	<p><u>Thủ tục:</u> NGHIÊN CỨU KHOA HỌC SINH VIÊN</p>	<p>Mã số: TT/P.QLPTKHCN/13/BM17 Ban hành lần: 06 Ngày hiệu lực:</p>
--	--	--

Xây dựng ứng dụng hỗ trợ đầu tư chứng khoán bằng trí tuệ nhân tạo và phân tích cơ bản

(Nguyễn Quang Huy, Lớp 23H50301, Khoa Công Nghệ Thông Tin)

(Nguyễn Trần Nhật AnHuy, Lớp 23H50301, Khoa Công Nghệ Thông Tin)

I. TÓM TẮT CÔNG TRÌNH

Nghiên cứu này trình bày việc xây dựng và phát triển một hệ thống hỗ trợ đầu tư chứng khoán thông minh, được thiết kế để nâng cao hiệu quả ra quyết định cho nhà đầu tư tại thị trường Việt Nam. Điểm đột phá của đề tài là việc kết hợp các mô hình Trí tuệ nhân tạo (AI) tiên tiến để phân tích đồng thời hai nguồn dữ liệu quan trọng: dữ liệu định lượng (giá lịch sử) và dữ liệu định tính (văn bản tài chính). Cụ thể, hệ thống tích hợp mạng LSTM (Long Short-Term Memory) để dự báo xu hướng giá và mô hình ngôn ngữ PhoBERT để phân tích, lượng hóa cảm xúc thị trường từ tin tức. Bằng cách tiếp cận đa chiều, nghiên cứu hướng tới việc cung cấp một công cụ phân tích sâu sắc, giúp nhà đầu tư giảm thiểu các quyết định cảm tính và xây dựng chiến lược dựa trên dữ liệu dự đoán.

II. QUÁ TRÌNH NGHIÊN CỨU VÀ KẾT QUẢ


2.1. Mở đầu

2.1.1. Mở đầu

Thị trường chứng khoán Việt Nam đang phát triển mạnh mẽ, thu hút một lượng lớn nhà đầu tư mới. Tuy nhiên, phần lớn các nhà đầu tư này thiếu kinh nghiệm, thường đưa ra quyết định cảm tính, gây rủi ro cho chính họ và sự ổn định của thị trường. Trí tuệ nhân tạo (AI) nổi lên như một công cụ tiềm năng, có khả năng xử lý dữ liệu lớn, phân tích các mẫu phức tạp và dự báo xu hướng với độ chính xác cao, giúp nhà đầu tư tiếp cận thị trường một cách khoa học hơn.

2.1.2. Lý do chọn đề tài

Việc ứng dụng AI vào đầu tư chứng khoán tại Việt Nam còn sơ khai, đa phần các công cụ chỉ tập trung vào phân tích kỹ thuật đơn thuần mà bỏ qua các yếu tố

	<p><u>Thủ tục:</u> NGHIÊN CỨU KHOA HỌC SINH VIÊN</p>	<p>Mã số: TT/P.QLPTKHCN/13/BM17 Ban hành lần: 06 Ngày hiệu lực:</p>
--	--	---

phân tích cơ bản và cảm xúc thị trường . Đề tài được thực hiện nhằm xây dựng một giải pháp toàn diện, kết hợp sức mạnh của AI với chiều sâu của phân tích cơ bản, cung cấp cho nhà đầu tư, đặc biệt là nhà đầu tư cá nhân, một cơ sở vững chắc hơn để ra quyết định .

2.1.3. Đối tượng và phạm vi nghiên cứu

Đối tượng: Các mô hình AI trong dự báo xu hướng (LSTM, PhoBERT) và các bài báo, báo cáo tài chính ảnh hưởng đến quyết định của nhà đầu tư .

Phạm vi: Nghiên cứu tập trung vào các mã cổ phiếu trong rổ VN30 trên thị trường chứng khoán Việt Nam, với dữ liệu được thu thập từ năm 2010 đến nay để đảm bảo tính bao quát và cập nhật.


2.2. Cơ sở lý thuyết

Nghiên cứu được xây dựng trên các nền tảng lý thuyết chính. Về tài chính, đề tài áp dụng Lý thuyết nghiên cứu sự kiện (Event Study) và Mô hình định giá tài sản vốn (CAPM) để tính toán hệ số Alpha, một thước đo lợi nhuận bất thường dùng để gán nhãn dữ liệu một cách khách quan . Về học máy, phương pháp Giám sát Yếu (Weak Supervision) với framework Snorkel được sử dụng để tự động tạo bộ dữ liệu huấn luyện lớn từ nhiều quy tắc (hàm gán nhãn) không hoàn hảo . Về mô hình, đề tài sử dụng mạng nơ-ron LSTM do khả năng vượt trội trong việc mô hình hóa dữ liệu chuỗi thời gian, và PhoBERT, mô hình ngôn ngữ tối ưu cho tiếng Việt, để thực hiện phân tích cảm xúc.

2.3. Phương pháp đề xuất

2.3.1. Kiến trúc hệ thống

Hệ thống có kiến trúc kết hợp hai luồng xử lý song song. Luồng thứ nhất, dữ liệu văn bản (tin tức) được đưa vào mô hình PhoBERT đã được tinh chỉnh để trích xuất đặc trưng cảm xúc. Luồng thứ hai, dữ liệu giá lịch sử được xử lý. Cuối cùng, đặc trưng cảm xúc được gộp với dữ liệu giá theo ngày và đưa vào mô hình LSTM để dự đoán giá đóng cửa của ngày tiếp theo .

	<p><u>Thủ tục:</u> NGHIÊN CỨU KHOA HỌC SINH VIÊN</p>	<p>Mã số: TT/P.QLPTKHCN/13/BM17 Ban hành lần: 06 Ngày hiệu lực:</p>
--	--	---

2.3.2. Tiền xử lý dữ liệu

Dữ liệu văn bản: Dữ liệu được thu thập từ các trang báo tài chính (CafeF, Báo Đầu Tư) bằng kỹ thuật crawling đa luồng. Sau đó, các mã cổ phiếu được nhận dạng bằng Regex, và ngữ cảnh (câu chứa mã và các câu lân cận) được trích xuất. Văn bản được tách từ bằng VnCoreNLP để chuẩn bị cho mô hình PhoBERT.

Dữ liệu giá: Dữ liệu giá lịch sử của các cổ phiếu VN30 được tải về thông qua thư viện vnstock. Dữ liệu được làm sạch, xử lý giá trị thiếu và chuẩn hóa bằng Min-Max Scaling về khoảng $[0,1]$.

Gán nhãn cảm xúc: Áp dụng phương pháp Giám sát Yếu, trong đó hệ số Alpha được dùng làm Hàm Gán Nhãn (LF) cốt lõi để xác định một tin tức là Tích cực ($\alpha > 0.01$), Tiêu cực ($\alpha < -0.01$) hay Trung tính. Các LF khác dựa trên từ khóa tài chính cũng được sử dụng để tăng cường chất lượng nhãn.

2.3.3. Huấn luyện mô hình

PhoBERT: Mô hình PhoBERT-base được tinh chỉnh (fine-tuned) trên bộ dữ liệu tin tức đã được gán nhãn cảm xúc để thực hiện bài toán phân loại ba lớp (Tích cực, Tiêu cực, Trung tính).

LSTM: Một mô hình LSTM nhiều lớp được xây dựng để dự đoán giá cổ phiếu. Mô hình được huấn luyện bằng hàm mất mát Mean Squared Error (MSE) và bộ tối ưu hóa Adam, với chiến lược dừng sớm (EarlyStopping) để tránh overfitting.

2.3.4. Gộp dữ liệu:

Chỉ số cảm xúc (sentiment score) đầu ra từ mô hình PhoBERT được đồng bộ hóa theo ngày với dữ liệu giá lịch sử. Cột dữ liệu cảm xúc mới này được thêm vào làm một đặc trưng đầu vào (feature) cho mô hình LSTM, tạo thành một bộ dữ liệu đa chiều hoàn chỉnh.

2.4 Kết quả thực nghiệm

2.4.1. Mô tả dữ liệu

Dữ liệu văn bản: Thu thập được khoảng 80.000 bài báo từ CafeF (89.7%) và Báo Đầu Tư (10.3%) . Sau khi gán nhãn và chia tách, tập huấn luyện có phân bố nhãn khoảng: 38% Tích cực, 31.6% Tiêu cực, và 30.4% Trung tính.

Dữ liệu giá: Thử nghiệm dự báo được thực hiện trên mã cổ phiếu VCB trong giai đoạn từ 03/06/2024 đến 26/05/2025. Hai bộ dữ liệu được tạo ra để so sánh: một bộ chỉ có dữ liệu giá và một bộ có kết hợp thêm thông tin cảm xúc.

2.4.2. Kết quả sentiment analysis

Mô hình PhoBERT sau khi tinh chỉnh đạt độ chính xác (Accuracy) là **0.55** trên tập kiểm thử. Chỉ số F1-Score cho thấy mô hình nhận diện tốt nhất ở nhãn "Tích cực" (0.610), và yếu hơn ở nhãn "Tiêu cực" (0.550) và "Trung tính" (0.450) . Kết quả này cho thấy sự phức tạp và mơ hồ của các văn bản tài chính, đặc biệt là các tin tức trung tính .

2.4.3. Kết quả dự báo giá VCB

Kết quả so sánh hai mô hình LSTM cho thấy sự cải thiện vượt trội khi tích hợp dữ liệu cảm xúc. Cụ thể:

- RMSE (Root Mean Square Error) giảm từ 0.917 xuống còn 0.263.
- MAE (Mean Absolute Error) giảm từ 0.791 xuống còn 0.215.

Mô hình có cảm xúc dự đoán bám sát hơn với đường giá thực tế, chứng tỏ việc tích hợp phân tích cảm xúc giúp mô hình nắm bắt tốt hơn các biến động của thị trường.

2.4.4. Ứng dụng thực tế


Hệ thống được đề xuất có thể được phát triển thành một ứng dụng hoàn chỉnh, cung cấp cho các nhà đầu tư cá nhân một công cụ hỗ trợ ra quyết định mạnh mẽ. Bằng cách cung cấp các dự báo dựa trên cả dữ liệu kỹ thuật và tâm lý thị trường, ứng dụng giúp giảm thiểu giao dịch cảm tính và nâng cao hiệu quả đầu tư.

III. KẾT LUẬN

Công trình đã xây dựng thành công một hệ thống dự báo giá cổ phiếu bằng cách kết hợp hiệu quả mô hình LSTM với dữ liệu cảm xúc trích xuất từ mô hình PhoBERT. Kết quả thực nghiệm đã chứng minh rõ ràng rằng việc bổ sung yếu tố cảm xúc thị trường vào mô hình dự báo giúp cải thiện đáng kể độ chính xác so với việc chỉ sử dụng dữ liệu giá lịch sử. Hướng phát triển trong tương lai bao gồm việc cải thiện phương pháp gán nhãn để nâng cao hiệu suất của mô hình phân tích cảm xúc, đặc biệt với các văn bản trung tính, và mở rộng hệ thống để áp dụng cho nhiều mã cổ phiếu hơn.

IV. TÀI LIỆU THAM KHẢO

- [1] Context-Aware Legal Citation Recommendation using Deep Learning, https://www.researchgate.net/publication/352642553_Context-Aware_Legal_Citation_Recommendation_using_Deep_Learning
- [2] Hệ số Alpha trong đầu tư chứng khoán, <https://pinetree.vn/post/20220313/he-so-alpha-trong-dau-tu-chung-khoan/>
- [4] Ultimate Guide Weak Supervision Data Mining, <https://www.numberanalytics.com/blog/ultimate-guide-weak-supervision-data-mining>
- [5] Weak Supervision: The End of Hand-Labeled Data? Ratner, A. và các cộng sự, <https://ajratner.github.io/assets/papers/ratner-sigmoddemo17.pdf>
- [6] Lãi suất phi rủi ro là gì? HSC Stock Insight, <https://stockinsight.hsc.com.vn/lai-suat-phi-rui-ro-la-gi/>
- [7] Vietnamese Sentiment Analysis. Lingvanex, <https://lingvanex.com/services/vietnamese-sentiment-analysis/>
- [8] Sentiment Analysis. MPT, <https://mpt.com.vn/sentiment-analysis/>
- [9] Phân tích cảm xúc trên thị trường chứng khoán. Tài chính doanh nghiệp, <https://taichinhdoanhnghiep.net.vn/phan-tich-cam-xuc-tren-thi-truong-truong-chung-khoan-d44822.html>

	<p><u>Thủ tục:</u> NGHIÊN CỨU KHOA HỌC SINH VIÊN</p>	<p>Mã số: TT/P.QLPTKHCN/13/BM17 Ban hành lần: 06 Ngày hiệu lực:</p>
--	--	--

- [10] Alpha - Corporate Finance Institute. Corporate Finance Institute,
<https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/alpha/>
- [11] Bộ nhớ dài-ngắn hạn. Wikipedia, [https://en.wikipedia.org/wiki/Long_short-term_memory#:~:text=March%202022\),and%20other%20sequence%20learning%20methods](https://en.wikipedia.org/wiki/Long_short-term_memory#:~:text=March%202022),and%20other%20sequence%20learning%20methods)
- [12] Hồi quy tuyến tính trong Machine Learning Nguyen Duong. Linear Regression - Hồi quy tuyến tính trong Machine Learning, <https://viblo.asia/p/hoi-quy-tuyen-tinh-trong-machine-learning-1VgZvaw7KAw>
- [13] Predictive Modeling of Stock Prices Using Transformer Model Mozaffari, L. and Zhang, J. 2024. Predictive Modeling of Stock Prices Using Transformer Model. ICMLT 2024, Oslo, Norway. ACM Digital Library, <https://dl.acm.org/doi/fullHtml/10.1145/3674029.3674037>
- [14] Multioutput Regression in Machine Learning, <https://www.geeksforgeeks.org/multioutput-regression-in-machine-learning/>
- [15] TF-IDF và vai trò của TF-IDF trong SEO, <https://vietmoz.edu.vn/tf-idf/>