ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

KHOA TÀI CHÍNH – NGÂN HÀNG

KHÓA LUẬN TỐT NGHIỆP

PHÂN TÍCH THỰC NGHIỆM TÁC ĐỘNG CỦA THÔNG TIN CƠ BẢN, KỸ THUẬT VÀ TÂM LÝ ĐẾN DỰ BÁO XU HƯỚNG CỔ PHIẾU THEO NGÀNH: ỨNG DỤNG MẠNG NƠ-RON ĐA LỚP (MLP)

GVHD: ThS. Ngô Phú Thanh

SVTH: Bùi Thị Ngọc Hà

MSSV: K224141657

TP.HCM, THÁNG 3/2026

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

KHOA TÀI CHÍNH – NGÂN HÀNG

KHÓA LUẬN TỐT NGHIỆP

PHÂN TÍCH THỰC NGHIỆM TÁC ĐỘNG CỦA THÔNG TIN CƠ BẢN, KỸ THUẬT VÀ TÂM LÝ ĐẾN DỰ BÁO XU HƯỚNG CỔ PHIẾU THEO NGÀNH: ỨNG DỤNG MẠNG NƠ-RON ĐA LỚP (MLP)

GVHD: ThS. Ngô Phú Thanh

SVTH: Bùi Thị Ngọc Hà

MSSV: K224141657

TP.HCM, THÁNG 3/2026

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời tri ân sâu sắc và chân thành nhất đến Thạc sĩ Ngô Phú Thanh. Thầy không chỉ là người hướng dẫn khoa học mà còn là người đã luôn đồng hành, sát cánh và truyền cảm hứng cho em trong suốt quá trình thực hiện khóa luận tốt nghiệp. Đề tài "Phân tích thực nghiệm tác động của thông tin cơ bản, kỹ thuật và tâm lý đến dự báo xu hướng cổ phiếu theo ngành: Ứng dụng mạng nơ-ron đa lớp (MLP)" chính là quả ngọt từ sự nỗ lực nghiên cứu của bản thân cùng với sự định hướng tận tình, tư duy phản biện sắc sảo và những kiến thức chuyên môn quý giá mà Thầy đã truyền đạt. Em thật sự biết ơn sự kiên nhẫn và những góp ý khắt khe nhưng đầy tâm huyết của Thầy để luận văn đạt được độ hoàn thiện cao nhất.

Tiếp theo, em xin trân trọng gửi lời cảm ơn đến Ban Giám hiệu, cùng toàn thể quý thầy cô Khoa Tài chính - Ngân hàng, Trường Đại học Kinh tế - Luật (ĐHQG-HCM). Xin tri ân các thầy cô đã tận tâm giảng dạy, trang bị cho em một nền tảng kiến thức vững chắc về tài chính định lượng và phương pháp nghiên cứu trong suốt những năm tháng học tập dưới mái trường này. Em vô cùng tự hào và may mắn khi được rèn luyện trong một môi trường học thuật chuyên nghiệp, nơi đã mở ra cho em vô vàn cơ hội phát triển bản thân.

Hơn thế nữa, em xin gửi lời biết ơn vô hạn đến gia đình, đặc biệt là cha mẹ hậu phương vững chắc và là nguồn động viên tinh thần lớn lao nhất của em. Sự hy sinh thầm lặng, tình yêu thương vô điều kiện và sự bao dung của gia đình chính là động lực to lớn giúp em vượt qua những áp lực học thuật, mạnh dạn dấn thân khám phá những hướng nghiên cứu mới và hoàn thành chặng đường này.

Do những giới hạn nhất định về mặt thời gian, nguồn dữ liệu và kinh nghiệm thực tiễn, nghiên cứu này chắc chắn không thể tránh khỏi những thiếu sót. Em rất mong nhận được những ý kiến đóng góp, phê bình quý báu từ Hội đồng bảo vệ và quý thầy cô để đề tài được hoàn thiện hơn, đồng thời mở ra những hướng phát triển sâu rộng hơn trong tương lai.

Cuối cùng, em xin kính chúc quý thầy cô luôn dồi dào sức khỏe, hạnh phúc và gặt hái thêm nhiều thành công trong sự nghiệp giảng dạy.

Trân trọng.

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

MỤC LỤC

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

STT	Viết tắt	Nguyên nghĩa
1	**AP**	Average Precision (Độ chuẩn xác trung bình)
2	**AUC**	Area Under the Curve (Diện tích dưới đường cong ROC)
3	**EMA**	Exponential Moving Average (Đường trung bình động hàm mũ)
4	**MACD**	Moving Average Convergence Divergence (Đường trung bình động hội tụ phân kỳ)
5	**MLP**	Multilayer Perceptron (Mạng thần kinh nhân tạo đa lớp)
6	**NLP**	Natural Language Processing (Xử lý ngôn ngữ tự nhiên)
7	**OHLC**	Open, High, Low, Close (Giá mở cửa, cao nhất, thấp nhất, đóng cửa)
8	**P/B**	Price-to-Book ratio (Hệ số giá trên giá trị sổ sách)
9	**P/E**	Price-to-Earnings ratio (Hệ số giá trên lợi nhuận một cổ phiếu)
10	**ReLU**	Rectified Linear Unit (Hàm kích hoạt tuyến tính chỉnh lưu)
11	**ROC**	Receiver Operating Characteristic (Đường cong đặc trưng hoạt động)
12	**ROE**	Return on Equity (Tỷ suất sinh lời trên vốn chủ sở hữu)
13	**RSI**	Relative Strength Index (Chỉ số sức mạnh tương đối)
14	**SHAP**	SHapley Additive exPlanations (Phương pháp giải thích giá trị Shapley)
15	**SMA**	Simple Moving Average (Đường trung bình động đơn giản)
16	**SMOTE**	Synthetic Minority Over-sampling Technique (Kỹ thuật sinh mẫu thiểu số nhân tạo)

TÓM TẮT

Nghiên cứu này ứng dụng mô hình Mạng thần kinh nhân tạo đa lớp (Multilayer Perceptron - MLP) nhằm phân tích thực nghiệm tác động của thông tin cơ bản, kỹ thuật và tâm lý thị trường đến năng lực dự báo xu hướng giá cổ phiếu ngắn hạn tại Việt Nam. Bằng việc ứng dụng mô hình ngôn ngữ học sâu PhoBERT để định lượng tâm lý truyền thông, kết quả từ phương pháp phân tích cắt bỏ (Ablation Study) chứng minh tổ hợp lai ghép giữa dữ liệu Kỹ thuật và Tâm lý đạt hiệu suất phân lớp tối ưu nhất. Năng lực dự báo này duy trì sự ổn định qua các kiểm định khắt khe bao gồm Bootstrap và kiểm định chéo chuỗi thời gian cuốn chiếu (Rolling TimeSeriesSplit). Phân tích cấu trúc nội tại mô hình bộc lộ sức mạnh chi phối của hiệu ứng đảo chiều về giá trị trung bình, rủi ro biến động và các cú sốc tâm lý; trong khi dữ liệu cơ bản tĩnh lại làm gia tăng độ nhiễu và suy giảm hiệu suất dự báo trong khung thời gian 5 ngày. Đáng chú ý, kết quả thực nghiệm làm nổi bật sự phân hóa sâu sắc về hiệu suất dự báo của thuật toán theo đặc thù từng nhóm ngành, đạt mức vượt trội tại nhóm Viễn thông và Tiện ích cộng đồng. Những phát hiện này cung cấp bằng chứng định lượng sắc bén ủng hộ lý thuyết tài chính hành vi, khẳng định sự cộng hưởng chặt chẽ giữa động lượng dòng tiền và trạng thái tâm lý đám đông trong việc định hình vi cấu trúc thị trường chứng khoán.

Từ khóa: Học máy (Machine Learning), Mạng nơ-ron đa lớp (MLP), Tài chính hành vi, Tâm lý thị trường (Market Sentiment), Dự báo chứng khoán..

CHƯƠNG 1. GIỚI THIỆU

1.1. Lý do chọn đề tài

Thị trường chứng khoán đóng vai trò trọng yếu trong việc huy động và phân bổ nguồn lực tài chính. Tại Việt Nam, với tỷ trọng giao dịch lớn từ nhà đầu tư cá nhân, giá cổ phiếu chịu tác động mạnh bởi yếu tố hành vi và thông tin truyền thông, dẫn đến các biến động giá có tính chất phi tuyến phức tạp. Trong bối cảnh đó, việc áp dụng độc lập phân tích cơ bản hoặc phân tích kỹ thuật bộc lộ nhiều hạn chế, đặc biệt khi thị trường xuất hiện các biến động bất thường do yếu tố tâm lý.

Trong thực tiễn, phân tích cơ bản và phân tích kỹ thuật là hai phương pháp tiếp cận chủ đạo nhưng thường bộc lộ hạn chế khi được áp dụng độc lập, đặc biệt trong các giai đoạn thị trường biến động mạnh bởi yếu tố tâm lý. Để khắc phục vấn đề này, các thuật toán Học máy (Machine Learning) được ứng dụng như một giải pháp thay thế hiệu quả. Cụ thể, Mạng thần kinh nhân tạo đa lớp (Multilayer Perceptron - MLP) thể hiện ưu điểm nổi bật nhờ khả năng mô hình hóa các hàm phi tuyến phức tạp và tích hợp đồng thời nhiều nguồn dữ liệu đa phương thức.

Nghiên cứu của Ballesteros và Martínez Miranda (2024) cho thấy mô hình mạng nơ-ron tích hợp dữ liệu cơ bản, kỹ thuật và tâm lý đem lại hiệu suất dự báo chỉ số S&P 500 cao hơn các phương pháp tiếp cận đơn lẻ. Dựa trên cơ sở lý thuyết này và thực trạng thiếu hụt các nghiên cứu ứng dụng Xử lý ngôn ngữ tự nhiên (NLP) để định lượng tâm lý thị trường tại Việt Nam, đề tài này đề xuất xây dựng mô hình MLP kết hợp đa nguồn dữ liệu. Nghiên cứu tập trung giải quyết khoảng trống học thuật hiện tại, đồng thời cung cấp phương pháp tiếp cận mới hỗ trợ thực tiễn đầu tư.

1.2. Mục tiêu và câu hỏi nghiên cứu

Mục tiêu tổng quát của nghiên cứu là xây dựng và đánh giá hiệu suất của mô hình Mạng thần kinh nhân tạo đa lớp (MLP) trong bài toán dự báo xu hướng giá cổ phiếu ngắn hạn, thông qua việc tích hợp đồng thời ba lăng kính thông tin: phân tích cơ bản, phân tích kỹ thuật và tâm lý thị trường.

Để đạt được mục tiêu tổng quát, nghiên cứu thiết lập ba mục tiêu cụ thể, tương ứng với ba câu hỏi nghiên cứu cốt lõi sau:

(1) Việc tích hợp đồng thời luồng dữ liệu Kỹ thuật, Cơ bản và Tâm lý thị trường có giúp nâng cao năng lực phân lớp và dự báo xu hướng giá cổ phiếu của mạng MLP so với việc sử dụng các nhóm biến đơn lẻ hay không?

(2) Trong khung thời gian dự báo ngắn hạn (5 ngày), nhóm đặc trưng nào (Kỹ thuật, Cơ bản hay Tâm lý) đóng vai trò chi phối và có mức độ đóng góp (Feature Importance) lớn nhất vào quyết định của thuật toán?

(3) Hiệu suất dự báo của mô hình học máy có sự phân hóa như thế nào khi áp dụng trên các nhóm ngành kinh tế mang đặc thù chu kỳ và cấu trúc vốn hóa khác nhau?

1.3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu trọng tâm của đề tài là năng lực dự báo xu hướng biến động giá cổ phiếu (tăng hoặc giảm) trong khung thời gian ngắn hạn của mô hình mạng thần kinh nhân tạo đa lớp dưới tác động của các hệ biến số kỹ thuật, cơ bản và tâm lý.

Về mặt không gian, nghiên cứu được thực hiện trên tập dữ liệu của các cổ phiếu tiêu biểu niêm yết trên thị trường chứng khoán Việt Nam, mang tính đại diện cho các lĩnh vực kinh tế trọng điểm.

Về mặt thời gian, dữ liệu nghiên cứu được thu thập và tổng hợp trong giai đoạn từ năm 2020 đến năm 2025 nhằm bao trọn các chu kỳ biến động lớn của thị trường. Cấu trúc của tập dữ liệu này bao gồm dữ liệu giao dịch lịch sử, báo cáo tài chính định kỳ và kho ngữ liệu tin tức truyền thông tài chính đã qua xử lý.

1.4. Đóng góp của nghiên cứu

Nghiên cứu đóng góp trên hai phương diện học thuật và thực tiễn. Về học thuật, đề tài bổ sung bằng chứng thực nghiệm cho lý thuyết tài chính hành vi tại thị trường cận biên Việt Nam, cho thấy sự tương tác giữa dòng tiền và yếu tố tâm lý truyền thông có thể giải thích biến động giá cổ phiếu hiệu quả hơn các tham số định giá truyền thống. Về thực tiễn, nghiên cứu thiết lập quy trình tiền xử lý và trích xuất đặc trưng dữ liệu, tích hợp điểm số cảm xúc từ mô hình ngôn ngữ lớn (PhoBERT). Khung phân tích định lượng này có khả năng ứng dụng trực tiếp vào các hệ thống giao dịch thuật toán và hỗ trợ quyết định trong quản trị rủi ro danh mục đầu tư.

CHƯƠNG 2. TỔNG QUAN LÝ THUYẾT VÀ CÁC NGHIÊN CỨU TRƯỚC

2.1. Tổng quan lý thuyết

Nền tảng của định giá tài sản truyền thống được xây dựng dựa trên Lý thuyết thị trường hiệu quả (Efficient Market Hypothesis - EMH) do Fama (1970) đề xuất. EMH lập luận rằng giá cổ phiếu tại mọi thời điểm luôn phản ánh đầy đủ toàn bộ thông tin sẵn có trên thị trường, do đó, các nỗ lực dự báo lợi suất vượt trội thông qua việc phân tích dữ liệu lịch sử là bất khả thi nếu không nắm giữ thông tin nội bộ. Tuy nhiên, sự phát triển của các thị trường tài chính, đặc biệt là các thị trường mới nổi, đã bộc lộ nhiều điểm bất thường mà EMH không thể giải thích. Khắc phục hạn chế này, lý thuyết Tài chính hành vi (Behavioral Finance) ra đời, nhấn mạnh rằng các quyết định đầu tư không hoàn toàn mang tính lý trí mà chịu ảnh hưởng sâu sắc bởi cảm xúc và hiệu ứng đám đông. Barberis, Shleifer và Vishny (1998) cùng Shiller (2003) đã chỉ ra rằng nhà đầu tư thường có xu hướng phản ứng thái quá (overreaction) hoặc phản ứng chậm (underreaction) trước các thông tin mới, từ đó tạo ra những sai lệch giá kéo dài so với giá trị nội tại và làm suy giảm hiệu quả dự báo của các mô hình kinh tế lượng tuyến tính truyền thống.

Để định lượng và dự báo các biến động này, giới học thuật và thực hành tài chính thường phân tách thành hai trường phái chủ đạo: phân tích cơ bản và phân tích kỹ thuật. Phân tích cơ bản tập trung đánh giá sức khỏe tài chính và năng lực cốt lõi của doanh nghiệp thông qua các hệ số định giá như P/E, P/B, ROE và EPS. Theo Damodaran (2012), nền tảng tài chính vững mạnh là mỏ neo quyết định dòng tiền kỳ vọng và giá trị nội tại dài hạn của tài sản. Ngược lại, phân tích kỹ thuật hoạt động dựa trên giả định rằng mọi yếu tố cơ bản và tâm lý đều đã được phản ánh vào diễn biến giá và khối lượng giao dịch. Theo Murphy (1999), phương pháp này khai thác các chỉ báo xu hướng (SMA, EMA), động lượng (RSI, MACD) và rủi ro biến động (Volatility) nhằm mô hình hóa chu kỳ lặp lại của tâm lý con người trên biểu đồ giá.

Nhằm xử lý khối lượng dữ liệu khổng lồ và giải mã các tương tác phi tuyến tính giữa ba lăng kính Cơ bản, Kỹ thuật và Tâm lý thị trường, Mạng thần kinh nhân tạo (Artificial Neural Network - ANN) nổi lên như một công cụ tối ưu. Cụ thể, Mạng thần kinh nhân tạo đa lớp (Multilayer Perceptron - MLP) là kiến trúc mạng truyền thẳng bao gồm lớp đầu vào, các lớp ẩn và lớp đầu ra. Theo Haykin (2009), MLP vận hành thông qua thuật toán lan truyền ngược (backpropagation) và các hàm kích hoạt phi tuyến, cho phép

mô hình tự động trích xuất các đặc trưng trừu tượng và xấp xỉ hóa các hàm quan hệ phức tạp. Đặc tính toán học này biến MLP trở thành thuật toán lý tưởng để thiết lập không gian đặc trưng đa chiều trong các bài toán dự báo rủi ro tài chính tần suất cao.

2.2. Lược khảo các nghiên cứu thực nghiệm trước đây

Ứng dụng của học máy trong dự báo chuỗi thời gian tài chính đã trải qua nhiều giai đoạn phát triển, bắt đầu từ việc chứng minh sự vượt trội của mạng nơ-ron so với các mô hình thống kê truyền thống. Công trình của Huang và các cộng sự (2005) và Wanjawa, Muchemi (2014) đã ứng dụng mạng MLP với thuật toán lan truyền ngược để dự báo chỉ số thị trường, cung cấp bằng chứng thực nghiệm rõ ràng về năng lực nắm bắt các nhiễu loạn phi tuyến của thuật toán mạng nơ-ron nhân tạo (ANN) so với phương pháp hồi quy tuyến tính cổ điển. Đánh giá toàn diện về xu hướng này, Atsalakis và Valavanis (2009) khẳng định các mô hình tính toán mềm đặc biệt phù hợp với cấu trúc dữ liệu tài chính có độ nhiễu cao.

Tiến thêm một bước trong việc giải mã vi cấu trúc thị trường, luồng nghiên cứu thứ hai tập trung vào việc số hóa và định lượng tâm lý nhà đầu tư thông qua Xử lý ngôn ngữ tự nhiên (NLP). Tetlock (2007) tiên phong lượng hóa nội dung báo chí tài chính, chứng minh sự bi quan trên truyền thông có hệ số tương quan đồng biến với sự sụt giảm lợi suất ngắn hạn. Tiếp nối xu hướng này, Halder (2022) đã đề xuất kiến trúc học sâu kết hợp mô hình ngôn ngữ tài chính FinBERT nhằm phân tích cảm xúc từ các bài báo của New York Times, qua đó gia tăng đáng kể độ chính xác trong việc dự báo chỉ số NASDAQ-100. Sự chuyển dịch sang các mô hình tích hợp toàn diện đạt đến đỉnh cao với công trình của Ballesteros và Martínez Miranda (2024). Các tác giả đã thiết kế mạng nơ-ron đa lớp (MLP) tích hợp đồng thời ba nguồn thông tin: cơ bản, kỹ thuật và tâm lý để dự báo xu hướng chỉ số S&P 500, chứng minh một cách thống kê rằng không gian đặc trưng lai ghép mang lại hiệu suất phân lớp vượt trội hoàn toàn so với các nhóm biến độc lập.

Đáng chú ý, tại thị trường chứng khoán Việt Nam, bộ môn học máy ứng dụng cũng đang thu hút sự quan tâm mạnh mẽ. Điển hình, Thao và các cộng sự (2024) đã công bố trên tạp chí PLOS One một nghiên cứu sử dụng học máy kết hợp kỹ thuật trích xuất đặc trưng tự động từ biến số cơ bản và kỹ thuật để dự báo tỷ suất sinh lời thị trường Việt Nam, khẳng định vai trò của máy học trong định giá tài sản. Về phương diện khai thác dữ liệu phi cấu trúc, Nguyễn và các cộng sự (2022) trên Tạp chí Khoa học Đại học Quốc gia Hà

Nội đã ứng dụng kỹ thuật khai phá văn bản (Text-mining) trên 70.000 bài báo tài chính trong nước, sử dụng Rừng ngẫu nhiên (Random Forest) và Máy học véc-tơ hỗ trợ (SVM) để dự báo thành công biến động của chỉ số VN-Index. Mới đây nhất, một nghiên cứu trên hệ thống tạp chí MDPI (2023) đã ứng dụng mô hình ngôn ngữ tiếng Việt PhoBERT để phân loại cảm xúc của gần 40.000 bài báo kinh tế, cung cấp bằng chứng định lượng sắc bén về việc tin tức tiêu cực có tác động trực tiếp làm thay đổi phương sai và cấu trúc biến động của thị trường chứng khoán trong nước.

2.3. Khoảng trống nghiên cứu

Tổng hợp lược khảo y văn cho thấy sự dịch chuyển tất yếu của các phương pháp định lượng từ mô hình đơn phân kỳ sang các kiến trúc học máy đa phương thức. Mặc dù đã có những nghiên cứu bước đầu tại Việt Nam áp dụng NLP (như PhoBERT hay Text-mining) để dự báo chỉ số thị trường chung (VN-Index), hoặc dùng học máy để phân tích các biến tài chính truyền thống nhưng sự giao thoa toàn diện giữa các trường phái này vẫn còn vắng bóng.

Phần lớn các công trình tích hợp xuất sắc cả ba lăng kính (Cơ bản, Kỹ thuật và Tâm lý) bằng mạng nơ-ron như Ballesteros và Martínez Miranda (2024) hay Halder (2022) đều chỉ được kiểm định trên dữ liệu của Mỹ (S&P 500, NASDAQ), nơi thông tin minh bạch và nhà đầu tư tổ chức chiếm ưu thế. Đặc thù của thị trường Việt Nam là tính bất cân xứng thông tin cao và sự chi phối áp đảo của dòng tiền cá nhân, khiến các sai lệch giá do hiệu ứng hành vi diễn ra với cường độ mạnh mẽ hơn. Việc thiếu vắng một nghiên cứu ứng dụng mô hình MLP tích hợp đồng thời dữ liệu giao dịch nội tại của từng cổ phiếu (stock-level) và điểm số cảm xúc truyền thông (được trích xuất từ PhoBERT) chính là một khoảng trống học thuật lớn. Nghiên cứu này được thực hiện nhằm lấp đầy khoảng trống đó, cung cấp cơ sở kiểm định định lượng về mức độ tương tác giữa nền tảng cơ bản, động lượng dòng tiền và cú sốc tâm lý trong hệ sinh thái đầu tư tại Việt Nam.

CHƯƠNG 3. PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Giới thiệu chung và thiết kế nghiên cứu

Chương này thiết lập một khung khổ phương pháp luận hệ thống nhằm giải quyết bài toán dự báo xu hướng giá cổ phiếu tại thị trường chứng khoán Việt Nam thông qua mô hình mạng thần kinh nhân tạo đa lớp (Multilayer Perceptron - MLP). Trên nền tảng triết lý thực chứng (Positivism), nghiên cứu vận dụng phương pháp định lượng thực nghiệm để kiểm chứng giả định rằng các biến động thị trường, dù tồn tại yếu tố ngẫu nhiên vẫn chứa đựng những quy luật phi tuyến và mẫu hình có thể nhận diện được qua việc phân tích dữ liệu lịch sử.

Thiết kế nghiên cứu được triển khai dựa trên quy trình suy diễn, bắt đầu từ việc kế thừa các lý thuyết nền tảng như Giả thuyết thị trường hiệu quả (EMH) của Fama (1970) và các học thuyết tài chính hành vi để xây dựng các giả thuyết khoa học về mối liên hệ giữa thông tin quá khứ và xu hướng giá tương lai. Để thực hiện kiểm định, một quy trình phân tích dữ liệu khép kín được thiết lập chặt chẽ, bắt đầu từ giai đoạn thu thập dữ liệu đa nguồn từ các tổ chức cung cấp uy tín, tiếp nối bằng công đoạn làm sạch và tiền xử lý dữ liệu thô nhằm đảm bảo tính chuẩn xác cho mô hình.

Đặc biệt, nghiên cứu chú trọng vào kỹ thuật đặc trưng để chuyển đổi các biến số thô thành tập hợp đầu vào có giá trị dự báo cao kết hợp giữa dữ liệu giao dịch tần suất cao, các chỉ số từ báo cáo tài chính định kỳ và dữ liệu phi cấu trúc từ tin tức kinh tế. Toàn bộ quá trình này tạo tiền đề cho việc thiết kế, huấn luyện và tối ưu hóa cấu trúc mạng MLP, từ đó cho phép đánh giá hiệu suất dự báo thông qua các kiểm định thống kê nghiêm ngặt, đảm bảo tính khách quan và khả năng tổng quát hóa của kết quả nghiên cứu trong bối cảnh thị trường chứng khoán đang phát triển.

3.2. Dữ liệu nghiên cứu

3.2.1. Chọn mẫu và khung thời gian nghiên cứu

Đối tượng khảo sát của nghiên cứu được xác định dựa trên danh sách 100 cổ phiếu có mức vốn hóa thị trường lớn nhất niêm yết tại Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh (HOSE). Nhằm đảm bảo tính liên tục của chuỗi thời gian và loại trừ rủi ro thiếu hụt dữ liệu (missing data), tập mẫu ban đầu tiếp tục được tiến hành sàng lọc, qua đó chỉ giữ lại những mã cổ phiếu đã được niêm yết chính thức và giao dịch xuyên suốt từ

tháng 01/2020. Việc tập trung vào nhóm cổ phiếu vốn hóa lớn không chỉ đảm bảo tính đại diện bao quát cho toàn thị trường mà còn đáp ứng tiêu chuẩn khắt khe về tính thanh khoản. Cụ thể, đặc tính thanh khoản cao giúp giảm thiểu tối đa các sai lệch phát sinh do chênh lệch giá mua - bán (bid-ask spread), từ đó củng cố tính khả thi thực tiễn cho các chiến lược giao dịch mô phỏng dựa trên tín hiệu dự báo của mô hình.

Khung thời gian nghiên cứu được xác định từ ngày 01/01/2020 đến ngày 30/09/2025. Đây là giai đoạn có biến động cấu trúc mạnh mẽ (structural breaks) với đầy đủ các chu kỳ thị trường: từ cú sốc sụt giảm do đại dịch Covid-19 (2020), giai đoạn tăng trưởng bùng nổ (2021), đến các đợt điều chỉnh sâu do thắt chặt tiền tệ và thanh lọc thị trường (2022) và cuối cùng là giai đoạn phục hồi tích lũy (2023 - quý 3/2025). Việc bao hàm các điều kiện thị trường đối lập này cho phép kiểm chứng chuẩn xác độ bền vững (robustness) và khả năng thích nghi của mô hình MLP trước các cú sốc ngoại sinh.

3.2.2. Nguồn dữ liệu và quy trình thu thập

Dữ liệu đầu vào phục vụ cho nghiên cứu được thu thập từ hai nguồn dữ liệu cấp một uy tín và phổ biến hàng đầu tại thị trường Việt Nam hiện nay là WiGroup và CafeF. Việc lựa chọn các nguồn cung cấp này nhằm đảm bảo tính toàn vẹn, độ chính xác và khả năng kiểm chứng độc lập của bộ dữ liệu.

Đầu tiên, đối với nhóm dữ liệu giao dịch và tài chính doanh nghiệp, nghiên cứu khai thác cơ sở dữ liệu từ nền tảng WiGroup. Dữ liệu giao dịch bao gồm chuỗi thời gian về giá mở cửa, giá cao nhất, giá thấp nhất, giá đóng cửa (OHLC) và khối lượng khớp lệnh được thu thập theo tần suất ngày (daily frequency). Một điểm quan trọng trong quá trình thu thập dữ liệu giá từ WiGroup là việc sử dụng chuỗi giá điều chỉnh (adjusted price). Giá cổ phiếu lịch sử được điều chỉnh kỹ thuật để loại bỏ các tác động nhân tạo do các sự kiện doanh nghiệp như chia tách cổ phiếu, trả cổ tức bằng tiền mặt hoặc cổ phiếu thưởng và phát hành quyền mua. Điều này đảm bảo rằng các biến động giá trong chuỗi dữ liệu chỉ phản ánh cung cầu thị trường thực tế, tránh gây nhiễu cho quá trình huấn luyện mô hình. Đồng thời, dữ liệu tài chính hàng quý của các doanh nghiệp trong tập mẫu được trích xuất từ Bảng cân đối kế toán, Báo cáo kết quả kinh doanh và Báo cáo lưu chuyển tiền tệ, toàn bộ được chuẩn hóa theo Chuẩn mực Kế toán Việt Nam (VAS).

Bảng 1. Phân loại và mô tả các biến dữ liệu kỹ thuật

Tên biến	Tên đầy đủ	Mô tả
close	Giá đóng cửa	Mức giá đóng cửa của cổ phiếu trong phiên giao dịch (đã biến đổi logarit).
volume	Khối lượng	Tổng khối lượng giao dịch trong phiên (đã biến đổi logarit).
rsi_14	Sức mạnh tương đối	Chỉ số RSI 14 ngày, đo lường trạng thái quá mua hoặc quá bán của cổ phiếu.
sma20_gap	Khoảng cách SMA	Mức độ chênh lệch giữa giá hiện tại và đường trung bình động 20 ngày..
sma_cross	Tín hiệu giao cắt	Tín hiệu giao cắt của các đường trung bình động (thường đại diện cho xu hướng ngắn hạn cắt dài hạn).
volatility_20d	Biến động giá	Độ lệch chuẩn của tỷ suất sinh lời trong 20 ngày qua, đại diện cho rủi ro hệ thống ngắn hạn.
ret_3d	Động lượng giá	Tỷ suất sinh lời trong 3 ngày giao dịch gần nhất (Momentum).
ret_5d_rank	Xếp hạng lợi suất	Vị thế xếp hạng tỷ suất sinh lời 5 ngày của cổ phiếu so với các mã khác trong rổ quan sát.
volume_ratio	Tỷ lệ khối lượng	Tỷ lệ khối lượng giao dịch hiện tại so với mức trung bình nền tảng, chuẩn hóa biến đổi bằng logarit.
momentum_vol	Động lượng khối lượng	Mức độ thay đổi động lượng của khối lượng giao dịch.

Nguồn: Tác giả tự tổng hợp

Bảng 2. Phân loại và mô tả các biến dữ liệu tài chính cơ bản

Tên biến	Tên đầy đủ	Mô tả
eps	Thu nhập cổ phần	Lợi nhuận sau thuế trên mỗi cổ phiếu.
pb	Định giá P/B	Tỷ lệ giữa giá thị trường và giá trị sổ sách của cổ phiếu.
pe	Định giá P/E	Hệ số giữa giá thị trường và lợi nhuận trên mỗi cổ phiếu.
lnst_yoy	Tăng trưởng LNST	Tốc độ tăng trưởng lợi nhuận sau thuế so với cùng kỳ năm trước.
nophaitra_vcsh	Debt-to-Equity ratio	Tỷ lệ nợ phải trả trên vốn chủ sở hữu, thể hiện đòn bẩy tài chính.
vonhoa_tts	Cấu trúc vốn hóa	Tỷ lệ giữa vốn hóa thị trường trên tổng tài sản của doanh nghiệp.
profitability	Khả năng sinh lời	Biến tổng hợp đo lường hiệu quả hoạt động, tính bằng trung bình cộng Z-score của hệ số ROA và ROE.

Nguồn: Tác giả tự tổng hợp

Thứ hai, đối với dữ liệu tin tức phục vụ phân tích tâm lý thị trường, nghiên cứu thực hiện quy trình thu thập và xử lý dữ liệu tự động từ chuyên trang tài chính CafeF. CafeF được lựa chọn vì đây là một trong những cổng thông tin tài chính lâu đời và có lượng truy cập lớn nhất tại Việt Nam, phản ánh kịp thời các thông tin tác động đến tâm lý nhà đầu tư cá nhân. Quy trình thu thập được thực hiện thông qua các thư viện lập trình Python như Requests và BeautifulSoup. Hệ thống được lập trình để truy cập vào kho lưu trữ tin tức của từng mã cổ phiếu, trích xuất và tiến hành xử lý tiêu đề bài viết, cùng với đoạn mô tả ngắn (sapo), nội dung chi tiết và thời gian đăng tải. Để đảm bảo tính phù hợp, chỉ các bài viết có gắn thẻ (tag) liên quan trực tiếp đến mã cổ phiếu hoặc tên doanh nghiệp thuộc mẫu nghiên cứu mới được giữ lại. Dữ liệu văn bản sau khi thu thập và làm sạch sẽ được lưu trữ dưới dạng cấu trúc bảng để phục vụ cho các bước xử lý ngôn ngữ tự nhiên tiếp theo.

Bảng 3. Mô tả tập dữ liệu tin tức từ CafeF

Chỉ tiêu thống kê	**Kết quả**
Tổng số lượng bài viết thu thập	35.481 bài
Số lượng doanh nghiệp (mã cổ phiếu)	100 mã
Số bài viết trung bình / 1 doanh nghiệp	~ 355 bài
Mã cổ phiếu có lượng tin tức nhiều nhất (Max)	HPG (1.125 bài)
Mã cổ phiếu có lượng tin tức ít nhất (Min)	DSE (1 bài)

Nguồn: Tác giả tự tổng hợp

Bảng 4. Các biến tâm lý thị trường và phân loại

Tên biến	**Tên đầy đủ**	**Mô tả**
news_dummy	Tín hiệu tin tức	Biến giả, nhận giá trị 1 nếu có tin tức về doanh nghiệp xuất hiện trong ngày và 0 nếu ngược lại.
sent_shock	Cú sốc tâm lý	Được tính bằng phần dư (residual) giữa điểm cảm xúc hiện tại ở ngày t và điểm tâm lý nền của chính giai đoạn đó: $S_t - sent\_score\_roll_{t-1}$, để nắm bắt các phản ứng thái quá bất ngờ của giới truyền thông
sent_score_roll	Điểm tâm lý nền	Điểm số cảm xúc Trung bình cộng trượt của điểm cảm xúc (S) trong 5 ngày: $\frac{1}{5}\sum_{i=0}^{4} S_{t-i}$, phản ánh tâm lý chung đang chi phối cổ phiếu.
momentum_sent	Động lượng theo tâm lý	Tích số giữa lợi suất 3 ngày (ret_3d) và điểm tâm lý nền (sent_score_roll): $ret\_3d_t \times sent\_score\_roll_t$, phản ánh tác động kép của đà tăng giá khi có tâm lý hỗ trợ.
volume_sent	Khối lượng theo tâm lý	Tích số giữa tỷ lệ khối lượng đã chuẩn hóa logarit (volume_ratio) và điểm tâm lý nền: $volume\_ratio_t \times sent\_score\_roll_t$, đo lường sự đồng thuận của dòng tiền dưới tác động của tin tức.

Nguồn: Tác giả tự tổng hợp

3.3. Xây dựng biến số và không gian đặc trưng

3.3.1. Tiền xử lý dữ liệu và đồng bộ hóa thời gian

Dữ liệu thô sau khi thu thập từ đa nguồn thường tồn tại các sai sót kỹ thuật, giá trị khuyết thiếu và sự không đồng nhất về tần suất thời gian. Do đó, quy trình tiền xử lý được thiết lập nhằm chuẩn hóa dữ liệu đầu vào, đảm bảo tính toàn vẹn cho mô hình học máy. Đối với dữ liệu giao dịch từ WiGroup, nghiên cứu thực hiện kiểm tra tính liên tục của chuỗi thời gian, loại bỏ các phiên giao dịch không phát sinh thanh khoản hoặc bị đình chỉ.

Đối với dữ liệu tài chính, thách thức lớn nhất nằm ở sự chênh lệch tần suất (dữ liệu quý so với dữ liệu ngày) và độ trễ công bố thông tin. Để triệt tiêu hiện tượng nhìn trước tương lai (look-ahead bias), một sai số nghiêm trọng trong kiểm định tài chính nghiên cứu không gán dữ liệu tài chính vào ngày kết thúc quý kế toán. Dữ liệu được ánh xạ vào ngày công bố thông tin thực tế trên thị trường. Kỹ thuật lấy mẫu lại và điền đầy dữ liệu về phía trước (Forward-filling) được áp dụng để chuyển đổi dữ liệu quý sang tần suất ngày, đảm bảo rằng tại mỗi thời điểm t, mô hình chỉ tiếp cận được các chỉ số tài chính đã được công bố chính thức trước đó.

Đối với dữ liệu văn bản từ CafeF, quy trình làm sạch tập trung vào việc loại bỏ các nhiễu thông tin như thẻ HTML, ký tự đặc biệt, URL và các nội dung trùng lặp. Văn bản tiếng Việt được chuẩn hóa theo bảng mã Unicode dựng sẵn, chuyển về dạng chữ thường và loại bỏ các từ dừng (stop words) không mang ý nghĩa phân loại. Bước này đóng vai trò quyết định trong việc tinh lọc các đặc trưng ngôn ngữ cho giai đoạn xử lý ngôn ngữ tự nhiên phía sau.

3.3.2. Xây dựng biến phụ thuộc

Khác với cách tiếp cận hồi quy truyền thống hướng tới dự báo mức giá trị tuyệt đối, nghiên cứu này định dạng bài toán dưới dạng mô hình phân lớp nhị phân (Binary Classification). Mục tiêu trọng tâm là nhận diện các cổ phiếu có khả năng xác lập mức sinh lời vượt trội so với thị trường. Biến phụ thuộc (Target variable) được xác định dựa trên tỷ suất sinh lời lũy kế trong khung thời gian 5 ngày giao dịch tiếp theo (t+5).

Để gán nhãn dữ liệu một cách khách quan, phương pháp phân vị cực đoan được áp dụng. Cụ thể, tại mỗi thời điểm quan sát t, phân phối tỷ suất sinh lời tương lai của toàn bộ rổ cổ phiếu trong mẫu nghiên cứu được tính toán. Nghiên cứu xác định hai ngưỡng phân

vị: ngưỡng trên (phân vị 80%_q_{80}) và ngưỡng dưới (phân vị 20%_q_{20}). Biến mục tiêu y sẽ nhận giá trị 1 (đại diện cho lớp tăng trưởng) nếu tỷ suất sinh lời lớn hơn hoặc bằng q_{80}, và nhận giá trị 0 (đại diện cho lớp suy giảm) nếu tỷ suất sinh lời nhỏ hơn hoặc bằng q_{20}.

Các quan sát có mức lợi suất dao động trong vùng trung vị (từ q_{20} đến q_{80}) được xem là vùng nhiễu hoặc trạng thái giá đi ngang (sideway) và được chủ động loại bỏ khỏi tập dữ liệu huấn luyện. Kỹ thuật này giúp tinh lọc không gian mẫu, ép buộc các thuật toán học máy phải tập trung tối đa nguồn lực để nhận diện các đặc trưng của những biến động giá xu hướng rõ nét nhất, qua đó cải thiện độ chính xác và khả năng ứng dụng thực tiễn của mô hình.

3.3.3. Xây dựng biến độc lập

Để giải quyết bài toán dự báo xu hướng giá cổ phiếu, một đại lượng chịu tác động của vô số các yếu tố ngẫu nhiên và phi tuyến tính, nghiên cứu áp dụng cách tiếp cận đa chiều (multi-modal approach) trong việc xây dựng biến số. Tập hợp các biến đầu vào được thiết kế nhằm bao quát ba khía cạnh cốt lõi của thị trường: động lượng giá quá khứ, giá trị nội tại của doanh nghiệp và tâm lý nhà đầu tư. Sự kết hợp này dựa trên giả thuyết rằng thị trường không hoàn toàn hiệu quả, và thông tin từ các khía cạnh khác nhau sẽ bù đắp khiếm khuyết cho nhau để tạo ra tín hiệu dự báo chính xác nhất.

Nhóm biến thứ nhất: Các chỉ báo Phân tích kỹ thuật

Dựa trên chuỗi dữ liệu giá (OHLC) và khối lượng giao dịch, nhóm biến phân tích kỹ thuật được thiết kế nhằm nắm bắt các mẫu hình biến động vi mô theo ba khía cạnh cốt lõi: xu hướng, động lượng và rủi ro biến động.

Đầu tiên, chỉ số Sức mạnh Tương đối (RSI 14 ngày) được sử dụng để định lượng trạng thái quá mua và quá bán, hỗ trợ dự báo các điểm đảo chiều tiềm năng. Về nhận diện xu hướng, nghiên cứu kết hợp hệ thống đường trung bình động (SMA, EMA đa chu kỳ) và chỉ báo MACD nhằm đo lường gia tốc dòng tiền. Để tối ưu hóa khả năng phân lớp phi tuyến tính của thuật toán, các biến phái sinh như độ lệch giá so với xu hướng nền (sma20_gap) và cường độ giao cắt động lượng (sma_cross) cũng được bổ sung.

Nhằm lượng hóa rủi ro, Dải Bollinger (với biên độ dao động $\pm 2\sigma$) và độ lệch chuẩn tỷ suất sinh lời 20 ngày (volatility_20d) được áp dụng để nhận diện các vùng nén giá và cảnh báo bùng nổ biến động. Đồng thời, dòng tiền được kiểm soát chặt chẽ thông qua tỷ

lệ đột biến khối lượng (volume_ratio) và hệ thống các biến trễ lợi suất đóng vai trò như các nhân tố tự hồi quy (autoregressive features). Cuối cùng, để tăng cường độ sắc bén của mô hình, nghiên cứu tích hợp các đặc trưng tương tác đa chiều: xếp hạng lợi suất chéo (ret_5d_rank) nhằm đánh giá sức mạnh tương đối của cổ phiếu và biến tương tác động lượng - biến động (momentum_vol) giúp thuật toán phân định rạch ròi giữa một đà tăng trưởng bền vững và những nhịp tăng giá đầu cơ đi kèm rủi ro hoảng loạn.

Nhóm biến thứ hai: Các chỉ báo Phân tích cơ bản

Trong khi nhóm phân tích kỹ thuật theo dõi sát sao diễn biến dòng tiền ngắn hạn, nhóm biến phân tích cơ bản đóng vai trò thiết lập mỏ neo giá trị, giúp mô hình đối chiếu mức giá thị trường hiện tại với năng lực cốt lõi của doanh nghiệp. Tập hợp các biến này được lựa chọn để phản ánh toàn diện sức khỏe tài chính thông qua bốn khía cạnh: định giá, khả năng sinh lời, đòn bẩy tài chính và tiềm năng tăng trưởng. Cụ thể, hệ số Giá trên Lợi nhuận (pe) và hệ số Giá trên Giá trị sổ sách (pb) được sử dụng làm thước đo định giá tương đối của thị trường; Tỷ suất sinh lời trên vốn chủ sở hữu (roe), Tỷ suất sinh lời trên tổng tài sản (roa) và Thu nhập trên mỗi cổ phần (eps) đại diện cho hiệu quả hoạt động; Tỷ lệ nợ phải trả trên vốn chủ sở hữu (nophaitra_vcsh) và cấu trúc vốn hóa trên tổng tài sản (vonhoa_tts) giúp định lượng rủi ro đòn bẩy; trong khi tốc độ tăng trưởng lợi nhuận sau thuế (lnst_yoy) đóng vai trò là động lực thúc đẩy giá trị kỳ vọng

Nhóm biến thứ ba: Chỉ số Tâm lý thị trường

Nhận thức được hạn chế của các phương pháp phân tích văn bản truyền thống dựa trên từ điển (Dictionary-based) vốn không thể nắm bắt được ngữ cảnh phức tạp của tiếng Việt, nghiên cứu này áp dụng kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) hiện đại dựa trên mô hình học sâu PhoBERT (Nguyen & Nguyen, 2020). PhoBERT là một mô hình ngôn ngữ dựa trên kiến trúc Transformer, được huấn luyện trước trên tập dữ liệu tiếng Việt quy mô lớn, cho phép hiểu sâu sắc ngữ nghĩa và sắc thái của văn bản. Quy trình lượng hóa tâm lý được thực hiện tự động: Các tiêu đề tin tức thu thập từ CafeF được đưa qua mô hình PhoBERT để phân loại thành ba trạng thái cảm xúc: tích cực, trung lập và tiêu cực. Đầu ra của mô hình là phân phối xác suất cho từng trạng thái. Thay vì chỉ đếm số lượng từ khóa đơn giản, điểm số cảm xúc tại mỗi thời điểm t được tính toán dựa trên hiệu số xác suất tin cậy của mô hình:

$$S_t = \frac{1}{N_t}\sum_{i=1}^{N_t}(P_{pos,i} - P_{neg,i})$$

Trong đó N_t là tổng số tin tức trong ngày $t, P_{pos,i}, P_{neg,i}$ lần lượt là xác suất dự báo của mô hình PhoBERT cho nhãn tích cực và tiêu cực của bản tin thứ i. Chỉ số này nhận giá trị liên tục trong khoảng $[-1, 1]$, cung cấp một tín hiệu định lượng tinh vi về mức độ lạc quan hay bi quan của đám đông nhà đầu tư, yếu tố thường dẫn dắt các biến động giá phi lý trí trong ngắn hạn.

Tiếp đó, nghiên cứu tiến hành tổng hợp dữ liệu văn bản theo tần suất ngày cho từng mã cổ phiếu để tạo ra một không gian đặc trưng tâm lý đa chiều. Hệ thống tính toán các đại lượng thống kê mô tả bao gồm: giá trị trung bình, độ lệch chuẩn, mức cực đại và cực tiểu của điểm số tâm lý tổng hợp (sent_score), cũng như phân phối xác suất dự báo cho các nhãn cảm xúc tích cực (sent_pos), tiêu cực (sent_neg) và trung lập (sent_neu). Đồng thời, cường độ thông tin cũng được kiểm soát thông qua biến đếm số lượng bản tin phát hành trong ngày (news_count).

Nhận thức rõ tính chất phi tuyến của thị trường, nghiên cứu thiết kế các biến tương tác để khai thác sự cộng hưởng giữa các nhân tố định lượng. Điển hình là biến động lượng theo tâm lý (momentum_sent), tích số giữa lợi suất 3 ngày và điểm tâm lý nền và khối lượng theo tâm lý (volume_sent). Các chỉ báo lai này đóng vai trò như bộ lọc tín hiệu một đà tăng giá chỉ được coi là có giá trị dự báo cao khi đi kèm với sự đồng thuận tích cực từ truyền thông, phản ánh sự nhập cuộc của dòng tiền thông minh (Smart money). Việc bổ sung các siêu phẳng (Hyperplanes) tương tác này giúp thuật toán MLP nhận diện chính xác hơn các điểm đảo chiều và xu hướng tăng trưởng dài hạn.

3.4. Làm sạch và chuẩn hóa dữ liệu

Dựa trên kết quả thống kê mô tả sơ bộ, tập dữ liệu được tiến hành xử lý qua các bước chuẩn hóa chuyên sâu nhằm giảm thiểu nhiễu và tối ưu hóa hiệu suất của các mô hình dự báo. Quá trình này được thực hiện theo một trình tự logic, bao gồm các bước sau:

Thứ nhất, nghiên cứu tiến hành làm sạch dữ liệu và xử lý các giá trị ngoại lai. Cụ thể, các quan sát vô nghĩa về mặt tài chính bị loại bỏ khỏi mẫu nghiên cứu, bao gồm những cổ phiếu ghi nhận hệ số định giá $P/E \leq 0$ hoặc $P/B \leq 0$, do các mức định giá này không phản ánh chuẩn xác tình trạng hoạt động liên tục của doanh nghiệp. Tiếp đó, để hạn chế

tác động tiêu cực của các giá trị cực đoan (outliers) lên quá trình huấn luyện, kỹ thuật Winsorization ở ngưỡng 1% và 99% được áp dụng. Phương pháp này cắt gọt hai vệt đuôi phân phối của các biến cơ bản dễ bị nhiễu như hệ số $P/E, P/B, EPS, ROA, ROE$, tỷ lệ nợ trên vốn chủ sở hữu, tỷ lệ khối lượng giao dịch và vốn hóa thị trường, giúp ngăn chặn hiện tượng các giá trị ngoại lai làm sai lệch trọng số của mô hình dự báo.

Thứ hai, nghiên cứu thực hiện các phép biến đổi logarit nhằm đưa phân phối dữ liệu về dạng tiệm cận chuẩn, qua đó hạn chế sự thiên lệch (bias) của thuật toán tối ưu hóa (gradient optimization). Phép biến đổi $\ln(x)$ (với giới hạn biên dưới rất nhỏ để tránh lỗi toán học) được áp dụng cho nhóm mức giá và đường trung bình động (Close, Open, High, Low, SMA). Đối với các tỷ số tài chính và tỷ lệ khối lượng có giá trị dương, phép biến đổi $\ln(1 + x)$ được sử dụng. Riêng biến tăng trưởng lợi nhuận (lnst_yoy) do có chứa cả giá trị âm và dương, nghiên cứu áp dụng phép biến đổi logarit có dấu nhằm thu hẹp biên độ biến động nhưng vẫn bảo toàn trọn vẹn chiều hướng tăng giảm gốc của dữ liệu:

$$x' = sign(x) \times \ln(1 + |x|)$$

Thứ ba, thay vì sử dụng các chỉ số sinh lời một cách đơn lẻ và thô, nghiên cứu tiến hành chuẩn hóa nội bộ (within-stock Z-score) đối với hệ số ROA và ROE của từng mã cổ phiếu riêng biệt. Phương pháp này giúp kiểm soát đặc tính cố hữu và quy mô của từng doanh nghiệp cụ thể. Sau khi chuẩn hóa, một biến tổng hợp đại diện cho khả năng sinh lời (Profitability) được tạo ra nhằm thay thế cho ROA và ROE đơn lẻ, mang lại một thước đo hiệu quả và ổn định hơn cho mô hình dựa trên công thức trọng số cân bằng:

$$Profitability_{i,t} = 0.5 \times ROA\_z_{i,t} + 0.5 \times ROE\_z_{i,t}$$

Thứ tư, đối với nhóm biến tâm lý thị trường (Sentiment), nghiên cứu áp dụng các kỹ thuật định lượng để chuyển hóa dữ liệu phi cấu trúc thành các tín hiệu giao dịch. Số lượng tin tức (news_count) được biến đổi logarit và tích hợp cùng một biến giả (news_dummy) nhằm ghi nhận trạng thái có hoặc không có tin tức xuất hiện trong phiên. Điểm số tâm lý hàng ngày (sent_score) được làm mượt bằng đường trung bình trượt 5 ngày (sent_score_roll). Từ nền tảng này, biến cú sốc tâm lý (sent_shock) được đo lường thông qua mức chênh lệch trực tiếp giữa điểm số hiện hành và trung bình trượt, giúp mô hình nắm bắt được những điểm uốn đột ngột trong cảm xúc của nhà đầu tư.

Thứ năm, nhằm tăng cường khả năng phân loại và dự báo của mô hình (đặc biệt là đối với thang đo độ chính xác AUC), nghiên cứu thiết kế thêm các biến tương tác bằng cách kết hợp chéo giữa yếu tố kỹ thuật và tâm lý. Hai biến tương tác cốt lõi bao gồm: động lượng tâm lý (momentum_sent) được tính bằng tích của lợi suất 3 ngày và điểm tâm lý trượt 5 ngày; cùng khối lượng theo tâm lý (volume_sent) được tính bằng tích của tỷ lệ khối lượng và điểm tâm lý trượt 5 ngày. Đồng thời, toàn bộ các giá trị khuyết thiếu còn sót lại trong nhóm biến cơ bản được lấp đầy một cách an toàn bằng giá trị trung vị của từng biến, đảm bảo không làm sai lệch xu hướng trung tâm của dữ liệu.

Cuối cùng, sau khi hoàn thiện các bước biến đổi, hệ thống tính năng được tái cấu trúc và phân nhóm lại một cách logic (bao gồm việc loại bỏ biến nhiễu stochastic_k và thiết lập lại các nhóm Phân tích Cơ bản, Phân tích Kỹ thuật, Tâm lý thị trường). Thống kê mô tả được thực hiện lại một lần nữa để xác nhận phân phối của dữ liệu đã đạt tiêu chuẩn trước khi chuyển sang giai đoạn huấn luyện mô hình (nơi phương pháp RobustScaler sẽ được áp dụng trực tiếp trong từng chu kỳ Cross-Validation để phòng tránh triệt để hiện tượng rò rỉ dữ liệu - data leakage).

3.5. Mô hình Mạng thần kinh nhân tạo đa lớp (MLP)

3.5.1. Cơ sở lý thuyết và kiến trúc mô hình

Nghiên cứu sử dụng Mạng thần kinh nhân tạo đa lớp (MLP) làm cấu trúc dự báo cốt lõi. Dựa trên Định lý xấp xỉ vạn năng (Hornik, 1989), mạng truyền thẳng MLP có khả năng xấp xỉ các hàm phi tuyến liên tục, cho phép mô hình hóa hiệu quả các mối quan hệ phức tạp giữa chỉ báo tài chính và giá cổ phiếu, qua đó khắc phục hạn chế của các mô hình kinh tế lượng truyền thống.

Về mặt kiến trúc, mô hình được thiết kế với kích thước giảm dần nhằm tối ưu hóa quá trình trích xuất đặc trưng. Lớp đầu vào có số nơ-ron tương đương số chiều của vector đặc trưng. Tín hiệu sau đó được truyền qua chuỗi bốn lớp ẩn liên tiếp với số lượng nơ-ron giảm dần lần lượt là: 256, 128, 64 và 32. Thiết kế dạng nén này hỗ trợ mô hình chọn lọc thông tin trọng yếu, giảm thiểu nhiễu và biểu diễn các đặc trưng ở mức độ trừu tượng cao. Tại mỗi nơ-ron, tín hiệu đầu ra được tính toán theo công thức tổng quát:

$$h_i = \sigma\left(\sum_j w_{ij}x_j + b_i\right)$$

trong đó, x_j đại diện cho tín hiệu đầu vào từ lớp trước đó, w_{ij}là trọng số liên kết, b_i là hệ số chệch và σ là hàm kích hoạt phi tuyến.

Để giải quyết vấn đề triệt tiêu đạo hàm thường gặp khi huấn luyện các mạng nơ-ron sâu, nghiên cứu sử dụng hàm kích hoạt ReLU (Rectified Linear Unit) với công thức $f(x) = \max(0, x)$ tại toàn bộ các lớp ẩn. ReLU không chỉ giúp quá trình tính toán diễn ra nhanh hơn mà còn hỗ trợ việc hội tụ của mô hình hiệu quả hơn so với các hàm Sigmoid hay Tanh truyền thống. Tại lớp đầu ra, mô hình sử dụng 2 nơ-ron tương ứng với hai trạng thái dự báo (Tăng/Giảm), kết hợp với hàm kích hoạt Softmax để chuyển đổi các giá trị đầu ra thành phân phối xác suất, đảm bảo tổng xác suất của các trường hợp luôn bằng 1.

3.5.2. Quá trình huấn luyện và tối ưu hóa

Cơ chế vận hành của quá trình huấn luyện dựa trên thuật toán lan truyền ngược, với mục tiêu tối cao là tối thiểu hóa sự sai biệt giữa dự báo của mô hình và thực tế. Đối với bài toán phân loại nhị phân, hàm mất mát được xác định là Binary Cross-Entropy, được biểu diễn qua phương trình:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

trong đó N là số lượng quan sát, $y_i$là nhãn thực tế và $\hat{y}_i$là xác suất được dự báo bởi mô hình.

Để giải bài toán tối ưu hóa hàm mất mát nêu trên, nghiên cứu áp dụng thuật toán Adam (Adaptive Moment Estimation). Đây là thuật toán tối ưu tiên tiến kết hợp được ưu điểm của cả Momentum và RMSProp, cho phép tự động điều chỉnh tốc độ học thích ứng cho từng tham số riêng biệt. Cơ chế này giúp mô hình vượt qua các điểm cực tiểu địa phương và hội tụ nhanh chóng, ổn định hơn trên bề mặt lỗi phức tạp của dữ liệu tài chính.

Song song với quá trình tối ưu hóa, vấn đề quá khớp một thách thức lớn khi áp dụng học máy vào dữ liệu chứng khoán nhiều nhiễu được kiểm soát chặt chẽ thông qua hai kỹ thuật điều chuẩn. Thứ nhất, kỹ thuật Dropout được áp dụng bằng cách ngẫu nhiên vô hiệu

hóa 20% số lượng nơ-ron trong mỗi bước huấn luyện, ngăn chặn việc các nơ-ron hình thành sự phụ thuộc lẫn nhau quá mức. Thứ hai, cơ chế Dừng sớm được thiết lập để theo dõi sai số trên tập kiểm định; quá trình huấn luyện sẽ tự động chấm dứt nếu hiệu suất không được cải thiện sau một khoảng kiên nhẫn là 20 vòng. Sự kết hợp này đảm bảo mô hình giữ lại được bộ tham số tối ưu nhất, cân bằng giữa khả năng học dữ liệu quá khứ và khả năng tổng quát hóa cho dữ liệu tương lai.

3.6. Phương pháp kiểm định và đánh giá

3.6.1. Chiến lược kiểm định chéo chuỗi thời gian

Một trong những thách thức cốt lõi khi mô hình hóa dữ liệu tài chính là tính phụ thuộc thời gian và sự tự tương quan của chuỗi dữ liệu. Do đó, việc áp dụng các phương pháp kiểm định chéo ngẫu nhiên truyền thống như K-Fold Cross-Validation được đánh giá là không phù hợp về mặt phương pháp luận, bởi kỹ thuật này phá vỡ cấu trúc thứ tự thời gian và dẫn đến hiện tượng rò rỉ dữ liệu (data leakage), tình trạng mô hình được huấn luyện trên thông tin tương lai để dự báo quá khứ. Để khắc phục hạn chế này và đảm bảo tính khách quan của kết quả thực nghiệm, nghiên cứu áp dụng chiến lược kiểm định chéo chuỗi thời gian cuốn chiếu (Rolling Time Series Split).

Cụ thể, tập dữ liệu tổng thể từ năm 2020 đến 2025 được phân chia thành các cửa sổ thời gian liên tiếp. Tại mỗi bước kiểm thử, tập huấn luyện được xác định bao gồm dữ liệu từ thời điểm bắt đầu cho đến thời điểm t, trong khi tập kiểm thử sẽ bao gồm các quan sát trong khoảng thời gian liền kề từ t đến $t + k$. Sau khi hoàn tất một vòng kiểm thử, cửa sổ thời gian được trượt tịnh tiến về phía trước; phần dữ liệu vừa được sử dụng để kiểm thử sẽ được gộp vào tập huấn luyện cho vòng tiếp theo. Quy trình mở rộng tập huấn luyện theo thời gian thực này giúp mô phỏng chính xác quy trình ra quyết định đầu tư trong thực tế, đảm bảo rằng tại bất kỳ thời điểm dự báo nào, mô hình chỉ được tiếp cận và học hỏi từ các thông tin đã phát sinh trong quá khứ, tuân thủ tuyệt đối nguyên tắc nhân quả trong phân tích tài chính.

3.6.2. Các thước đo hiệu suất

Để đánh giá toàn diện hiệu quả dự báo của mô hình MLP, nghiên cứu sử dụng hệ thống các chỉ số đo lường được xây dựng từ Ma trận nhầm lẫn (Confusion Matrix). Chỉ số cơ bản nhất là Độ chính xác (Accuracy), phản ánh tỷ lệ tổng số dự báo đúng trên toàn bộ

mẫu nghiên cứu. Tuy nhiên, trong bối cảnh đầu tư tài chính nơi chi phí cơ hội và rủi ro thua lỗ có trọng số khác nhau, nghiên cứu chú trọng sâu hơn vào Độ chính xác của lớp tích cực (Precision) và Độ nhạy (Recall). Precision đo lường tỷ lệ các dự báo tăng giá là chính xác, đóng vai trò quan trọng trong việc giảm thiểu rủi ro mua sai (False Positive), trong khi Recall phản ánh khả năng của mô hình trong việc không bỏ lỡ các cơ hội tăng giá thực tế trên thị trường.

Để cân bằng giữa hai mục tiêu mâu thuẫn là tối đa hóa độ chính xác và độ bao phủ, điểm F1 (F1-Score) trung bình điều hòa giữa Precision và Recall được sử dụng làm thước đo tổng hợp. Bên cạnh đó, khả năng phân loại tổng quát của mô hình được đánh giá thông qua diện tích dưới đường cong đặc trưng hoạt động (AUC-ROC). Với giá trị dao động từ 0.5 (ngẫu nhiên) đến 1.0 (hoàn hảo), AUC cung cấp một cái nhìn khách quan về năng lực xếp hạng cổ phiếu của mô hình tại các ngưỡng xác suất cắt (threshold) khác nhau, không phụ thuộc vào một điểm cắt cụ thể nào, qua đó khẳng định độ bền vững của khả năng dự báo.

3.6.3. Kiểm định thống kê và phân tích so sánh

Nhằm đảm bảo độ tin cậy khoa học và tính vững chắc của các kết luận, nghiên cứu thực hiện các kiểm định thống kê suy diễn chuyên sâu. Phương pháp Bootstrap Confidence Intervals được áp dụng để ước lượng khoảng tin cậy 95% cho chỉ số AUC. Thông qua quy trình lấy mẫu lại có hoàn lại 1.000 lần trên tập kết quả dự báo, nghiên cứu xác định được biên độ dao động của hiệu suất, từ đó đánh giá mức độ ổn định của mạng nơ-ron trước các nhiễu loạn ngẫu nhiên.

Nhằm đánh giá hiệu quả của việc tích hợp đa nguồn dữ liệu, nghiên cứu tiến hành phân tích cắt bỏ (Ablation Study). Quá trình huấn luyện và kiểm thử được thực hiện trên 7 tổ hợp không gian đặc trưng, bao gồm các nhóm biến đơn lẻ (Kỹ thuật, Cơ bản, Tâm lý) và các mô hình lai ghép. Sự khác biệt về hiệu suất dự báo giữa các tổ hợp được đánh giá thông qua kiểm định phi tham số Wilcoxon Signed-Rank Test với mức ý nghĩa thống kê $p < 0,05$. Cuối cùng, để khắc phục tính hộp đen của học sâu, phương pháp SHAP (SHapley Additive exPlanations) dựa trên lý thuyết trò chơi được ứng dụng nhằm lượng hóa chính xác mức độ đóng góp biên của từng biến số vào quyết định dự báo, minh bạch hóa cơ chế suy luận bên trong của mô hình.

CHƯƠNG 4. KẾT QUẢ PHÂN TÍCH VÀ THẢO LUẬN

4.1. Thống kê mô tả

Để đảm bảo chất lượng dữ liệu trước khi đưa vào các mô hình học máy, nghiên cứu bắt đầu bằng việc phân tích thống kê mô tả chi tiết. Quá trình này đánh giá phân phối tĩnh của biến phụ thuộc (bao gồm biến phân loại mục tiêu và biến liên tục thể hiện tỷ suất sinh lời) cùng toàn bộ các biến độc lập. Các đại lượng thống kê cơ bản được trích xuất bao gồm số lượng quan sát, giá trị trung bình, độ lệch chuẩn, các điểm cực trị, mức độ khuyết thiếu và độ lệch phân phối. Bước phân tích này đóng vai trò nền tảng giúp nhận diện những bất thường trong dữ liệu ban đầu, làm cơ sở vững chắc cho các quyết định làm sạch tiếp theo.

Kết quả thống kê mô tả ở Bảng 6 và Bảng 7 dưới đây, cho thấy tập dữ liệu có quy mô lớn với 50.829 quan sát và đạt chất lượng độ sạch rất cao. Đối với biến phụ thuộc, phân phối của biến mục tiêu phân loại (target) đạt trạng thái cân bằng lý tưởng với 50,99% quan sát thuộc lớp 1 và 49,01% thuộc lớp 0. Sự cân bằng này tạo điều kiện tối ưu cho các thuật toán học máy học hỏi mà không bị thiên lệch, loại bỏ sự cần thiết phải áp dụng các kỹ thuật tái lấy mẫu. Trong khi đó, biến mục tiêu hồi quy (future_return) ghi nhận mức lợi suất kỳ vọng trung bình là 0,92%, với độ lệch chuẩn 8,81% và biên độ dao động từ -8,19% đến 51,69%, phản ánh đầy đủ các chu kỳ biến động đa dạng của thị trường. Về phía các biến độc lập, sau các bước biến đổi, hầu hết các nhóm biến phân tích cơ bản, kỹ thuật và tâm lý đều có độ lệch phân phối (skewness) nằm trong ngưỡng kiểm soát. Đặc biệt, các chỉ báo đại diện cho tâm lý thị trường (sent_score_roll, sent_shock) đều dao động quanh giá trị 0, cho thấy trạng thái cảm xúc chung của nhà đầu tư xuyên suốt tập dữ liệu ở mức độ trung lập và các cú sốc phân bổ đối xứng ở cả hai chiều.

Bảng 5. Thống kê mô tả các biến đặc trưng

Biến	Count	Mean	Std	Min	25%	Median	75%	Max	Skew
close	50829	10.0586	0.7372	7.5496	9.5418	10.0328	10.6100	12.3445	0.0098
volume	50829	14.3855	1.7330	0.6931	13.4273	14.5318	15.6270	19.2214	-0.8758
eps	50829	6.2965	1.0608	2.9014	5.7267	6.4564	7.0051	8.3018	-0.8403
pb	50829	1.0856	0.3862	0.3842	0.8214	1.0288	1.3004	2.2893	0.8079
pe	50829	4.1356	0.9323	2.5253	3.5037	4.0126	4.5655	7.4867	1.1309
lnst_yoy	50829	0.3086	0.7383	-0.6849	-0.1964	0.1805	0.5724	3.1315	1.5297
nophaitra_vcsh	50829	1.0792	0.7323	0.1074	0.5531	0.8980	1.3616	2.9236	0.9553
vonhoa_tts	50829	0.5347	0.3305	0.0564	0.2810	0.5011	0.7335	1.5406	0.6958
profitability	50829	-0.0001	0.9873	-3.5266	-0.7328	-0.1324	0.6079	7.4097	0.6615
rsi_14	50829	52.6826	18.2351	0.0000	39.3443	52.9520	66.1290	100.000	-0.0515
volume_ratio	50829	0.6817	0.2474	0.2013	0.5090	0.6485	0.8179	1.4768	0.7631
volatility_20d	50827	0.0256	0.0107	0.0021	0.0175	0.0242	0.0322	0.0995	0.7522
sma20_gap	50829	0.0080	0.0790	-0.5265	-0.0308	0.0082	0.0483	0.5087	0.0254
sma_cross	50829	0.0188	0.1073	-0.6095	-0.0350	0.0188	0.0750	0.6554	0.0174
ret_5d_rank	50829	0.5191	0.2973	0.0100	0.2600	0.5253	0.7849	1.0000	-0.0523

momentum_vol	50827	0.0002	0.0024	-0.0281	-0.0006	0.0001	0.0009	0.0344	0.5636
ret_3d	50829	0.0033	0.0526	-0.2915	-0.0220	0.0025	0.0292	0.3998	0.0274
momentum_sent	50829	-0.0003	0.0218	-0.2424	-0.0031	0.0000	0.0025	0.3083	0.0981
volume_sent	50829	-0.0243	0.2738	-1.2279	-0.1581	0.0000	0.0444	1.4404	0.3850
sent_score_roll	50829	-0.0349	0.3773	-0.8765	-0.2673	0.0000	0.0727	0.9804	0.3869
sent_shock	50829	0.0006	0.3790	-1.4677	-0.1245	0.0000	0.1119	1.4796	0.0944
ret_5d_rank	50829	0.5191	0.2973	0.0100	0.2600	0.5253	0.7849	1.0000	-0.0523
news_dummy	50829	0.1694	0.3751	0.0000	0.0000	0.0000	0.0000	1.0000	1.7625

Nguồn: Tác giả tự phân tích và tổng hợp

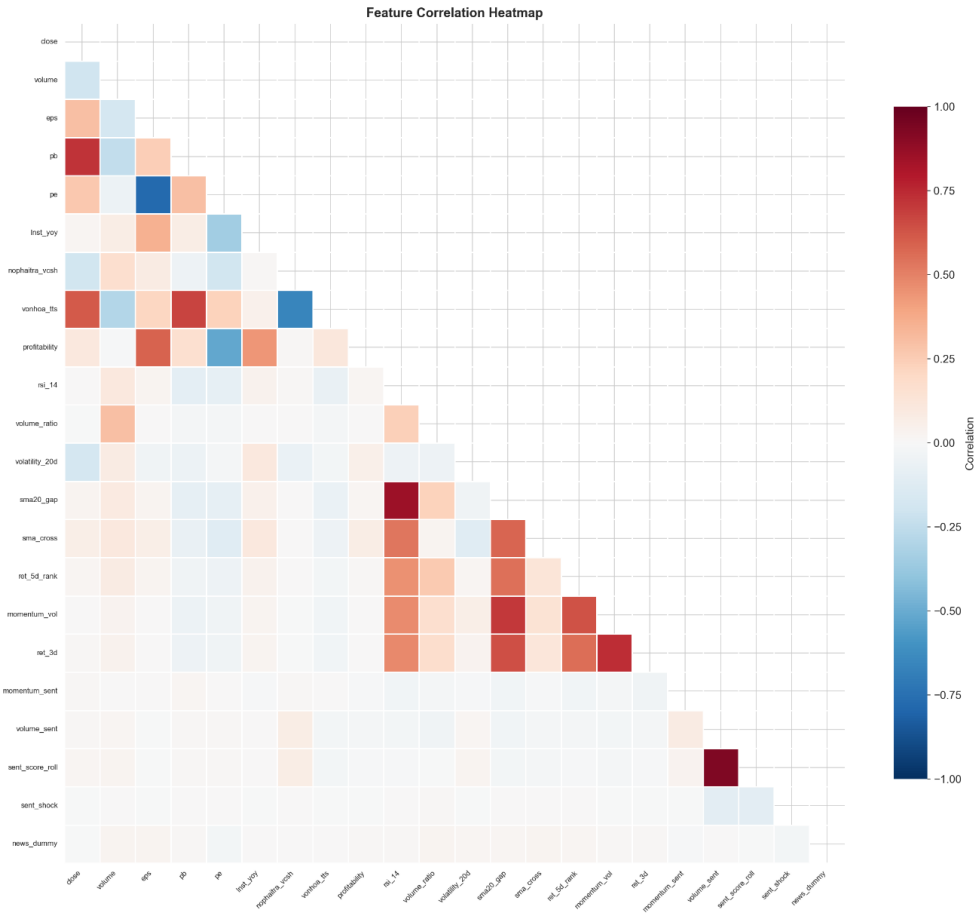
Bảng 6. Thống kê mô tả các biến mục tiêu

Biến	Count	Mean	Std	Min	25%	Median	75%	Max	Value Counts
target	50829	0.5099	0.4999	0	0	1	1	1	{0: 24911, 1: 25918}
future_return	50829	0.0092	0.0881	-0.3819	-0.0550	0.0433	0.0720	0.5169	*Không áp dụng*

Nguồn: Tác giả tự phân tích và tổng hợp

4.2. Phân tích tương quan đặc trưng

Nghiên cứu tiến hành đánh giá mức độ tương quan tuyến tính giữa các biến độc lập thông qua ma trận tương quan. Việc phân tích này đóng vai trò thiết yếu trong việc nhận diện hiện tượng đa cộng tuyến, qua đó cung cấp cơ sở khoa học cho quá trình chọn lọc đặc trưng và tối ưu hóa các thuật toán học máy. Về tổng thể, các hệ số tương quan thấp chiếm ưu thế trên toàn bộ không gian ma trận, cho thấy phần lớn các biến độc lập không có sự tương quan tuyến tính mạnh với nhau. Điều này khẳng định rằng mỗi chỉ báo được thiết kế đều phản ánh một khía cạnh thông tin riêng biệt, góp phần mở rộng không gian đặc trưng mà không gây nhiễu loạn hay trùng lặp dữ liệu.



Hình 1. Biểu đồ nhiệt tự tương quan các biến

Phân tích chi tiết vào nhóm biến cơ bản, kết quả cho thấy một số mối liên hệ tuyến tính đặc thù hoàn toàn đồng nhất với lý thuyết tài chính truyền thống. Cụ thể, ma trận ghi nhận mối tương quan nghịch biến rõ nét giữa hệ số định giá P/E và thu nhập trên cổ phần (EPS), cũng như tương quan thuận biến ở mức cao giữa hệ số P/B, quy mô vốn hóa thị

trường và mức giá đóng cửa. Tương tự, biến tổng hợp đại diện cho khả năng sinh lời cốt lõi (profitability) cũng thể hiện mối tương quan âm với hệ số định giá P/E, phản ánh thực trạng rằng các doanh nghiệp có tỷ suất sinh lời cao trong mẫu nghiên cứu thường duy trì mức định giá tương đối thấp hoặc ở mức hợp lý. Sự tồn tại của các mối tương quan này xác nhận tính logic và độ tin cậy của tập dữ liệu sau khi trải qua các quy trình tiền xử lý nghiêm ngặt.

Trái ngược với nhóm biến cơ bản, đối với nhóm biến phân tích kỹ thuật và động lượng, kết quả ghi nhận một cụm tương quan thuận biến ở mức tương đối cao. Cụm liên kết này bao gồm các chỉ báo như sức mạnh tương đối (RSI 14), khoảng cách đến đường trung bình động 20 ngày (SMA20 gap), tín hiệu giao cắt của đường trung bình (SMA cross) cùng các biến lợi suất ngắn hạn (lợi suất 3 ngày, động lượng khối lượng). Sự gắn kết chặt chẽ này là một hệ quả toán học tất yếu, xuất phát từ việc các chỉ báo này đều được dẫn xuất trực tiếp từ chuỗi thời gian của mức giá và khối lượng trong các khung thời gian tiệm cận nhau. Mặc dù mức độ tương quan nội bộ cao có thể gây ra thách thức về độ trễ và sai số đối với các mô hình hồi quy tuyến tính cổ điển, nhưng đối với các kiến trúc học máy phi tuyến tính mà nghiên cứu áp dụng, tập hợp các đặc trưng này vẫn cung cấp những tín hiệu dự báo giá trị mà không bị suy giảm hiệu suất bởi hiện tượng cộng tuyến.

Điểm đáng chú ý nhất trong phân tích tương quan là đặc tính của nhóm biến đại diện cho tâm lý thị trường. Các chỉ báo tâm lý, bao gồm điểm tâm lý trượt, cú sốc tâm lý, động lượng tâm lý và biến giả tin tức, gần như không thể hiện bất kỳ sự tương quan tuyến tính nào với hai nhóm biến cơ bản và kỹ thuật. Đặc tính trực giao này là một phát hiện có ý nghĩa quan trọng, minh chứng rõ ràng cho việc tích hợp dữ liệu phân tích văn bản và tâm lý học hành vi đã cung cấp một nguồn thông tin hoàn toàn độc lập với các dữ liệu giao dịch định lượng truyền thống.

4.3. Kết quả thực nghiệm so sánh các nhóm biến

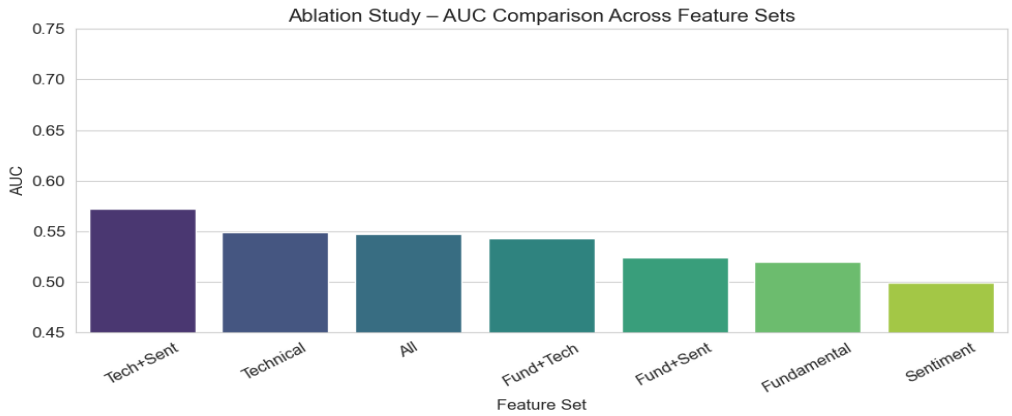
Để xác định tổ hợp dữ liệu tối ưu và đánh giá đóng góp biên của từng nguồn thông tin, nghiên cứu đã thực hiện 7 kịch bản thử nghiệm độc lập. Kết quả chi tiết về các thước đo hiệu suất trên tập kiểm thử được tổng hợp trong Bảng 8 dưới đây.

Bảng 7. Tổng hợp hiệu suất các mô hình dự báo trên tập kiểm thử

Model	Accuracy	AUC	Precision	Recall	F1	AP
Fundamental	0.512	0.520	0.502	0.607	0.550	0.508
Technical	0.536	0.549	0.529	0.491	0.509	0.549
Sentiment	0.494	0.499	0.491	0.905	0.637	0.490
Fund+Tech	0.535	0.543	0.529	0.470	0.498	0.541
Fund+Sent	0.513	0.524	0.503	0.587	0.542	0.512
Tech+Sent	0.554	0.573	0.553	0.473	0.510	0.564
All	0.536	0.547	0.535	0.418	0.469	0.540

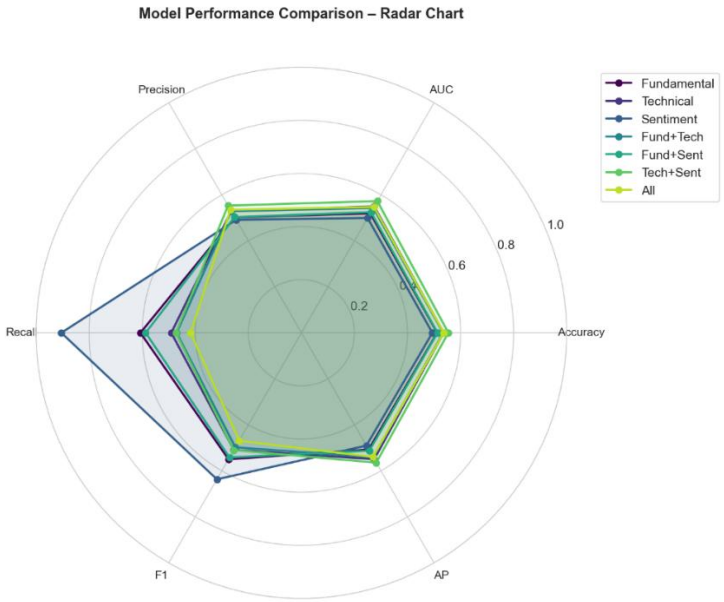
Nguồn: Tác giả tự phân tích và tổng hợp

Kết quả thực nghiệm từ phương pháp phân tích cắt bỏ (Ablation Study) cung cấp những bằng chứng định lượng quan trọng về mức độ đóng góp của từng lăng kính thông tin vào năng lực dự báo xu hướng giá cổ phiếu ngắn hạn. Đánh giá trên các nhóm đặc trưng đơn lẻ, lăng kính Phân tích kỹ thuật (Technical) thể hiện sức mạnh phân lớp tốt nhất với hệ số AUC đạt 0,549 và Độ chính xác (Accuracy) đạt 0,536. Trái lại, nhóm biến Phân tích cơ bản (Fundamental) chỉ đạt mức AUC khiêm tốn 0,519, tiệm cận với tỷ lệ của một bộ phân loại ngẫu nhiên. Đáng chú ý, nhóm chỉ báo Tâm lý thị trường (Sentiment) khi đứng độc lập lại cho kết quả dự báo kém nhất với độ đo AUC rơi xuống ngưỡng 0,499. Sự bất thường của nhóm biến này còn thể hiện qua điểm Recall cao đột biến (0,905) đi kèm với Precision thấp (0,491), cho thấy mô hình khi chỉ học từ tin tức đã bị rơi vào trạng thái thiên lệch dự báo (Model Bias), có xu hướng dự báo tăng giá cho hầu hết các quan sát mà thiếu đi khả năng chọn lọc tín hiệu.



Hình 2. Biểu đồ so sánh chỉ số AUC giữa các tổ hợp đặc trưng

Sự bứt phá về năng lực dự báo của mạng nơ-ron chỉ thực sự xuất hiện khi các lăng kính thông tin được lai ghép với nhau. Tổ hợp Kỹ thuật và Tâm lý (Tech+Sent) đã chứng minh sự vượt trội toàn diện khi xác lập mức hiệu suất cao nhất trên toàn bộ các thang đo: AUC đạt 0,573, Độ chính xác đạt 0,554 và điểm Average Precision (AP) đạt 0,564. Phát hiện này củng cố mạnh mẽ cho giả thuyết tài chính hành vi đã thiết lập: hành vi giao dịch thuần túy (Technical) mang lại bộ lọc tín hiệu về giá, nhưng chính sự xác nhận từ tâm lý đám đông (Sentiment) mới đóng vai trò là chất xúc tác để xác định một đà tăng trưởng có độ tin cậy cao. Sự đồng thuận giữa biến động dòng tiền và bức tranh truyền thông đã triệt tiêu hiệu quả các tín hiệu nhiễu (Whipsaws) vốn thường xuyên xảy ra nếu chỉ thuần túy quan sát biểu đồ giá.



Hình 3. Biểu đồ Radar đánh giá toàn diện các thang đo hiệu suất của các tổ hợp biến

Một phát hiện mang tính phản trực giác (Counter-intuitive) nhưng cực kỳ thú vị từ kết quả thực nghiệm là hiện tượng lời nguyền dữ liệu khi tích hợp nhóm Phân tích cơ bản. Trái với kỳ vọng ban đầu, việc kết hợp nhóm Phân tích cơ bản vào nhóm Kỹ thuật (Fund+Tech) lại làm suy giảm hệ số AUC từ 0,549 xuống còn 0,543. Càng rõ ràng hơn, khi tích hợp toàn bộ các nguồn dữ liệu vào tổ hợp toàn phần (All Features), hiệu suất tổng thể của mô hình sụt giảm đáng kể (AUC chỉ còn 0,547) so với mô hình tối ưu Tech+Sent.

Hiện tượng này có thể được giải thích thông qua cấu trúc vi mô của thị trường và độ trễ của thông tin. Bài toán của nghiên cứu được thiết lập để dự báo tỷ suất sinh lời trong khung thời gian cực ngắn (5 ngày giao dịch). Trong khung thời gian này, các biến số cơ

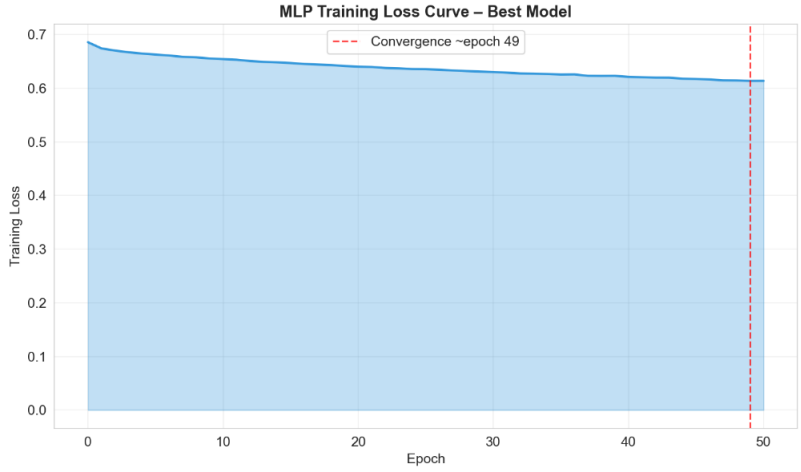
bản (như $P/E, P/B$ hay ROE) mang đặc tính khá tĩnh do chỉ được cập nhật theo quý. Việc cố gắng ép mạng nơ-ron học các dữ liệu tĩnh này để dự đoán các biến động vi mô hàng ngày không những không mang lại giá trị thông tin biên (Marginal information) mà còn đưa thêm nhiễu (Noise) vào hệ thống. Hệ quả tất yếu là rủi ro quá khớp (Overfitting) cấu trúc gia tăng, làm mờ nhạt đi các tín hiệu động lượng đang trực tiếp chi phối thị trường.

4.4. Phân tích chi tiết mô hình tối ưu

Sau khi xác lập tổ hợp Kỹ thuật và Tâm lý (Tech+Sent) là không gian đặc trưng tối ưu với hệ số AUC đạt 0,573, nghiên cứu tiến hành giải mã hộp đen của mạng nơ-ron đa lớp (MLP). Quá trình này nhằm đánh giá chi tiết hành vi học, năng lực phân lớp, mức độ đóng góp của các biến số và sự phân hóa hiệu suất dự báo giữa các nhóm ngành trên thị trường.

4.4.1. Đánh giá khả năng hội tụ và phân lớp của mô hình

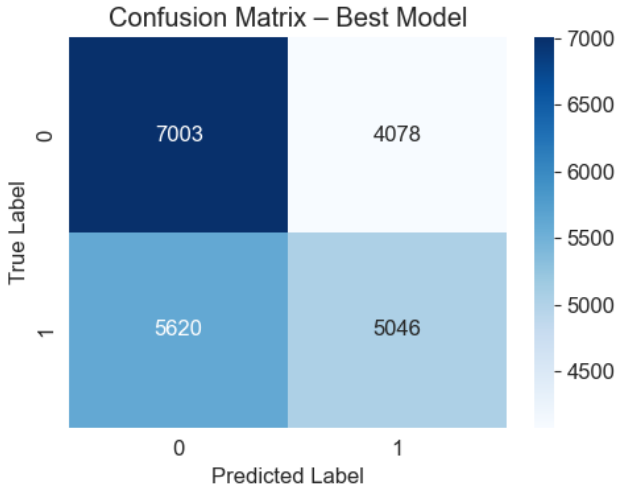
Khảo sát đường cong học tập (Learning Curve) cho thấy thuật toán tối ưu hóa Adam đã kiểm soát quá trình cập nhật trọng số một cách trơn tru và ổn định. Hàm mất mát (Loss function) trên tập huấn luyện suy giảm đều đặn từ mức 0,68 và chính thức hội tụ tại chu kỳ (epoch) thứ 49. Tại ngưỡng này, cơ chế Dừng sớm (Early Stopping) được kích hoạt kịp thời, giúp ngăn chặn mạng nơ-ron rơi vào trạng thái quá khớp (overfitting), qua đó duy trì độ khái quát hóa cao nhất cho mô hình trên các tập dữ liệu chưa từng quan sát.



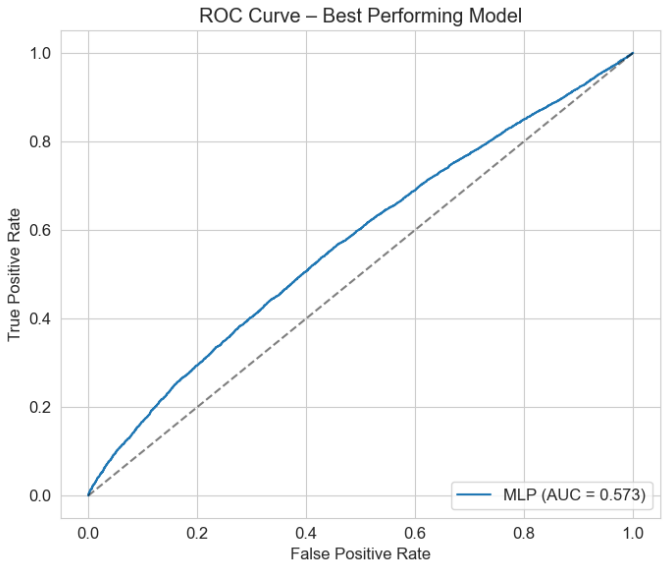
Hình 4. Đường cong mất mát trong quá trình huấn luyện mô hình Tech+Sent

Đi sâu phân tích Ma trận nhầm lẫn (Confusion Matrix), mô hình bộc lộ thiên hướng dự báo thận trọng, hoàn toàn phù hợp với triết lý quản trị rủi ro trong đầu tư tài chính. Cụ

thể, số lượng dự báo sai tích cực (False Positives) được nén lại ở mức 4.078 quan sát, thấp hơn đáng kể so với 5.620 quan sát thuộc nhóm dự báo sai tiêu cực (False Negatives). Đặc tính phân phối này giúp mô hình đạt Độ chuẩn xác (Precision) là 0,553, vượt trội hơn hẳn so với Độ nhạy (Recall). Ý nghĩa kinh tế của kết quả này vô cùng sâu sắc: thuật toán thà chấp nhận bỏ lỡ một số cơ hội tăng giá chưa thực sự rõ ràng (chấp nhận Recall thấp) để đổi lấy việc hạn chế tối đa các tín hiệu mua sai lầm có thể gây tổn thất vốn cho nhà đầu tư (ưu tiên Precision cao). Đồng thời, đường cong ROC luôn duy trì vị thế nằm vắt ngang trên đường chéo ngẫu nhiên xuyên suốt mọi ngưỡng cắt xác suất, một lần nữa tái khẳng định năng lực phân loại xu hướng vững chắc của tổ hợp lai ghép.



Hình 5. Ma trận nhầm lẫn (Confusion Matrix) của mô hình phân lớp tối ưu

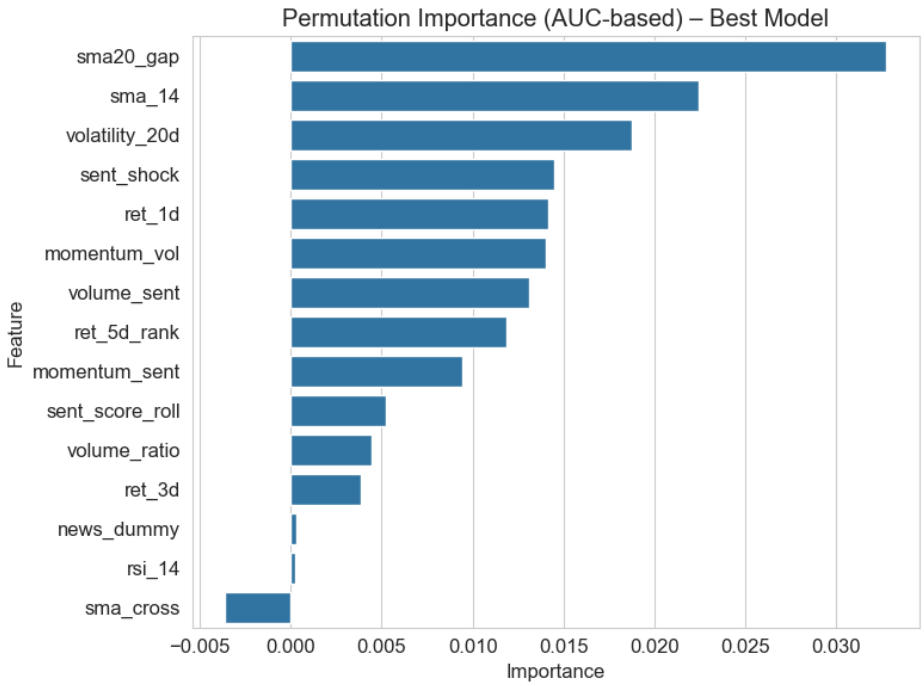


Hình 6. Đường cong ROC và hệ số AUC của mô hình Tech+Sent

4.4.2. Đánh giá mức độ quan trọng của đặc trưng

Để lượng hóa mức độ đóng góp của từng biến số vào kết quả phân lớp, nghiên cứu áp dụng kỹ thuật Hoán vị đặc trưng (Permutation Importance) dựa trên sự suy giảm của thang đo hiệu suất AUC. Kết quả phân tích mang lại những phát hiện mang tính đột phá về cấu trúc vi mô của thị trường trong ngắn hạn. Dẫn đầu tuyệt đối về sức mạnh dự báo là biến sma20_gap (mức chênh lệch giữa giá hiện tại và đường trung bình động 20 ngày). Phát hiện này cung cấp bằng chứng thực nghiệm vững chắc cho thấy hiệu ứng Đảo chiều về giá trị trung bình (Mean Reversion) chính là quy luật vận động chi phối mạnh mẽ nhất trên thị trường chứng khoán Việt Nam; cụ thể, khi giá cổ phiếu bị kéo dãn quá đà so với nền tảng trung hạn, một lực hút điều chỉnh tất yếu sẽ xuất hiện. Đứng ở các vị trí tiếp theo trong thang đo tầm quan trọng là đường trung bình sma_14 và biến động rủi ro volatility_20d.

Đặc biệt, kết quả phân tích tầm quan trọng là minh chứng đanh thép cho sự thành công của chiến lược tích hợp biến tâm lý vào không gian đặc trưng. Điển hình, biến Cú sốc tâm lý (sent_shock) vươn lên vị trí thứ 4 trong rổ đặc trưng, vượt qua hàng loạt các chỉ báo giá truyền thống. Hơn thế nữa, các biến tương tác lai ghép như Khối lượng theo tâm lý (volume_sent) và Động lượng theo tâm lý (momentum_sent) cũng ghi nhận sự góp mặt trong nhóm các biến có sức mạnh phân loại cao nhất. Trái ngược lại, chỉ báo giao cắt trung bình động (sma_cross) lại ghi nhận mức đóng góp âm, minh chứng cho việc các tín hiệu xu hướng chậm (Lagging indicators) thường sinh ra tín hiệu nhiễu thay vì mang lại giá trị dự báo trong một dải thời gian ngắn hạn 5 ngày. Kết quả này tái khẳng định một cách mạnh mẽ luận điểm: một đà tăng giá dù đi kèm với khối lượng bùng nổ cũng chỉ thực sự đáng tin cậy nếu nó được cộng hưởng bởi một cú sốc truyền thông tích cực từ thị trường.



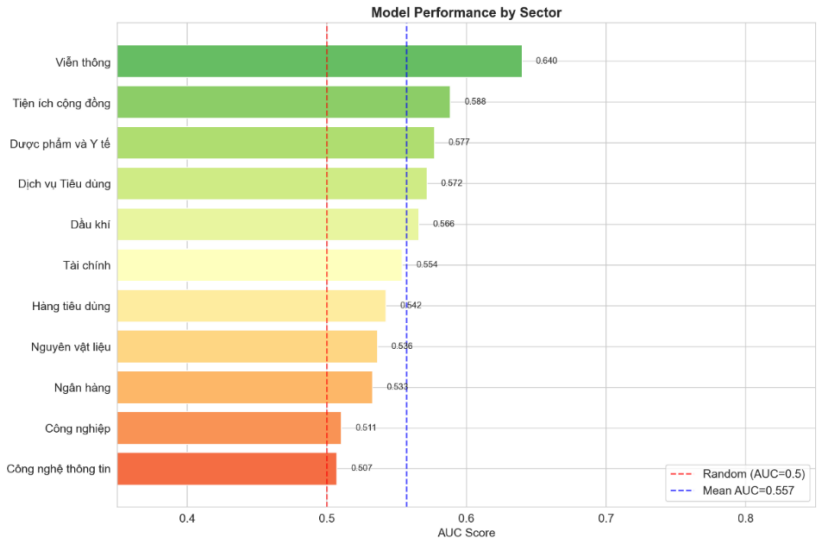
Hình 7. Mức độ đóng góp của các biến số (Permutation Importance) trong mô hình tối ưu

4.4.3. Phân tích hiệu suất theo nhóm ngành (Sector Performance)

Tiến hành kiểm định kết quả phân lớp theo từng nhóm ngành kinh tế, hiệu suất dự báo của mô hình bộc lộ sự phân hóa vô cùng rõ rệt. Cụ thể, thể hiện qua Hình 9 dưới đây, mô hình đạt chỉ số AUC cao nhất tại các nhóm ngành sở hữu tính ổn định cao trong hoạt động kinh doanh, điển hình như Viễn thông (0,640), Tiện ích cộng đồng (0,588) và Dược phẩm & Y tế (0,577). Kết quả này minh chứng rằng các cổ phiếu mang tính phòng thủ thường có cấu trúc biến động giá tuân theo các quy luật kỹ thuật và phản ứng với tâm lý đám đông một cách tương đối nhất quán trong ngắn hạn.

Trái ngược với nhóm phòng thủ, mạng nơ-ron gặp nhiều thách thức và ghi nhận mức hiệu suất sụt giảm đáng kể khi dự báo các nhóm ngành mang tính chu kỳ cao, có cấu trúc phức tạp và quy mô vốn hóa lớn như Ngân hàng (0,533), Công nghiệp (0,511) và Công nghệ thông tin (0,507). Sự chênh lệch hiệu suất này có thể được lý giải bởi bản chất vận động của các cổ phiếu đầu ngành, đặc biệt là nhóm tài chính - ngân hàng. Quỹ đạo giá của nhóm này bị chi phối nặng nề bởi các yếu tố kinh tế vĩ mô mang tính hệ thống (như chính sách tiền tệ, dư địa tăng trưởng tín dụng, hay biến động tỷ giá), những tham số định lượng phức tạp vốn chưa được phản ánh đầy đủ trong giới hạn của không gian đặc trưng về Kỹ thuật và luồng tin tức Tâm lý thông thường.

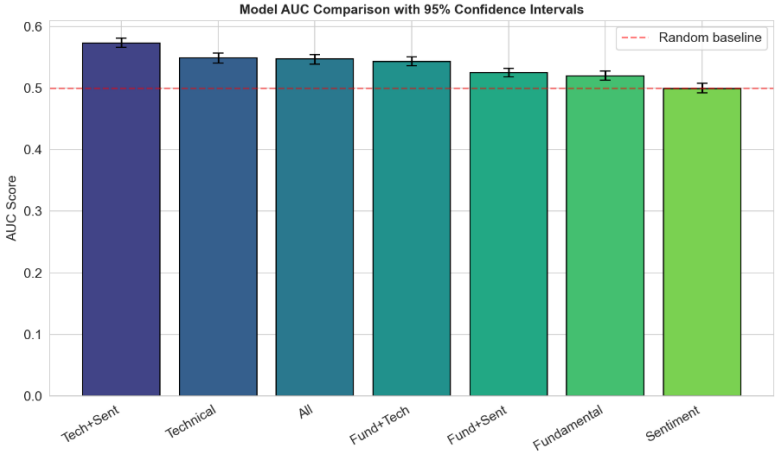
Từ những quan sát thực nghiệm trên, nghiên cứu rút ra một hàm ý quan trọng dành cho các nhà thực hành tài chính định lượng: hoàn toàn không tồn tại một mô hình chén thánh (Holy Grail) chung mang lại hiệu quả đồng nhất cho toàn bộ thị trường. Thay vào đó, việc ứng dụng các thuật toán học máy đòi hỏi tư duy tùy biến chiến lược sâu sắc, trong đó không gian đặc trưng cần được thiết kế và tinh chỉnh linh hoạt sao cho bám sát với đặc thù cấu trúc thông tin của từng hệ sinh thái ngành riêng biệt.



Hình 8. Phân hóa hiệu suất dự báo (AUC) của mô hình theo các nhóm ngành

4.5. Kiểm định độ bền vững và thống kê

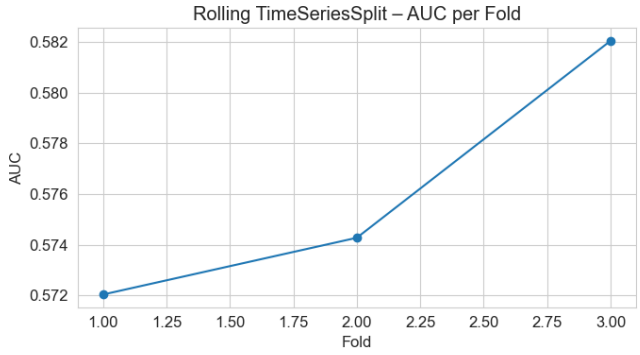
Để củng cố độ tin cậy của kết quả thực nghiệm và loại trừ khả năng hiệu suất phát sinh do tính ngẫu nhiên của mẫu dữ liệu, nghiên cứu tiến hành ước lượng khoảng tin cậy (Confidence Interval) cho hệ số AUC thông qua phương pháp lấy mẫu lại Bootstrap.



Hình 9. Hệ số AUC của các tổ hợp mô hình với khoảng tin cậy 95%

Kết quả kiểm định Bootstrap với 1.000 lần lấy mẫu lại cho thấy mô hình tối ưu (Tech+Sent) đạt giá trị AUC trung bình là 0,5727, với khoảng tin cậy 95% dao động trong vùng [0,5659; 0,5804]. Việc cận dưới của khoảng tin cậy (0,5659) duy trì một khoảng cách an toàn so với ngưỡng phân loại ngẫu nhiên (0,500) đã xác nhận năng lực dự báo của tổ hợp Kỹ thuật và Tâm lý là thực sự có ý nghĩa thống kê. Biên độ dao động tương đối hẹp của khoảng tin cậy cũng phản ánh việc mô hình không bị phụ thuộc quá mức vào bất kỳ một tập con dữ liệu cụ thể nào, qua đó đảm bảo tính vững (Robustness) khi ứng dụng trên các tập dữ liệu ngoài mẫu.

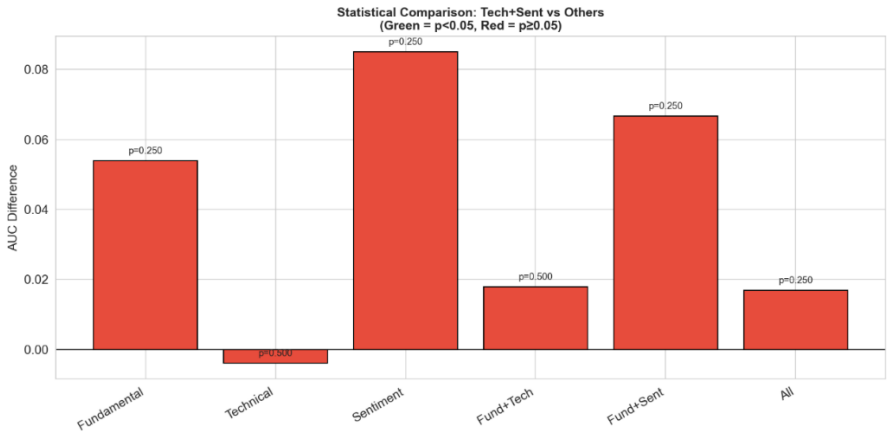
Bên cạnh đó, nghiên cứu tiến hành kiểm định độ bền vững thông qua chiến lược kiểm định chéo chuỗi thời gian cuốn chiếu (Rolling TimeSeriesSplit). Đường cong biểu diễn hệ số AUC trên 3 nếp gấp (folds) thời gian liên tiếp cho thấy hiệu suất của mô hình tối ưu (Tech+Sent) không những duy trì sự ổn định mà còn có xu hướng cải thiện, tăng từ mức 0,572 ở nếp gấp đầu tiên lên 0,582 ở nếp gấp thứ ba. Kết quả này khẳng định mạng nơ-ron không bị rơi vào trạng thái quá khớp (Overfitting) đối với một giai đoạn thị trường cụ thể, đồng thời giữ vững năng lực tổng quát hóa và khả năng dự báo khi tập dữ liệu kiểm thử được trượt dần về tương lai.



Hình 10. Kết quả chỉ số AUC qua các nếp gấp thời gian (Rolling TimeSeriesSplit)

Cuối cùng, nghiên cứu áp dụng kiểm định phi tham số Wilcoxon nhằm đánh giá ý nghĩa thống kê về sự chênh lệch hiệu suất giữa tổ hợp tối ưu (Tech+Sent) và các kịch bản còn lại. Kết quả trên biểu đồ phân tích cho thấy sự chênh lệch tuyệt đối về giá trị AUC là hiện hữu. Tuy nhiên, các giá trị p-value tương ứng (ví dụ: so với nhóm Fundamental $p = 0{,}250$, so với nhóm All $p = 0{,}250$) đều lớn hơn mức ý nghĩa $\alpha = 0{,}05$. Dữ liệu này chỉ ra rằng, mặc dù việc tích hợp Kỹ thuật và Tâm lý mang lại điểm số AUC cao nhất về mặt thực nghiệm, nhưng ở độ tin cậy 95%, chưa có đủ bằng chứng thống kê để khẳng định sự

vượt trội tuyệt đối của tổ hợp này so với các nhóm biến khác. Việc thừa nhận kết quả này thể hiện sự minh bạch và tính thận trọng học thuật cần thiết trong nghiên cứu định lượng.

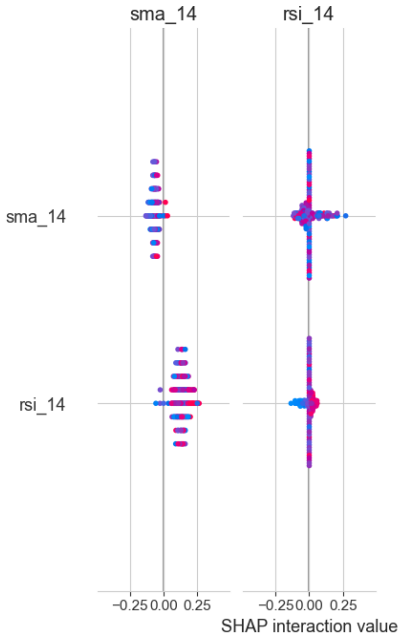


Hình 11. Kết quả kiểm định thống kê sự khác biệt AUC giữa tổ hợp Tech+Sent và các tổ hợp khác

4.6. Phân tích tương tác đặc trưng và sự phân hóa theo ngành

4.6.1. Phân tích tương tác đặc trưng thông qua mô hình SHAP

Nhằm khắc phục rào cản về khả năng diễn giải vốn là một hạn chế cố hữu của các kiến trúc mạng nơ-ron hộp đen, nghiên cứu ứng dụng phương pháp SHAP (SHapley Additive exPlanations) để bóc tách cơ chế tác động nội tại và mức độ đóng góp của các biến số.

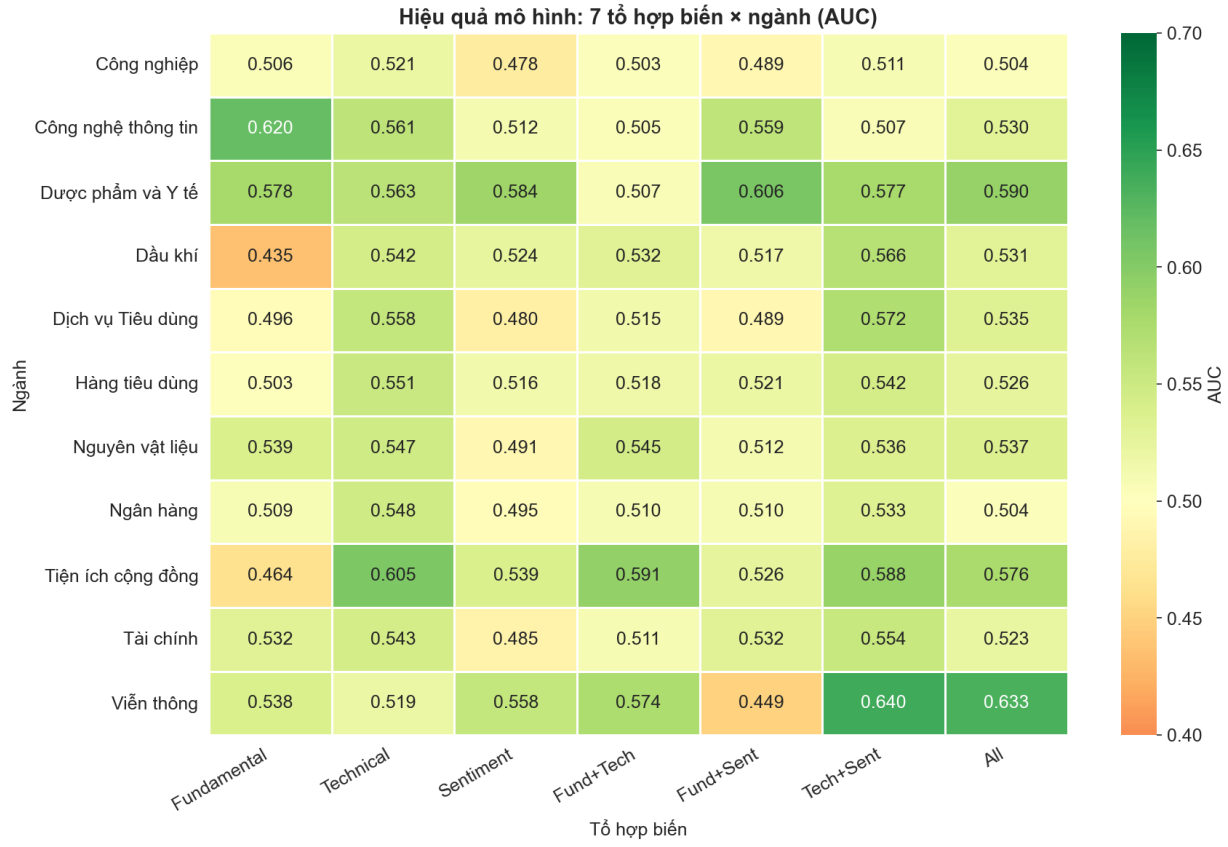


Hình 12. Biểu đồ tương tác SHAP của các chỉ báo kỹ thuật SMA_14 và RSI_14

Phân tích giá trị tương tác SHAP đối với các chỉ báo kỹ thuật cốt lõi như sma_14 và rsi_14 bộc lộ một mối quan hệ phân tán phi tuyến tính vô cùng rõ rệt. Thay vì hội tụ thành một đường xu hướng tuyến tính đơn điệu, các điểm dữ liệu phân bổ trải rộng dọc theo trục giá trị SHAP tương tác. Kết quả thực nghiệm này phản ánh chính xác bản chất phức tạp của vi cấu trúc thị trường: tác động của động lượng giá lên tỷ suất sinh lời trong tương lai không phải là một hằng số cố định. Ngược lại, cường độ và chiều hướng của tác động này biến thiên liên tục, phụ thuộc chặt chẽ vào điều kiện biên và sự cộng hưởng của các đặc trưng đồng thời khác tại từng thời điểm quan sát.

4.6.2. Phân tích hiệu suất theo tổ hợp biến và nhóm ngành

Bản đồ nhiệt tại Hình 14 cung cấp cái nhìn thực nghiệm chi tiết về tính không đồng nhất của thị trường, chứng minh rằng năng lực dự báo của các tổ hợp biến số phụ thuộc chặt chẽ vào đặc thù của từng nhóm ngành kinh tế.



Hình 13. Bản đồ nhiệt thể hiện mức độ phân hóa hiệu suất (AUC) theo tổ hợp biến và nhóm ngành

Kết quả phân tích cho thấy sự áp đảo rõ rệt của các tổ hợp lai ghép tại những nhóm ngành có tính chất thâm dụng thông tin cao như Viễn thông và Tiện ích cộng đồng. Cụ thể, ngành Viễn thông xác lập chỉ số AUC cao nhất toàn thị trường với mức 0,640 khi áp dụng tổ hợp Tech+Sent. Hiện tượng này minh chứng cho đặc tính của dòng tiền trong ngành thường có xu hướng phản ứng nhạy bén với các luồng tin tức sự kiện, nơi mà các chỉ báo kỹ thuật thuần túy (với mức AUC chỉ đạt 0,519) không thể phản ánh trọn vẹn các động lực chi phối. Tương tự, tại nhóm ngành Tiện ích cộng đồng, mặc dù tổ hợp Technical đơn lẻ mang lại hiệu suất tương đối ổn định ở mức 0,605, việc tích hợp đa nguồn dữ liệu trong mô hình toàn phần (All) vẫn duy trì được năng lực phân loại khả quan với AUC đạt 0,576 nhờ vào khả năng lọc nhiễu dựa trên sự đồng thuận giữa biến động giá và các tín hiệu tâm lý truyền thông.

Một phát hiện mang tính phản trực giác nhưng sở hữu giá trị khoa học cao được ghi nhận tại ngành Công nghệ thông tin, khi tổ hợp Phân tích cơ bản (Fundamental) lại mang lại hiệu suất tối ưu với AUC đạt 0,620, vượt xa mức 0,507 của tổ hợp tối ưu chung Tech+Sent. Kết quả này phản ánh một thực tế đặc thù tại thị trường Việt Nam đối với các doanh nghiệp công nghệ đầu ngành (như FPT hay CMG); theo đó, giá trị nội tại và các chỉ số tài chính cốt lõi mới là động lực chính yếu dẫn dắt quỹ đạo giá cổ phiếu, thay vì các mẫu hình kỹ thuật ngắn hạn hay biến động tâm lý đám đông. Do đó, việc vận dụng các mô hình học sâu để khai thác các đặc trưng từ báo cáo tài chính định kỳ cho nhóm ngành này mang lại giá trị dự báo thực tiễn sắc bén hơn đáng kể so với các cách tiếp cận chỉ dựa trên dữ liệu giao dịch tần suất cao.

4.7. Thảo luận kết quả

Kết quả thực nghiệm thu được từ mô hình mạng thần kinh nhân tạo đa lớp (MLP) cung cấp những bằng chứng định lượng quan trọng ủng hộ lý thuyết Tài chính hành vi trong cấu trúc vi mô của thị trường chứng khoán Việt Nam. Trong khung thời gian ngắn hạn 5 ngày giao dịch, các biến động giá cổ phiếu được chứng minh là chịu sự chi phối chủ yếu bởi động lượng dòng tiền và trạng thái tâm lý đám đông thay vì các giá trị nội tại doanh nghiệp. Sự vượt trội của tổ hợp lai ghép giữa phân tích kỹ thuật và tâm lý (Tech+Sent) đã củng cố luận điểm rằng: các chỉ báo kỹ thuật đóng vai trò nhận diện trạng thái cung - cầu hiện hành, trong khi tín hiệu tâm lý trích xuất từ mô hình PhoBERT đóng vai trò là bộ lọc nhiễu, xác nhận độ tin cậy và sức mạnh của các nhịp bứt phá về giá.

Trái ngược với kỳ vọng về việc bổ sung thông tin sẽ làm tăng độ chính xác, việc tích hợp dữ liệu phân tích cơ bản (Fundamental) lại dẫn đến sự suy giảm hiệu suất dự báo (AUC). Do đặc tính tĩnh và tần suất công bố theo quý có độ trễ lớn, các chỉ số tài chính cốt lõi như P/E hay P/B phần lớn đã được thị trường chiết khấu vào giá trước thời điểm công bố chính thức. Việc ép buộc mạng nơ-ron học các tham số tĩnh này để dự báo các dao động tần suất cao không những không mang lại giá trị thông tin biên mà còn làm gia tăng độ nhiễu hệ thống, gây pha loãng không gian đặc trưng và làm giảm khả năng nhận diện các tín hiệu động lượng của mô hình.

Đi sâu vào phân tích mức độ đóng góp của các biến số, nghiên cứu làm nổi bật hiệu ứng đảo chiều về giá trị trung bình (Mean Reversion) một quy luật vận hành cốt lõi tại thị trường Việt Nam. Điều này thể hiện qua sức mạnh phân loại áp đảo của biến chênh lệch giá so với xu hướng nền (sma20_gap). Đồng thời, sự dẫn đầu của chỉ báo cú sốc tâm lý (sent_shock) trong bảng xếp hạng tầm quan trọng đặc trưng khẳng định mức độ nhạy cảm cao của nhà đầu tư nội địa với các thông tin bất ngờ từ truyền thông. Ngược lại, các chỉ báo xác nhận xu hướng có độ trễ lớn như sma_cross bị chứng minh là kém hiệu quả khi dải thời gian dự báo bị thu hẹp, thường tạo ra các tín hiệu giả trong các nhịp biến động ngắn.

Cuối cùng, sự phân hóa hiệu suất mô hình theo từng nhóm ngành kinh tế phản ánh rõ nét tính không đồng nhất và cấu trúc phức tạp của thị trường. Mạng nơ-ron hoạt động tối ưu ở các ngành có tính chất phòng thủ như Viễn thông và Tiện ích cộng đồng, nhưng suy giảm hiệu suất rõ rệt ở nhóm có tính chu kỳ cao như Ngân hàng và Công nghiệp. Sự khác biệt này cho thấy các cổ phiếu vốn hóa lớn thường chịu sự chi phối mạnh mẽ bởi các biến số vĩ mô hệ thống chưa được bao phủ hoàn toàn trong bộ dữ liệu hiện tại. Phát hiện này nhấn mạnh yêu cầu tất yếu đối với các nhà thực hành định lượng trong việc cá nhân hóa không gian đặc trưng học máy cho từng hệ sinh thái ngành riêng biệt, thay vì áp dụng một mô hình phổ quát duy nhất cho toàn bộ thị trường.

CHƯƠNG 5. KẾT LUẬN VÀ KIẾN NGHỊ

5.1. Kết luận chung

Nghiên cứu đã hoàn thành mục tiêu thiết lập và đánh giá thực nghiệm mô hình mạng thần kinh nhân tạo đa lớp (MLP) trong bài toán dự báo xu hướng giá cổ phiếu ngắn hạn tại thị trường chứng khoán Việt Nam. Thông qua quy trình thực nghiệm nghiêm ngặt với các phương pháp kiểm định ngoài mẫu, kết quả từ phân tích cắt bỏ (Ablation Study) đã khẳng định rằng sự lai ghép giữa dữ liệu Phân tích kỹ thuật và Tâm lý thị trường (Tech+Sent) cấu thành một không gian đặc trưng tối ưu nhất, xác lập hệ số AUC đạt 0,573. Sự kết hợp giữa động lượng dòng tiền từ các chỉ báo kỹ thuật và tín hiệu xác nhận từ tâm lý truyền thông đã tạo ra một bộ lọc nhiễu hiệu quả, giúp mô hình nhận diện chính xác hơn các điểm đảo chiều và xu hướng tăng trưởng bền vững. Ngược lại, dữ liệu phân tích cơ bản bộc lộ những hạn chế rõ rệt khi không mang lại giá trị gia tăng trong dải thời gian dự báo 5 ngày, chủ yếu do đặc tính tĩnh và độ trễ phản ánh thông tin so với tốc độ chiết khấu của thị trường.

Việc bóc tách cơ chế nội tại của mô hình thông qua phương pháp định lượng tầm quan trọng đặc trưng (Permutation Importance) đã mang lại những phát hiện quan trọng về quy luật vận động của thị trường nội địa. Nghiên cứu phát hiện sức mạnh chi phối tuyệt đối của hiệu ứng đảo chiều về giá trị trung bình (Mean Reversion) thông qua biến số mức chênh lệch giá (sma20_gap), đồng thời xác nhận các cú sốc tâm lý đột ngột từ truyền thông là nhân tố kích hoạt rủi ro biến động giá cực kỳ mạnh mẽ. Bên cạnh đó, các đánh giá đa phân khúc chỉ ra cấu trúc thị trường Việt Nam mang tính phân mảnh sâu sắc. Thuật toán học máy đạt hiệu suất tối ưu tại các nhóm ngành phòng thủ có mô hình kinh doanh ổn định như Viễn thông và Tiện ích cộng đồng, nhưng lại suy giảm năng lực phân lớp đáng kể ở các nhóm ngành chu kỳ có vốn hóa lớn như Ngân hàng và Công nghiệp. Sự phân hóa này phản ánh giới hạn của tập dữ liệu hiện tại trước các tác động của yếu tố vĩ mô hệ thống, đồng thời bác bỏ quan điểm về một mô hình dự báo phổ quát cho toàn bộ thị trường.

5.2. Hạn chế của nghiên cứu

Mặc dù đạt được những kết quả có ý nghĩa thống kê về năng lực dự báo, nghiên cứu vẫn tồn tại những hạn chế nhất định về phương pháp luận và cấu trúc dữ liệu, đòi hỏi tư duy phản biện khi ứng dụng vào thực tiễn đầu tư.

Thứ nhất, hệ thống không gian đặc trưng hiện tại chưa tích hợp các chỉ báo kinh tế vĩ mô cốt lõi như lãi suất liên ngân hàng, tỷ giá hối đoái hay tăng trưởng tín dụng. Sự thiếu khuyết này trực tiếp dẫn đến sự suy giảm năng lực dự báo đối với các nhóm cổ phiếu ngành tài chính, vốn có mức độ nhạy cảm cao với các biến số hệ thống. Bên cạnh đó, nguồn dữ liệu tâm lý trong nghiên cứu hiện mới chỉ được trích xuất từ các kênh báo chí chính thống, vô hình trung bỏ qua một lượng lớn thông tin phi cấu trúc và đa chiều từ mạng xã hội hoặc các diễn đàn đầu tư cá nhân, nơi thường xuyên phản ánh xung lực tâm lý của các nhà đầu tư nhỏ lẻ tại Việt Nam.

Thứ hai, về mặt thiết kế thuật toán, Mạng nơ-ron đa lớp (MLP) với bản chất là kiến trúc truyền thẳng (Feed-forward) tuy sở hữu năng lực xấp xỉ hàm phi tuyến vượt trội, nhưng chưa thực sự tối ưu trong việc khai thác các phụ thuộc chuỗi thời gian dài hạn (Long-term sequential dependencies). Cuối cùng, việc cố định biên độ dự báo ở mức 5 ngày giao dịch đã làm hạn chế khả năng quan sát độ trễ phản ứng của thị trường, đặc biệt là đối với các thay đổi mang tính chu kỳ trong cấu trúc tài chính cơ bản của doanh nghiệp.

5.3. Kiến nghị và hướng nghiên cứu tiếp theo

Về mặt thực tiễn trong lĩnh vực công nghệ tài chính (FinTech), các kết quả định lượng cung cấp cơ sở phương pháp luận vững chắc để thiết kế các hệ thống quản trị danh mục đầu tư tự động hóa (Robo-advisors). Các chiến lược giao dịch ngắn hạn cần được xây dựng dựa trên sự giao thoa giữa thuật toán nhận diện tín hiệu kỹ thuật và bộ lọc rủi ro tâm lý từ mô hình PhoBERT. Đặc biệt, việc phân bổ tỷ trọng tài sản và tinh chỉnh tham số mô hình bắt buộc phải được cá nhân hóa cho từng hệ sinh thái ngành cụ thể, tuyệt đối tránh việc áp dụng một quy tắc định giá chung cho toàn bộ thị trường.

Về định hướng nghiên cứu tiếp theo, các công trình tương lai được khuyến nghị thử nghiệm những kiến trúc học sâu chuyên biệt cho chuỗi thời gian như Mạng nơ-ron hồi quy (LSTM, GRU) hoặc mô hình chú ý (Transformer) nhằm khai thác triệt để cấu trúc tự tương quan của dữ liệu giao dịch. Song song đó, việc mở rộng mô hình sang bài toán dự báo đa khung thời gian (Multi-timeframe prediction), kết hợp cùng dữ liệu vĩ mô và dữ liệu thay thế (Alternative data) như lưu lượng truy cập mạng xã hội sẽ là bước tiến cần thiết để hoàn thiện bức tranh toàn cảnh về cơ chế định giá tài sản tài chính trong kỷ nguyên số.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Nguyễn, & cộng sự. (2022). [Ứng dụng kỹ thuật khai phá văn bản để dự báo biến động chỉ số VN-Index]. *Tạp chí Khoa học, Đại học Quốc gia Hà Nội*.

Tiếng Anh

2. Arauco Ballesteros, M. A., & Martínez Miranda, E. A. (2024). Stock market forecasting using a neural network through fundamental indicators, technical indicators and market sentiment analysis. *Computational Economics*.
3. Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques–Part II: Soft computing methods. *Expert Systems with Applications*, 36(3), pp. 5932–5941.
4. Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49(3), pp.307-343.
5. Damodaran, A. (2012). *Investment valuation: Tools and techniques for determining the value of any asset* (3rd ed.). John Wiley & Sons.
6. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), pp. 383-417.
7. Halder, S. (2022). FinBERT-LSTM: Deep Learning based stock price prediction using News Sentiment Analysis. *arXiv preprint, arXiv:2211.07392*.
8. Haykin, S. (2009). Neural networks and learning machines (3rd ed.). Pearson Education.
9. Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), pp. 359-366.
10. Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32(10), pp. 2513-2522.
11. Murphy, J. J. (1999). *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. New York Institute of Finance.

12. Nguyen, D. Q., & Nguyen, A. T. (2020). PhoBERT: Pre-trained language models for Vietnamese. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1037-1042.
13. Nguyen-Trang, T., Do-Thi, N., & Nhon, H. T. (2024). Predicting stock returns using machine learning combined with data envelopment analysis and automatic feature engineering: A case study on the Vietnamese stock market. *PLOS One*, 20(9), e0332154.
14. Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, 17(1), pp. 83-104.
15. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), pp. 1139-1168.
16. Vu, L. T., Pham, D. N., Kieu, H. T., & Pham, T. T. T. (2023). Sentiments extracted from news and stock market reactions in Vietnam. *International Journal of Financial Studies*, 11(3), pp. 101.
17. Wanjawa, B. W., & Muchemi, L. (2014). ANN model to predict stock prices at stock exchange markets. *arXiv preprint arXiv:1502.06434.*