậm chí ảnh hưởng đến kết quả giao dịch của họ (Ackert và Deaves, 2010). Trên thực tế, nhiều nghiên cứu đã được tiến hành để xác định sự ảnh hưởng đến tâm lý nhà đầu tư lên thị trường tài chính được phản ánh thông qua các dữ liệu sẵn có trên thị trường.

Theo Simon và Wiggins III (2001), sự biến động của thị trường thường được các chuyên gia gọi là "thước đo nỗi sợ hãi của nhà đầu tư" và khi các nhà đầu tư trong giai đoạn lo lắng tột độ có thể mang đến những cơ hội lớn để mua chứng khoán trên thị trường (Simon và Wiggins III, 2001). Hơn nữa, nghiên cứu của Ritter và Welch (2002) đã chỉ ra rằng trong các đợt phát hành cổ phiếu lần đầu ra công chúng (IPO), do sự lạc quan quá mức của các nhà đầu tư đã khiến cho giá

IPO thường bị đẩy cao hơn giá trị thực của chúng.

Hiện nay, nhờ sự phát triển của khoa học máy tính và xử lý ngôn ngữ tự nhiên, các nhà nghiên cứu đã có thể đo lường tâm lý nhà đầu tư thông qua dữ liệu văn bản với một lượng dữ liệu lớn. Với việc áp dụng các mô hình NLP, các nhà nghiên cứu đã đạt được độ chính xác cao trong việc dự đoán tâm lý dựa trên các dữ liệu văn bản tài chính. Dương và nnk. (2016) sử dụng mô hình Support Vector Machine và đạt được tỉ lệ chính xác là 67.6% trong việc phân tích cảm xúc của các tiêu đề báo tài chính trong khoảng thời gian 1 năm.

Xử lý ngôn ngữ tự nhiên cũng được sử dụng trong nhiều lĩnh vực như bình luận phim, nhà hàng, khách sạn. Ở Việt Nam, Duyên và nnk. (2015) đã sử dụng mô hình áp dụng thuật toán Naive Bayes để phân tích cảm xúc xuất hiện các đánh giá khách sạn. Họ đạt được độ chính xác trung bình là 70,4%. Quân và nnk. (2017) sử dụng phương pháp deep learning, đề xuất mô hình kết hợp LSTM và CNN, để phân tích tâm lý mua hàng của người tiêu dùng Việt Nam, đạt tỷ lệ chính xác là 59.61% trên bộ dữ liệu VLSP 2016 về các đánh giá sản phẩm bằng tiếng Việt.

2. Phương pháp nghiên cứu

2.1. Thu thập và xử lý dữ liệu

Đối với dữ liệu văn bản, nghiên cứu thu thập phần tin tức ngắn gọn từ 4 trang tin tài chính Việt Nam được công nhận rộng rãi: "baodautu.vn", "vneconomy.vn", "thoibaotaichinhvietnam.vn" và "thoibaokinhte". Nghiên cứu đã sử dụng Python 3.7 để phát triển trình thu thập dữ liệu web trích xuất dữ liệu văn bản từ các bài đăng trên diễn đàn. Bộ dữ liệu bao gồm 15 nghìn bài đăng từ ngày 04/01/2013 đến ngày 02/02/2024.

Nghiên cứu áp dụng bộ công cụ NLP chuyên dụng Underthesea để xử lý văn bản tiếng Việt bằng cách phân đoạn văn bản thành từ riêng biệt một cách hiệu quả và loại bỏ dấu phụ, cũng như chuẩn hóa chính tả. Sử dụng Underthesea trong quy trình tiền xử lý giúp giải quyết các vấn đề cụ thể của tiếng Việt. Nghiên cứu cũng tích hợp kỹ thuật deep learning Word2vec của Google vào để tạo ra các vectơ tương ứng (Mikolov & nnk, 2013).

2.2. Chú thích dữ liệu

Để chú thích dữ liệu của nghiên cứu này, văn bản tiếng Việt được phân loại một cách có phương pháp thành các loại cảm xúc riêng biệt, bao gồm thái độ tích cực, trung lập và tiêu cực. Văn bản được phân loại và mã hóa chỉ số cảm xúc như sau:

Bảng 1. Chú thích dữ liệu

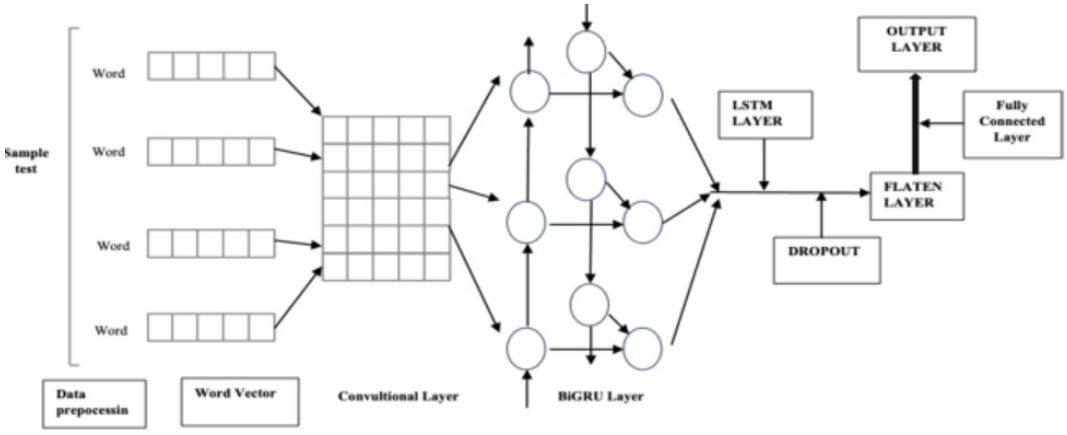
Câu	Mã hóa cảm xúc[1]
Thị trường đã có được một tuần bùng nổ thành công	1
Thị trường đã có phiên giảm điểm thứ hai trong tuần	-1
Thị trường có tăng điểm nhưng chưa khả quan	0

[1]: Cảm xúc tích cực, 0: Cảm xúc trung lập, -1: Cảm xúc tiêu cực.

2.3. Mã hoá One-hot

Mã hoá One-hot là kỹ thuật chuyển đổi nhãn chỉ số cảm xúc thành dạng vectơ. Phương pháp này biến đổi dữ liệu phân loại thành các vectơ nhị phân, mỗi vectơ có chiều dài phụ thuộc vào số lượng lớp và mỗi phần tử trong vectơ tượng trưng cho một lớp cụ thể.

2.4. Mô hình CNN-BiGRU-LSTM



Hình 1. Mô hình tích hợp CNN-BiGRU-LSTM

Mô hình tích hợp CNN-BiGRU-LSTM gồm 9 lớp.

Lớp đầu vào: Trong lớp này, dữ liệu câu đầu vào được làm sạch trước khi phân tách thành các đoạn từ. Sử dụng mã hóa one-hot, kết quả được chuyển đổi thành các vectơ dày đặc có kích thước cố định. Kích thước từ vựng được đặt ở mức 200.000 và thứ nguyên nhúng được đặt là 200 thứ nguyên.

Tiếp theo sau đó, là các lớp ẩn (hidden layer), bao gồm:

Lớp Convolutional: Lớp này trích xuất các tính năng từ các vectơ từ được nhúng. Sau lớp đầu vào, các vectơ từ phải chịu sự tích chập bằng cách sử dụng một tập hợp các hạt tích chập có kích thước h × k.

Lớp mạng Bi-GRU: Lớp đầu vào tạo điều kiện thuận lợi cho việc truyền dữ liệu tới cả GRU chuyển tiếp và GRU ngược.

Lớp mạng LSTM: Khi các chuỗi đầu vào đi qua lớp LSTM, các biểu diễn dữ liệu đầu vào ở cấp độ cao hơn sẽ được hình thành.

Theo sau đó là **lớp Dropout, lớp Flatten, lớp Dense** được áp dụng để chuẩn hóa, định hình lại dữ liệu và trích chọn thuộc tính, giúp đơn giản hóa mô hình.

Lớp được kết nối đầy đủ (Fully Connected Layer): Trong lớp này, các chuỗi kết quả tổng hợp được liên kết với nhau để tạo thành ma trận riêng, được xây dựng theo đặc tả đầu vào Bi-GRU.

Lớp đầu ra: Lớp cuối cùng bao gồm ba đơn vị với kích hoạt softmax, đại diện cho ba loại tình cảm (tiêu cực, trung tính và tích cực). Softmax đảm bảo rằng xác suất đầu ra tổng cộng bằng một, tạo điều kiện cho việc phân loại nhiều lớp.

3. Kết quả và thảo luận

Trong nghiên cứu này, dữ liệu tin tức được thu thập từ năm 2013 đến 2023. Bằng cách áp dụng mô hình kết hợp CNN-Bi-GRU-LSTM để phân tích cảm xúc tại thị trường Việt Nam,

nghiên cứu đã đạt được độ chính xác 74,5%. Mặc dù tỷ lệ chính xác cao hơn có thể được tìm thấy trong nghiên cứu phân tích cảm xúc bằng tiếng Anh khác (Rehman và nnk., 2019, Amrani và nnk., 2018), chúng tôi khẳng định rằng độ chính xác 74,5% là con số đáng kể của một mô hình trong việc trích xuất cảm xúc từ văn bản, được hỗ trợ bởi một số lập luận.

Thứ nhất, trong các tóm tắt bài báo được sử dụng làm dữ liệu đầu vào cho mô hình, sẽ có một lượng dữ liệu nhất định có các tín hiệu gây nhiễu cho mô hình và làm giảm khả năng phân tích cảm xúc trong câu văn.

Bảng 2. Ví dụ cho kết quả phân tích cảm xúc bởi mô hình

Ví dụ	Tóm tắt bài báo	Phân tích cảm xúc bởi con người	Phân tích cảm xúc bởi mô hình
1	Không phải tất cả các cổ phiếu bất động sản hôm nay đều giảm, cũng không phải tất cả bị bán tháo, nhưng hành động xả hàng là rất rõ ràng. Rất nhiều cổ phiếu sụt giảm làm mắc kẹt khối lượng cổ phiếu khổng lồ.	-1	-1
2	Thị trường chứng khoán hôm nay (05/12) duy trì đà tăng ngay từ đầu phiên, nhưng áp lực bán tăng tại ngưỡng kháng cự 1.125 điểm khiến chỉ số giảm điểm dần về cuối phiên.	-1	1

Trong các ví dụ về kết quả phân tích cảm xúc của mô hình, ta có thể thấy rằng mô hình dự đoán chính xác các câu văn có tín hiệu rõ ràng (tích cực, tiêu cực, trung tính). Trong ví dụ 1, khá nhiều từ mang ý nghĩa tiêu cực xuất hiện như "giảm", "bán tháo", "xả hàng", "sụt giảm", "mắc kẹt" khiến cho mô hình có thể dự đoán chính xác cảm xúc trong câu văn là tiêu cực. Tuy nhiên, ở ví dụ 2, khi ở trong câu văn xuất hiện cả hai từ ngữ mang tính tích cực "tăng" và tiêu cực "giảm" đã gây nhiễu cho mô hình và khiến cho mô hình không thể dự đoán chính xác cảm xúc được hàm ý ở trong ví dụ. Từ đó, giảm tỷ lệ dự đoán cảm xúc chính xác của mô hình được đề xuất.

Thứ hai, khi được so sánh với các mô hình cùng sử dụng bộ dữ liệu đầu vào tiếng Việt, mô hình được đề xuất cho thấy sự cải thiện hơn trong việc phân tích cảm xúc cho các câu văn.

Bảng 3. Tỷ lệ chính xác của các mô hình dự đoán cảm xúc bằng tiếng Việt

Mô hình	Độ chính xác
Naive Bayes (Duyên và nnk., 2015)	70,40%
SVM (Dương và nnk., 2016)	67,60%
LSTM+CNN (Quân và nnk., 2017)	59,10%
Bi-GRU+CNN+LSTM (Mô hình đề xuất)	74,5%

So với các mô hình sử dụng cho ngôn ngữ tiếng Việt ở trong Bảng 3 với tỉ lệ chính xác là 70,4% và 67,6%, mô hình của chúng tôi đã cho thấy sự cải thiện trong việc phân tích cảm xúc trong các dữ liệu văn bản tiếng Việt. Đặc biệt là đối với mô hình sử dụng thuật toán Naive Bayes khi cùng sử dụng bộ dữ liệu là các bài báo liên quan đến lĩnh vực tài chính, mô hình được đề xuất đã cải thiện được 4% độ chính xác trong quá trình phân tích cảm xúc trong các dữ liệu văn bản.

4. Kết luận và hướng nghiên cứu trong tương lai

Các nhà nghiên cứu tài chính hành vi từ lâu đã quan tâm đến mối quan hệ giữa những thay đổi trên thị trường chứng khoán và cảm xúc của nhà đầu tư. Ngày nay, vấn đề không còn là liệu cảm xúc của nhà đầu tư có ảnh hưởng đến giá cổ phiếu hay không, thay vào đó, vấn đề là làm thế nào để đo lường và đánh giá tác động của nó (Baker & Wurgler, 2007). Nghiên cứu của chúng tôi đã đề xuất một phương pháp tích hợp kết hợp CNN với Bi-GRU và LSTM cho phân tích cảm xúc dựa trên dữ liệu tiếng Việt.

Kết quả của nghiên cứu đã chỉ ra rằng việc ứng dụng xử lý ngôn ngữ tự nhiên trong phân tích cảm xúc cung cấp một khung phân tích tiềm năng về cảm xúc thị trường dựa trên dữ liệu tin tức bằng tiếng Việt. Bên cạnh đó, cách tiếp cận tích hợp CNN, Bi-GRU và LSTM để xử lý ngôn ngữ tiếng Việt đã giải quyết được những vấn đề tồn đọng của những mô hình độc lập trước đó, phương pháp này cho phép trích xuất các đặc trưng chi tiết từ bộ dữ liệu mở rộng, bao gồm cả thông tin văn bản bằng tiếng Việt. Nghiên cứu cũng đề xuất một hướng đi khả quan trong khâu tiền xử lý dữ liệu tiếng Việt bằng cách chuyển đổi văn bản thành vector để xác định và đo lường tâm lý của nhà đầu tư, từ đó, nâng cao các mô hình dự đoán cho thị trường tài chính.

Với những hạn chế về khả năng đại diện của dữ liệu, bộ dữ liệu có thể không phản ánh đầy đủ sự đa dạng của ngôn ngữ và văn hóa trong tiếng Việt. Điều này có thể dẫn đến việc mô hình không hiểu được một số ngữ cảnh hoặc ngôn ngữ đặc thù. Ở các ngôn ngữ khác, đặc biệt là tiếng Anh, nhiều từ điển cảm xúc đã được xây dựng để cải thiện hiệu quả của các mô hình phân tích cảm xúc. Tại Việt Nam, lĩnh vực phân tích cảm xúc vẫn đang ở giai đoạn đầu, và thiếu một từ điển cảm xúc tiếng Việt đầy đủ, phục vụ riêng cho quá trình ứng dụng mô hình phân tích cảm xúc. Do đó, quá trình tiền xử lý dữ liệu văn bản gặp nhiều khó khăn, chủ yếu do có quá nhiều cụm từ khác nhau thường được sử dụng trong lĩnh vực tài chính hoặc cụ thể hơn là trong bối cảnh thị trường chứng khoán. Chính vì vậy, việc xây dựng một từ điển cảm xúc chuyên dụng cho tiếng Việt có thể đơn giản hóa quá trình chuẩn bị dữ liệu và nâng cao độ chính xác của phân tích cảm xúc bằng cách cung cấp các nguồn tài nguyên ngôn ngữ được tùy chỉnh.

Tài liệu tham khảo

Ackert, L. F., & Deaves, R. (2010). Behavioral finance: Psychology, decision-making, and markets. Mason, OH: South-Western Cengage Learning.

Al Amrani, Y., Lazaar, M., & El Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science, 127,*

511–520. https://doi.org/10.1016/j.procs.2018.01.150.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *The Journal of Economic Perspectives,* 21(2), 129-151.

Duc Duong, Toan Nguyen, & Minh Dang. (2016). Stock market prediction using financial news articles on Ho Chi Minh Stock Exchange. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication (IMCOM '16)* (pp. 1–6). Association for Computing Machinery. https://doi.org/10.1145/2857546.2857619.

Duyên, Nguyễn, Bach, N. X., & Phuong, T. M. (2015). An empirical study on sentiment analysis for Vietnamese. *International Conference on Advanced Technologies for Communications,* 309–314. https://doi.org/10.1109/ATC.2014.7043403.

Huang, A. G., Tan, H., & Wermers, R. (2020). Institutional trading around corporate news: Evidence from textual analysis. *The Review of Financial Studies,* 33(10), 4627-4675.

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing & Management,* 57(5), 102212.

Liu, J., Wu, K., & Zhou, M. (2023). News tone, investor sentiment, and liquidity premium. *International Review of Economics & Finance,* 84, 167-181.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ICLR Workshop.*

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications,* 42(24), 9603-9611.

Petropoulos, A., & Siakoulis, V. (2021). Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique. *Central Bank Review,* 21, 141–153.

Quan, H. V., Huy, T. N., Bac, L., & Minh, L. N. (2017). Multi-channel LSTM-CNN model for Vietnamese sentiment analysis. *9th International Conference on Knowledge and Systems Engineering* (KSE) (pp. 24-29). https://doi.org/10.1109/KSE.2017.8119429.

Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications.* https://doi.org/10.1007/s11042-019-07788-7.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. *Journal of Banking & Finance,* 84, 25-40.

Ritter, J. R., & Welch, I. (2002). A review of IPO activity, pricing, and allocations. *Journal of Finance,* 57(4), 1795-1828.

Simon, D. P., & Wiggins III, R. A. (2001). S&P futures returns and contrary sentiment indicators. *Journal of Futures Markets,* 21(5), 447-462.

Tran, T. U., Hoang, H. T. T., & Huynh, H. X. (2019). Aspect extraction with bidirectional GRU and CRF. IEEE-RIVF *International Conference on Computing and Communication Technologies (RIVF)* (pp. 1-5).

Truc, P., Bertrand, P., Hai, P. H., & Vinh, V. X. (2021). Investor sentiment and stock return: Evidence from Vietnam stock market. T*he Quarterly Review of Economics and Finance,* 87. https://doi.org/10.1016/j.qref.2021.07.001.