

CUỘC THI DATA EXPLORER 2025



BÁO CÁO KỸ THUẬT

Chủ đề: Tiếp cận đa mô hình - Dự báo giá cổ phiếu ngắn hạn, dài hạn và khai thác thông tin với AI

Đội thi: BareFooths

Nguyễn Đức Dũng

Đại học Bách Khoa Hà Nội

Nguyễn Hoàng Quân

Đại học Bách Khoa Hà Nội

Tô Lê Quang

Đại học Bách Khoa Hà Nội

Ngô Thị Thùy Duyên

Đại học Bách Khoa Hà Nội

Tống Diệu Linh

Đại học Kinh tế Quốc dân

Hà Nội, 2025

MỤC LỤC

Chương 1: Giới Thiệu Chung	4
1.1 Tính cấp thiết của vấn đề	4
1.2 Mục tiêu của báo cáo	4
1.3 Cấu trúc báo cáo kĩ thuật	5
Chương 2: Tổng Quan Về Cổ Phiếu Ngành Công Nghệ Và 2 Công Ty	5
2.1 Tổng quan về Ngành	5
2.1.1. Phân cấp ngành	5
2.1.2. Phân tích xu hướng dòng tiền và độ ổn định của các nhóm ngành	6
2.1.3. Sự tương quan giữa các ngành	9
2.2 Tổng quan về công ty FPT	10
2.2.1. Sức khỏe tài chính của công ty	10
2.3 Tổng quan về công ty CMC	16
2.3.1. Sức khỏe tài chính của công ty	16
Chương 3: Xây Dựng Mô Hình Dự Đoán Giá Cổ Phiếu	22
3.1 Phân tích dữ liệu	22
3.1.1. Thu thập dữ liệu	22
3.1.2. Mô tả dữ liệu	22
3.1.3. Tiền xử lý	23
3.1.4. Trực quan hóa dữ liệu	23
3.1.5. Giả thuyết và kiểm định	31
3.2 Mô hình phân loại xu hướng cổ phiếu	31
3.3 Mô hình dự đoán giá cổ phiếu	33

3.3.1. Giải pháp sơ khai	33
3.3.2. Mô hình phức hợp	38
Chương 4: Triển Khai RAG	50
4.1 Thu thập dữ liệu	50
4.2 Phương pháp thực hiện	51
4.2.1. Tiền xử lý	51
4.2.2. Embedding	53
4.2.3. Vector search	54
4.2.4. Quy trình tạo đầu ra của hệ thống RAG	56
4.3 Kết quả phân tích và định hướng	57
4.3.1. Kết quả	57
4.3.2. Huấn luyện tinh chỉnh mô hình Qwen:	57
4.3.3. Định hướng phát triển	60
Chương 5: Huấn Luyện Tinh Chỉnh Mô Hình Ngôn Ngữ Lớn	60
5.1 Thu thập dữ liệu	60
5.2 Gán nhãn cảm xúc của dữ liệu	61
5.3 Xử lý dữ liệu	63
5.4 Kiểm thử mô hình	64
5.5 Tinh chỉnh mô hình ngôn ngữ lớn	65
5.5.1. Phi-3 3,8B, Llama guard 8B, Llama 3 8B Instruct, Llama 3 13B hf	65
5.5.2. Gemma 2 2,1B phi	66
5.6 Đánh giá mô hình	66
5.6.1. Dánh giá sai số dự đoán	66
5.6.2. Dánh giá trên các khía cạnh khác	67

5.7	Kết luận	68
Chương 6: Hồi Kết		68
6.1	Đánh giá	68
6.1.1.	Xử lý dữ liệu	69
6.1.2.	Sử dụng học máy, học sâu dự đoán giá đóng cửa ngắn, trung và dài hạn	69
6.1.3.	RAG truy vết thông tin cổ phiếu và tài chính	69
6.1.4.	Tinh chỉnh mô hình ngôn ngữ lớn	69
6.2	Hướng phát triển trong tương lai	70
6.2.1.	Nâng cấp mô hình dự báo	70
6.2.2.	Cải tiến RAG	70
6.2.3.	Tối ưu LLM	70
6.2.4.	Hệ thống đầu tư tự động thông minh	70
6.2.5.	Dánh giá khách quan hơn	70

Chương 1: Giới Thiệu Chung

1.1 Tính cấp thiết của vấn đề

Thị trường chứng khoán Việt Nam là kênh dẫn vốn quan trọng, phản ánh sức khỏe nội tại doanh nghiệp và kỳ vọng tăng trưởng chung của nền kinh tế. Trong đó giá trị cổ phiếu là yếu tố cốt lõi trong quyết định đầu tư, chịu ảnh hưởng tổng hợp từ nội lực doanh nghiệp, biến động của nền kinh tế vĩ mô, tâm lý nhà đầu tư. Đặc biệt, trong bối cảnh nền kinh tế thế giới nói chung, Việt Nam nói riêng đang trải qua nhiều thay đổi sâu sắc về chính sách và cấu trúc, tính bất định của thị trường tài chính ngày càng gia tăng. Điều này đặt ra yêu cầu cấp thiết cho việc ứng dụng các phương pháp phân tích dữ liệu hiện đại, học máy và trí tuệ nhân tạo nhằm dự báo chính xác hơn xu hướng giá cổ phiếu, từ đó hỗ trợ nhà đầu tư và doanh nghiệp đưa ra các quyết định chiến lược phù hợp với bối cảnh mới.

1.2 Mục tiêu của báo cáo

- Báo cáo hướng đến giải quyết 5 mục tiêu lớn:
 - + Phân tích thị trường về ngành công nghệ và hai công ty FPT, CMG.
 - + Xây dựng, huấn luyện và đánh giá mô hình học máy nhằm dự báo xu hướng biến động giá cổ phiếu và dự báo giá đóng cửa trong nhiều ngày tới.
 - + Ứng dụng RAG để trích xuất và tổng hợp thông tin định tính và định lượng hỗ trợ tra cứu thông tin.
 - + Tinh chỉnh mô hình ngôn ngữ lớn (LLM) để kết hợp các yếu tố định tính và định lượng dự đoán giá đóng cửa cổ phiếu trong nhiều ngày tới.
 - + Đề xuất một quy trình khai thác dữ liệu và phát triển mô hình AI có khả năng áp dụng thực tiễn trong lĩnh vực tài chính - chứng khoán.
 - Thông qua đó, báo cáo hướng tới việc nâng cao khả năng dự báo thị trường, góp phần hỗ trợ các nhà đầu tư và tổ chức tài chính ra quyết định đầu tư hiệu quả trong bối cảnh thị trường chứng khoán nhiều biến động.

1.3 Cấu trúc báo cáo kĩ thuật

Chương 2: Tổng Quan Về Cổ Phiếu Ngành Công Nghệ Và 2 Công Ty

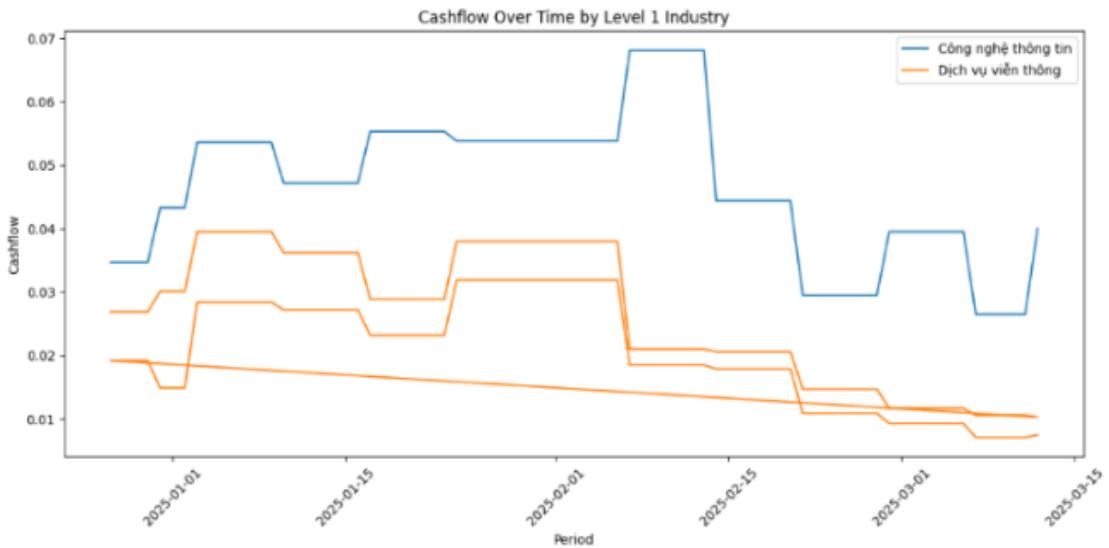
2.1 Tổng quan về Ngành

2.1.1. Phân cấp ngành

FPT và CMC là hai tập đoàn đa ngành trong lĩnh vực ICT. Dựa trên phân tích dữ liệu của hai công ty, có thể đưa ra nhận xét về các ngành mà 2 công ty tham gia:

- + Công nghệ thông tin
- + Dịch vụ viễn thông
- + Dịch vụ viễn thông
- + Phân phối và bán lẻ hàng lâu bền
- + Phần mềm và dịch vụ
- + Dịch vụ công nghệ thông
- + Dịch vụ viễn thông không dây
- + Dịch vụ viễn thông đa dạng
- + Phần mềm

Trong đó, phân cấp ngành chính mà 2 công ty tham gia dựa trên mức độ tổng quát và chuyên sâu của hoạt động sản xuất, cung cấp dịch vụ như sau:



Hình 1: Dòng tiền hai ngành level 1

Cấp ngành	Phạm vi	Cụ thể
Level 1 (ngành tổng quát)	Nhóm ngành theo lĩnh vực kinh tế lớn	+ Công nghệ thông tin + Dịch vụ viễn thông
Level 2 (ngành trung gian)	Phân nhóm theo sản phẩm/dịch vụ chính trong lĩnh vực kinh tế lớn.	+ Dịch vụ viễn thông + Phân phối và bán lẻ hàng lâu bền + Phần mềm và dịch vụ.
Level 3 (ngành cụ thể)	Chi tiết các lĩnh vực cụ thể hai công ty tham gia cung cấp sản phẩm/dịch vụ	+ Dịch vụ công nghệ thông tin + Dịch vụ viễn thông không dây + Dịch vụ viễn thông đa dạng + Phần mềm.

Bảng 1: Phân cấp ngành mà 2 công ty tham gia

2.1.2. Phân tích xu hướng dòng tiền và độ ổn định của các nhóm ngành

- Sau khi phân tích dữ liệu dòng tiền của 2 công ty trong thời gian từ ngày 01/01/2025 - 15/03/2025, ta có thể nhận xét như sau:
- Đối với các ngành level 1:
 - + Ngành công nghệ thông tin:

		Industry	Stability_STD
19	Phân phối và bán lẻ hàng lâu bền		1.031886
7	Công nghệ thông tin		4.850527
73	Phần mềm		4.949522
27	Phần mềm và dịch vụ		5.027499
72	Dịch vụ công nghệ thông tin		6.314680
8	Dịch vụ viễn thông		19.429811
78	Dịch vụ viễn thông không dây		34.919304
30	Dịch vụ viễn thông		36.417562
77	Dịch vụ viễn thông đa dạng		46.707796

Hình 2: Biến động giá của 10 ngành ảnh hưởng tới hai công ty

- Xu hướng: Dòng tiền tăng đều từ đầu tháng 1 đến giữa tháng 2 (0.035-0.07) sau đó giảm dần từ giữa tháng 2 đến giữa tháng 3. Trong đó, tăng cao nhất vào thời gian từ 01/02/2025-15/02/2025.
- Mức độ biến động của ngành thấp (stabilit_STD = 4.850527)

→ Ngành đem lại dòng tiền cao, mức độ biến động thấp, ổn định.

+ Ngành dịch vụ viễn thông:

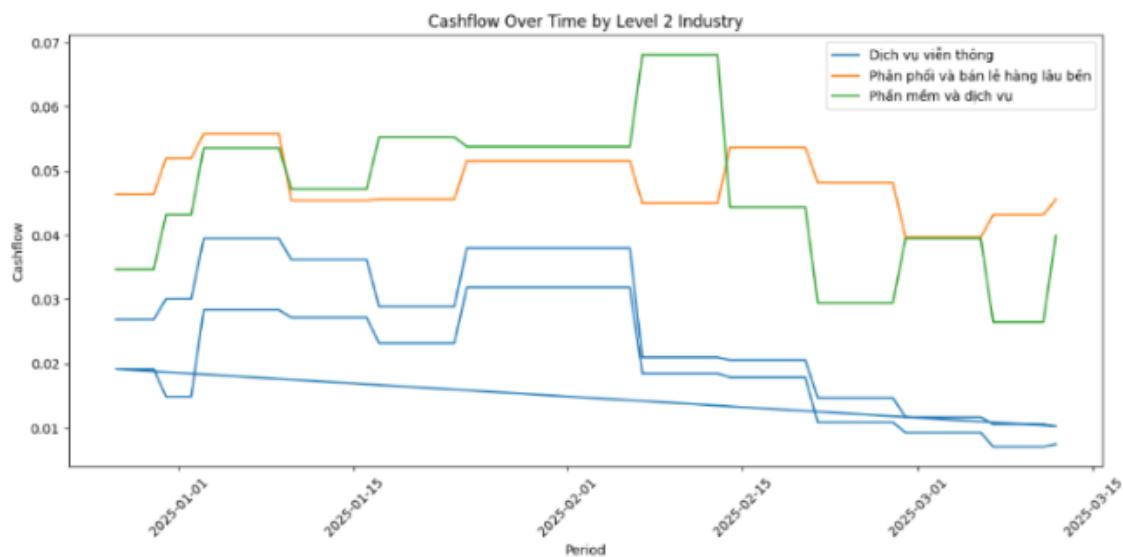
- Xu hướng: Dòng tiền dao động trong (0.01-0.04), tăng nhẹ trong giai đoạn tháng 1, giảm mạnh từ đầu tháng 2 đến 15/03/2025.
- Mức độ biến động của ngành cao (stability_STD = 19.429811).

→ Ngành đem lại dòng tiền trung bình, mức độ biến động cao, rủi ro cao.

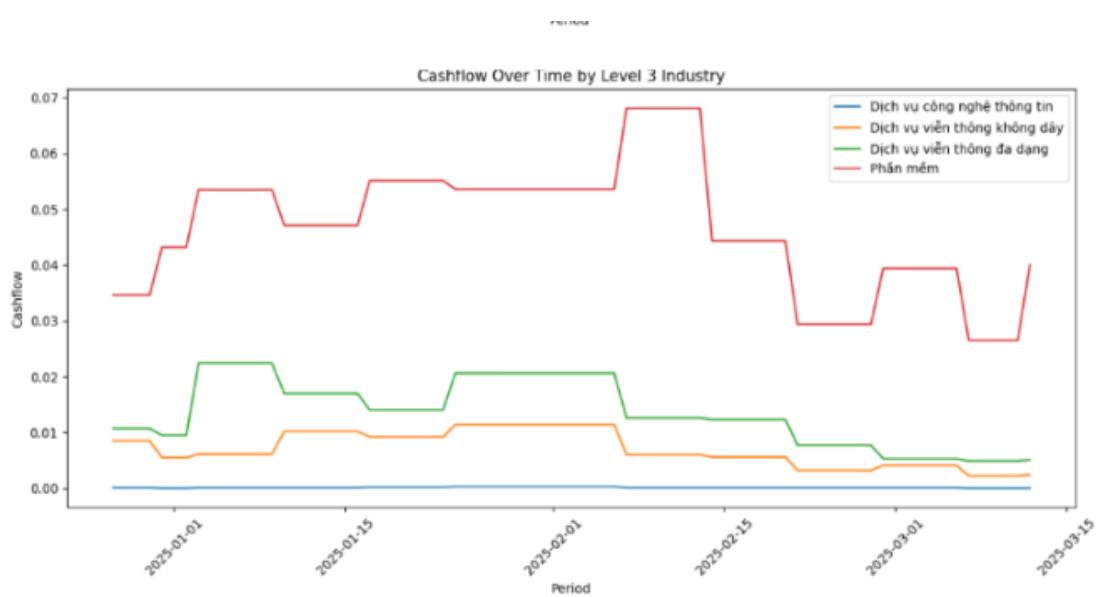
- Đối với ngành level 2 và 3:

- Dịch vụ viễn thông: Dòng tiền ở mức ổn định, giao động trong khoảng (0.03-0.04) trong tháng, giảm dần từ đầu tháng 1 đến giữa tháng 3 (0.01). Trong đó 2 ngành đem lại doanh thu chính là:

+ Dịch vụ viễn thông không dây: độ biến động (stability_STD = 34.919304) → Dòng tiền biến động mạnh, không ổn định.



Hình 3: Dòng tiền các ngành level 2



Hình 4: Dòng tiền các ngành level 3

- + Dịch vụ viễn thông đa dạng: độ biến động (stability_STD = 46.707796) → Dòng tiền biến động mạnh, rủi ro lớn, tuy nhiên có khả năng mở rộng.
- Phân phối và bán lẻ hàng lâu bền: Dòng tiền ổn định, giao động trong khoảng (0.04-0.055), có thể thấy dòng tiền của phân phối và bán lẻ hàng lâu bền lớn, duy trì ổn định với stability_STD = 1.031886 → Có thể thấy đây là nhóm ngành ổn định, ít rủi ro và có khả năng tạo doanh thu ổn định cho công ty.
- Phần mềm và dịch vụ: Dòng tiền dao động trong khoảng (0.03-0.065), trong đó tăng cao nhất vào nửa đầu tháng 2 và giảm mạnh nhất ở nửa đầu tháng 3. Xét về mức độ ổn định, giá trị stability_STD = 5.027499 → ngành ổn định, ít biến động. → Như vậy, nhóm ngành hai công ty có dòng tiền nhiều, ổn định từ hai ngành phân phối và bán lẻ hàng lâu bền, phần mềm và dịch vụ, dòng tiền đến từ ngành dịch vụ viễn thông không ổn định.

2.1.3. Sự tương quan giữa các ngành

Dựa vào tính toán độ tương quan và chỉ số thống kê, chúng tôi rút ra được những ngành có mức độ liên quan tới các ngành mà hai công ty tham gia (corr value >0.3 và p-value<0.5 là):

- + Dịch vụ thương mại và vật tư
- + Vận tải hàng không và logistics
- + Dịch vụ viễn thông đa dạng
- + Dịch vụ viễn thông
- + Dịch vụ viễn thông không dây
- + Dịch vụ chuyên biệt và thương mại
- + Cơ sở hạ tầng giao thông
- + Vận tải
- + Chất bán dẫn và thiết bị bán dẫn
- + Dịch vụ công nghệ thông tin
- + Phần mềm và dịch vụ
- + Phần mềm
- + Công nghệ thông tin
- + Đồ dùng cá nhân

+ Dịch vụ chuyên biệt

2.2 Tổng quan về công ty FPT

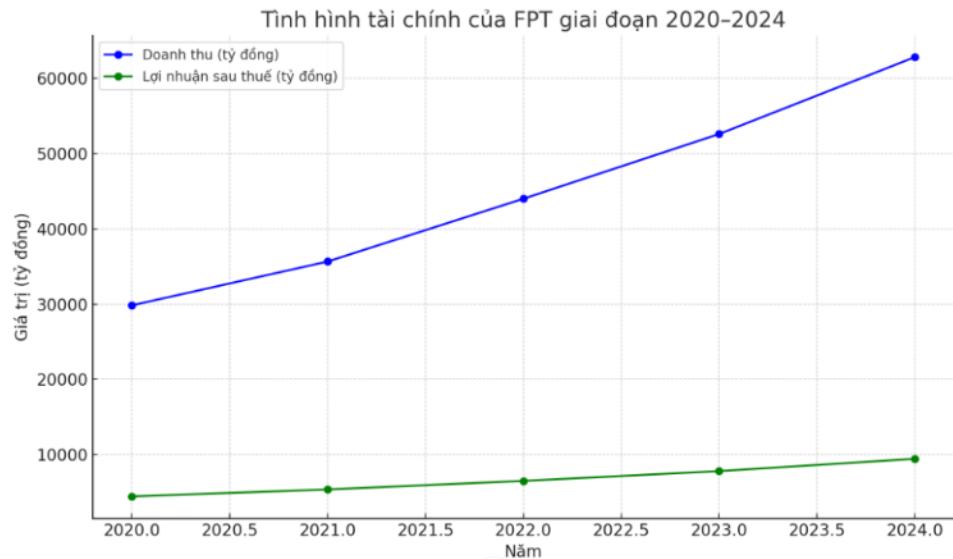
Công ty Cổ phần FPT (FPT) có tiền thân là Công ty Công nghệ Thực phẩm được thành lập năm 1988. Công ty hoạt động chính trong lĩnh vực phần mềm, công nghệ thông tin, tích hợp hệ thống, viễn thông, và giáo dục đào tạo. FPT chính thức hoạt động theo mô hình công ty cổ phần từ năm 2002. FPT sở hữu hơn 250 giải pháp phần mềm được cấp bản quyền trong các lĩnh vực chuyên biệt, mạng lưới hạ tầng internet phủ rộng khắp 63 tỉnh thành của cả nước. FPT tham gia vào lĩnh vực giáo dục đào tạo thông qua Trường Đại học FPT, Trường Đại học Trực tuyến FUNiX - trường đại học trực tuyến đầu tiên của Việt Nam. Bên cạnh đó, Công ty đầu tư vào các công ty liên kết trong lĩnh vực phân phối bán lẻ các sản phẩm công nghệ, chứng khoán và quản lý quỹ đầu tư. Trên lĩnh vực viễn thông, FPT nằm trong TOP 3 nhà cung cấp dịch vụ internet hàng đầu Việt Nam. FPT được niêm yết trên Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh (HOSE) từ tháng 12/2006. Cổ phiếu của công ty FPT (mã FPT): Vốn hóa thị trường khoảng 5,5–6 tỷ USD → thuộc nhóm vốn hóa lớn trên sàn HOSE, có tác động đáng kể tới chỉ số VN-Index.

2.2.1. Sức khỏe tài chính của công ty

Doanh thu và Lợi nhuận sau thuế

Bảng 2: Bảng doanh thu và lợi nhuận qua các năm (Đơn vị: tỷ đồng)

	2020	2021	2022	2023	2024
Doanh thu	29,830.4	35,657.26 (+19.5%)	44,009.53 (+23.4%)	52,617.9 (+19.6%)	62,848.79 (+19.4%)
Lợi nhuận sau thuế	4,423.75	5,349.3	6,491.34	7,788.05	9,427.42
Biên lợi nhuận sau thuế	14.8%	15.0%	14.7%	14.8%	15.0%



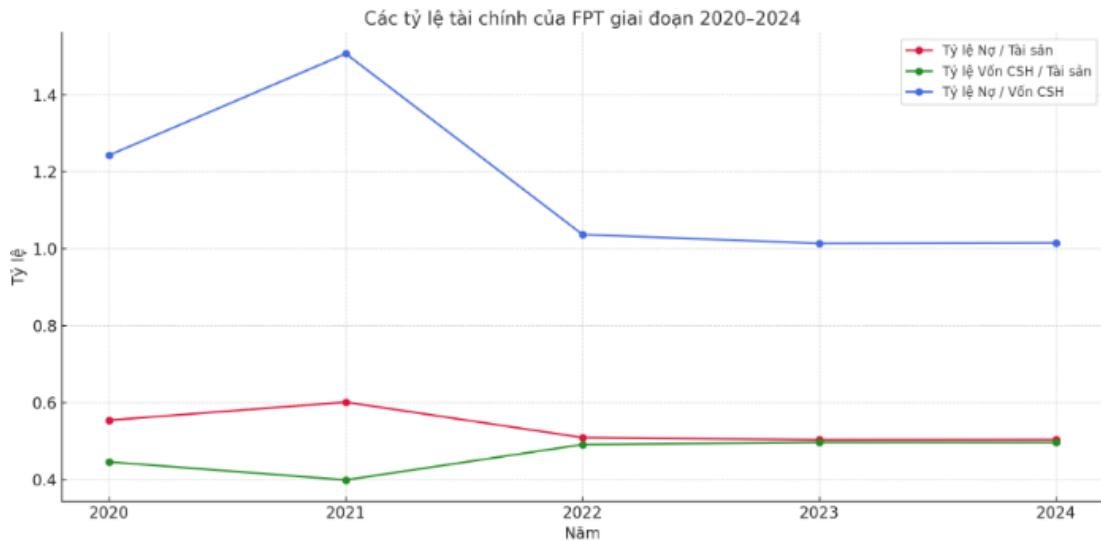
Nhận xét: Tốc độ tăng trưởng trung bình hàng năm (CAGR) khoảng 20%, cho thấy công ty đang mở rộng quy mô kinh doanh bền vững. FPT đang có sức khỏe tài chính tốt, với tốc độ tăng trưởng cả về doanh thu và lợi nhuận sau thuế ổn định, biên lợi nhuận cao. Điều này phản ánh hiệu quả hoạt động và khả năng thích nghi tốt trong bối cảnh thị trường có thể nhiều biến động.

Các chỉ tiêu phản ánh khả năng huy động vốn

Bảng 3: Bảng cân đối kế toán qua các năm (Đơn vị: nghìn tỷ đồng)

	2020	2021	2022	2023	2024
Tổng tài sản	41,734.32	53,697.94	51,650.40	60,282.83	71,999.99
Nợ phải trả	23,128.66	32,279.96	26,294.28	30,349.82	36,272.46
Vốn chủ sở hữu	18,605.67	21,417.99	25,356.12	29,933.01	35,727.54
Tỷ lệ nợ	55.4%	60.1%	50.9%	50.3%	50.4%

* Số liệu được làm tròn đến 2 chữ số thập phân



Nhận xét:

- Tổng tài sản tăng đều qua các năm, từ: 41.734 nghìn tỷ (2020) → 71.999 nghìn tỷ (2024).
 - Mức tăng trưởng bình quân khoảng 14,6%/năm, thể hiện năng lực mở rộng quy mô đầu tư và hoạt động kinh doanh của FPT.
- Tỷ lệ nợ trên tổng tài sản giảm dần sau 2021, duy trì quanh mức 50–55%, cho thấy cơ cấu vốn khá lành mạnh, không quá phụ thuộc vào nợ vay.
- Vốn chủ sở hữu tăng từ 18.606 nghìn tỷ (2020) → 35.728 nghìn tỷ (2024), gấp gần 2 lần trong 5 năm.
- Tốc độ tăng trưởng vốn chủ sở hữu 17%/năm, tăng nhanh hơn nợ phải trả, cho thấy FPT ưu tiên giữ lại lợi nhuận để tái đầu tư thay vì vay nợ.

Dánh giá khả năng huy động vốn

- Tích cực:
 - + Hệ số nợ/vốn chủ sở hữu (D/E) giảm dần (năm 2024 là 1,01, tức là mỗi 1 đồng vốn chủ sở hữu chỉ đi kèm khoảng 1 đồng nợ) → cho thấy dư địa huy động nợ còn nhiều, công ty đã biết tận dụng đòn bẩy tài chính.
 - + Năng lực sinh lời ổn định (đã phân tích ở bảng trước) giúp FPT dễ dàng tiếp cận vốn từ ngân hàng, trái phiếu hoặc đối tác đầu tư.
- Huy động qua cổ phần hóa/niêm yết thêm:
 - + Tăng trưởng vốn chủ sở hữu mạnh có thể đến từ phát hành thêm cổ phiếu – cho

thấy FPT có niềm tin từ nhà đầu tư trên thị trường vốn.

Định giá

Bảng 4: Bảng chỉ số định giá cổ phiếu

Chỉ số	2020	2021	2022	2023	2024
EPS (VND)	4,512.84	4,779.25	4,840.46	5,090.83	5,340.86
P/E	19.52	16.74	19.63	30.13	21.08
P/B	4.39	4.05	4.95	7.80	5.56
P/Cash Flow	119.57	114.67	110.30	108.21	119.57

Nhận xét:

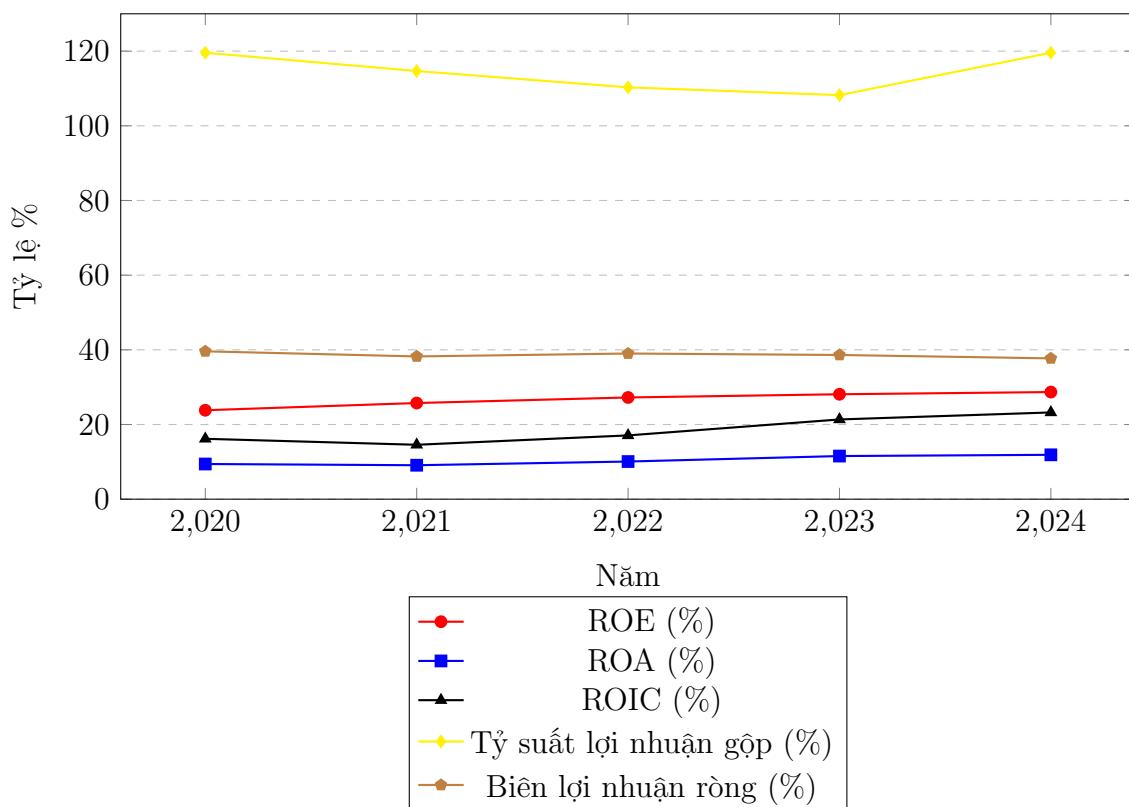
Chỉ số	Mức độ	Nhận xét
EPS (lợi nhuận/mỗi cổ phiếu)	Tăng 18.3% trong 5 năm, tốc độ tăng ổn định	Riêng năm 2023 và 2024 tăng mạnh hơn, cho thấy hiệu quả kinh doanh được cải thiện rõ rệt. Đây là cơ sở nội tại tốt để định giá cổ phiếu cao hơn trong tương lai.
P/E	Năm 2024, P/E = 21	Năm 2023 hơi cao, nhưng năm 2024 hợp lý với mặt bằng chung ngành công nghệ
P/B	4.39 (2020) → 7.80 (2023) rồi giảm xuống 5.56 (2024)	P/B > 5 cho thấy nhà đầu tư sẵn sàng trả giá cao hơn nhiều so với giá trị tài sản ròng → phản ánh kỳ vọng tăng trưởng hoặc thương hiệu mạnh.
P/Cash Flow	110-120	Dòng tiền từ hoạt động kinh doanh chưa tăng tương ứng với lợi nhuận. Hoặc thị trường kỳ vọng lớn vào tương lai tăng trưởng của doanh nghiệp.

Khả năng sinh lời

Bảng 5: Các chỉ số tài chính

Chỉ số	2020	2021	2022	2023	2024
ROE	23.82%	25.75%	27.24%	28.10%	28.69%
ROA	9.42%	9.09%	10.08%	11.55%	11.88%
ROIC	16.17%	14.57%	17.08%	21.35%	23.23%
Tỷ suất lợi nhuận gộp	119.57%	114.67%	110.30%	108.21%	119.57%
Biên lợi nhuận ròng	39.60%	38.23%	39.01%	38.62%	37.71%

Biểu đồ hiệu quả hoạt động (2020–2024)



Tiêu chí	Nhận xét
ROE > 25%	Tạo giá trị cao cho cổ đông
ROIC tăng liên tục	Sử dụng vốn hiệu quả
ROA > 10%	Quản lý tài sản tốt
Biên lợi nhuận ròng cao & ổn định	Mô hình kinh doanh hiệu quả
Tỷ suất lợi nhuận gộp giữ vững	Khả năng kiểm soát chi phí tốt

→ FPT là một cổ phiếu tăng trưởng với nền tảng sinh lời vững chắc, phù hợp cho cả nhà đầu tư dài hạn và các quỹ đầu tư giá trị.

Khả năng thanh toán

Bảng 6: Các chỉ số thanh toán

Chỉ số	2020	2021	2022	2023	2024
Thanh toán nhanh	0.49	0.41	0.61	0.61	0.49
Thanh toán hiện hành	0.79	0.75	0.84	0.94	0.95
Chỉ số thanh toán tiền mặt	0.21	0.18	0.26	0.28	0.27

Nhận xét:

1. Chỉ số thanh toán nhanh

- Mức dưới 1 cho thấy FPT chưa đủ tài sản lưu động trừ hàng tồn kho để thanh toán nợ ngắn hạn ngay lập tức.
- Năm 2022–2023 có cải thiện rõ rệt (0.61) → tăng tính thanh khoản, nhưng đến 2024 lại giảm về 0.49 → cảnh báo có rủi ro thanh khoản nếu áp lực trả nợ tăng cao. → DN cần kiểm soát nợ, tránh vay nhiều quá

2. Khả năng thanh toán hiện hành

- Vẫn dưới ngưỡng an toàn (thường là 1.0–1.5), nhưng xu hướng tăng dần và ổn định.
- Từ 2022–2024: đạt gần 1 → cho thấy FPT đang cải thiện khả năng đáp ứng nghĩa vụ ngắn hạn bằng tài sản lưu động.

3. Chỉ số thanh toán tiền mặt

- Thấp hơn mức lý tưởng (0.5) → FPT không giữ quá nhiều tiền mặt → có thể đang ưu tiên đầu tư mở rộng, không để dòng tiền "chết".
- Tuy nhiên, chỉ số tăng đều từ 2021 → 2023 và duy trì ở mức 0.27 năm 2024 cho thấy dòng tiền khả dụng cải thiện nhẹ.

Dánh giá:

- FPT có tiến bộ rõ rệt về khả năng thanh toán, đặc biệt là trong 3 năm gần đây.
- Tuy vẫn chưa đạt mức an toàn tối ưu, nhưng xu hướng ổn định và cải thiện dần, cho thấy công ty quản lý vốn lưu động tương đối tốt.
- Rủi ro thanh khoản chưa quá lớn, nhưng cần chú ý đến biến động của chỉ số thanh toán nhanh nếu áp lực nợ ngắn hạn tăng trong tương lai.

2.3 Tổng quan về công ty CMC

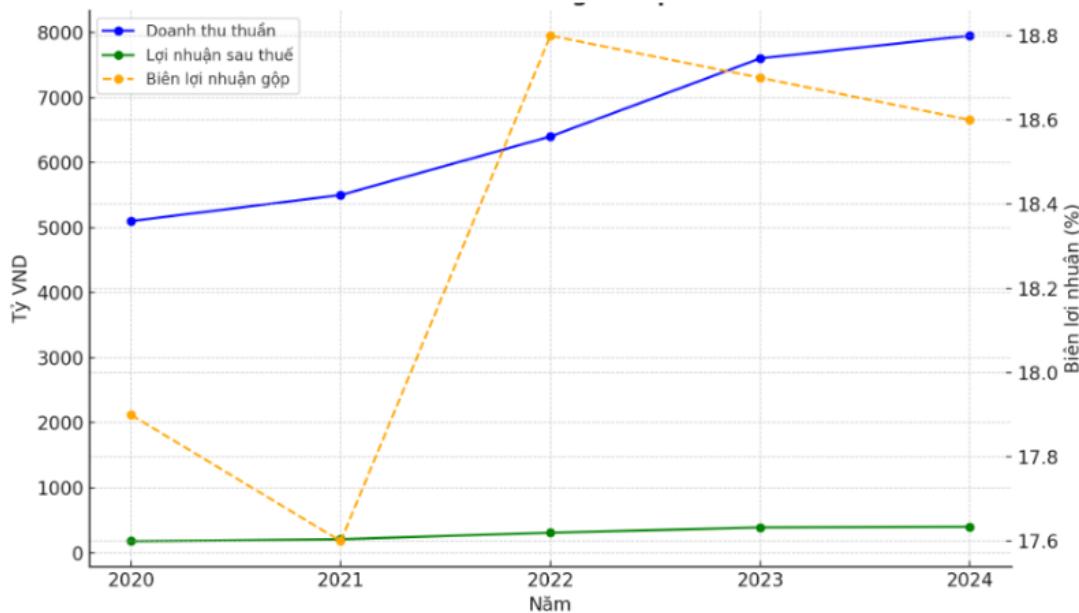
Tập đoàn Công nghệ CMC (mã CMG) là một trong những doanh nghiệp công nghệ thông tin lớn tại Việt Nam, hoạt động trong các lĩnh vực tích hợp hệ thống, phần mềm, dịch vụ đám mây, an ninh mạng và viễn thông. Với hơn 30 năm phát triển, CMC đang hướng tới mục tiêu trở thành tập đoàn công nghệ số toàn cầu, đồng hành cùng quá trình chuyển đổi số của chính phủ và doanh nghiệp. CMC sở hữu đội ngũ kỹ sư công nghệ lớn mạnh, hệ sinh thái dịch vụ đa dạng và định hướng tăng trưởng bền vững. CMC được đánh giá là có nền tảng tài chính ổn định, doanh thu và lợi nhuận tăng trưởng đều, quản lý chi phí hiệu quả, và thị trường vẫn đánh giá tốt về triển vọng. Mặc dù cổ phiếu đang điều chỉnh, nhưng về dài hạn, CMG vẫn là một doanh nghiệp khỏe mạnh và có tiềm năng tăng trưởng tốt trong ngành công nghệ.

2.3.1. Sức khỏe tài chính của công ty

Doanh thu và lợi nhuận sau thuế

Bảng 7: Báo cáo doanh thu và lợi nhuận qua các năm (Đơn vị: tỷ VND)

Chỉ tiêu	2020	2021	2022	2023	2024
Doanh thu thuần	5,100	5,400	6,200	7,950	7,957
Lợi nhuận sau thuế	180	200	300	390	400
Biên lợi nhuận gộp (%)	18.0	17.6	18.8	18.6	18.5



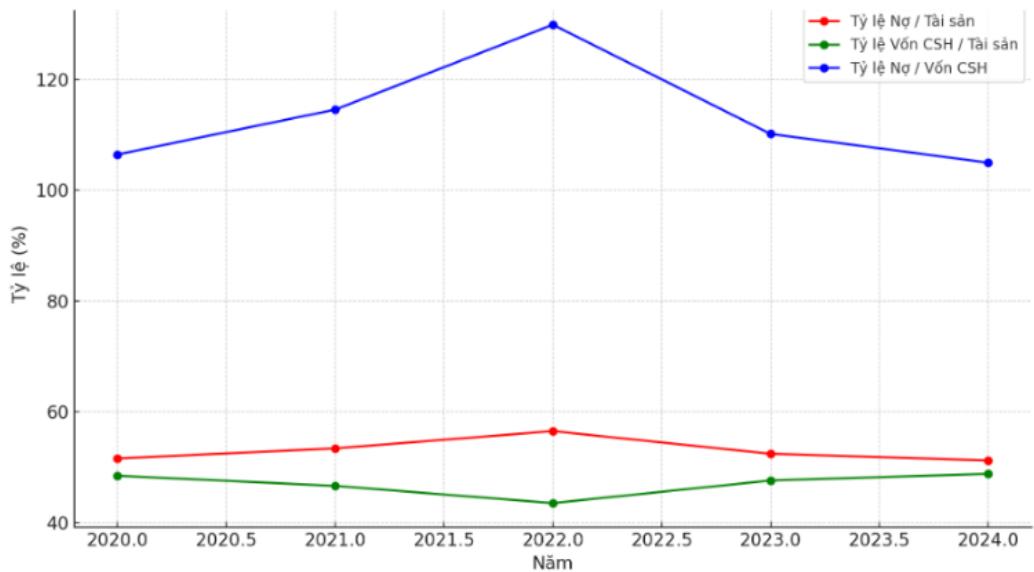
Nhận xét:

- Tăng trưởng ổn định và bền vững:
 - + Doanh thu thuần và lợi nhuận sau thuế đều tăng trưởng liên tục qua các năm, cho thấy công ty có chiến lược mở rộng thị trường và gia tăng hiệu quả hoạt động tốt.
 - + Tốc độ tăng trưởng không đột biến nhưng vững chắc, phản ánh nội lực mạnh.
 - + Tuy nhiên, biên lợi nhuận đang có dấu hiệu giảm nhẹ, cần theo dõi thêm để đánh giá xu hướng dài hạn.

Các chỉ tiêu phản ánh khả năng huy động vốn

Bảng 8: Cơ cấu tài sản và nợ phải trả

	2020	2021	2022	2023	2024
Tổng tài sản	4,649,385	4,983,477	6,255,925	6,561,871	6,853,773
Nợ phải trả	2,396,983	2,660,453	3,534,164	3,439,129	3,509,625
Vốn chủ sở hữu	2,252,403	2,323,023	2,721,761	3,122,743	3,344,148
Tỷ lệ nợ (%)	51.6	53.4	56.5	52.4	51.2



Nhận xét:

- Tổng tài sản tăng liên tục từ 4.650 tỷ VND (2020) lên 6.854 tỷ VND (2024) → cho thấy CMC đang mở rộng quy mô hoạt động.
- Vốn chủ sở hữu cũng tăng đều, từ 2.252 tỷ VND lên 3.344 tỷ VND, thể hiện khả năng giữ lại lợi nhuận tốt và/hoặc gọi vốn hiệu quả.
- Nợ phải trả tăng tương ứng nhưng không vượt trội → tỷ lệ nợ vẫn duy trì trong vùng an toàn (51–56%).
- Tỷ lệ nợ/tổng tài sản ổn định ở mức trung bình khoảng 52–53%, phản ánh cơ cấu vốn cân bằng, không phụ thuộc quá nhiều vào vay nợ.

→ Khả năng huy động vốn:

- CMC có nền tảng tài chính ổn định, tạo điều kiện thuận lợi để huy động vốn qua ngân hàng hoặc phát hành trái phiếu nếu cần thiết.
- Dư địa vay vốn còn nhiều do tỷ lệ nợ chưa cao → tăng khả năng tiếp cận các nguồn vốn bên ngoài.
- Vốn chủ sở hữu tăng đều → khả năng tự tài trợ đầu tư tốt, giúp giảm áp lực huy động vốn khẩn cấp.
- Với uy tín và kết quả kinh doanh tích cực, CMC có khả năng huy động vốn ở mức khá, phục vụ các mục tiêu tăng trưởng và chuyển đổi số dài hạn.

Định giá

Bảng 9: Các chỉ số tài chính

	2020	2021	2022	2023	2024
EPS (VND)	1,700	1,370	1,080	1,060	1,480
P/E Ratio	18.10	23.52	33.90	29.95	20.80
P/B Ratio	1.00	1.55	2.52	2.45	2.29
P/Cash Flow Ratio	8.38	10.35	10.96	11.33	13.88

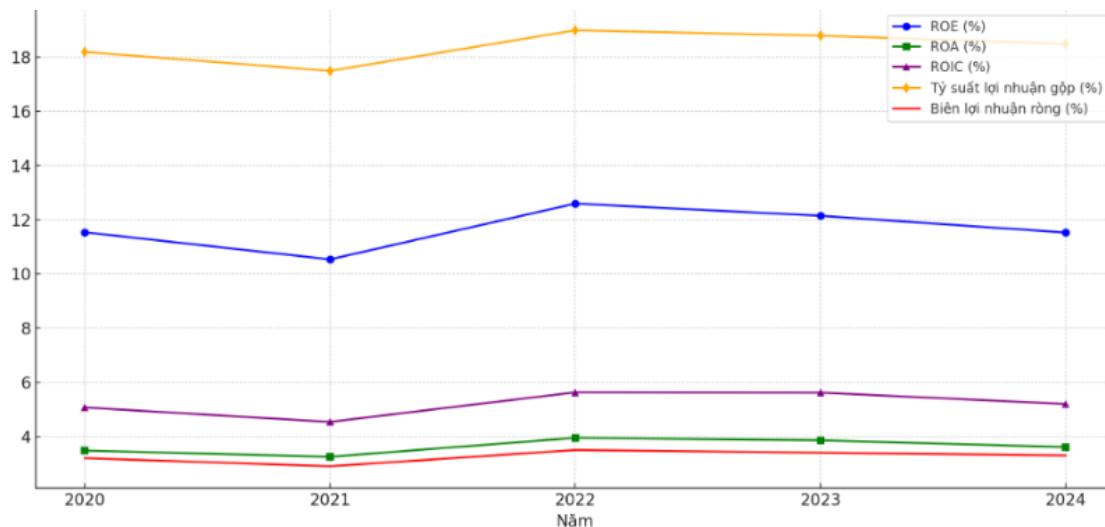
**Nhận xét:**

Chỉ số	Phân tích	Nhận xét
EPS (VND)	Giảm từ 1,700 (2020) xuống 1,060 (2023), tăng lên 1,480 (2024)	Giai đoạn 2020-2023 gặp khó khăn về lợi nhuận, nhưng sự phục hồi năm 2024 cho thấy hoạt động kinh doanh đang cải thiện.
P/E	Tăng mạnh đến 33.9 (2022), giảm dần còn 20.8 (2024)	Định giá cổ phiếu có lúc quá cao so với lợi nhuận thực tế. Việc điều chỉnh P/E về mức hợp lý giúp tăng sức hấp dẫn đầu tư năm 2024.
P/B	Tăng từ 1.00 (2020) lên 2.52 (2022), giảm nhẹ xuống 2.29 (2024)	P/B cao phản ánh kỳ vọng tích cực từ thị trường, cho thấy cổ phiếu được định giá cao hơn so với giá trị sổ sách.
P/Cash Flow	Tăng liên tục từ 8.38 (2020) lên 13.88 (2024)	Nhà đầu tư sẵn sàng trả giá cao hơn cho dòng tiền, cho thấy niềm tin vào khả năng tạo dòng tiền bền vững của doanh nghiệp.

Khả năng sinh lời

Bảng 10: Các chỉ số tài chính

	2020	2021	2022	2023	2024
ROE	11.54	10.54	12.60	12.15	11.53
ROA	3.48	3.25	3.95	3.86	3.61
ROIC	5.07	4.54	5.63	5.62	5.20
Tỷ suất lợi nhuận gộp	18.20	17.50	19.00	18.80	18.50
Biên lợi nhuận ròng	3.20	2.90	3.50	3.40	3.30



Nhận xét:

	Đánh giá	Nhận xét
ROE	ROE chỉ dao động quanh 11.5%–12.6%	Chưa tạo ra giá trị cao cho cổ đông, nhưng vẫn ổn so với ngành công nghệ có tính cạnh tranh cao
ROIC	ROIC tăng từ 5.07% (2020) → 5.63% (2022), sau đó giảm nhẹ còn 5.20% (2024)	Có cải thiện nhưng chưa duy trì liên tục
ROA	ROA dao động từ 3.25%–3.95%	Hiệu quả sử dụng tài sản chưa cao, nhưng có tính ổn định
Biên lợi nhuận ròng	Biên lợi nhuận ròng ổn định quanh mức 3.2%–3.5%	Ôn định nhưng chưa cao
Tỷ suất lợi nhuận gộp	Duy trì ổn định từ 17.5%–19.0%	Khả năng kiểm soát chi phí khá tốt

→ Hiệu quả tài chính của CMC trong giai đoạn 2020–2024 nhìn chung ổn định nhưng chưa thực sự nổi bật. Các chỉ số như ROE, ROA và ROIC đều duy trì ở mức trung bình, phản ánh doanh nghiệp hoạt động hiệu quả ở mức vừa phải, nhưng chưa tạo ra giá trị vượt trội cho cổ đông. Trong khi đó, tỷ suất lợi nhuận gộp và biên lợi nhuận ròng duy trì ổn định, cho thấy CMC có khả năng kiểm soát chi phí tốt và sở hữu mô hình kinh doanh khá bền vững. Để nâng cao vị thế cạnh tranh và thu hút nhà đầu tư dài hạn, CMC cần tiếp tục cải thiện hiệu suất sinh lời và tối ưu hóa hiệu quả sử dụng vốn.

Khả năng thanh toán

Bảng 11: Tỷ số thanh khoản qua các năm

	2020	2021	2022	2023	2024
Tỷ số thanh toán hiện hành	1.45	1.36	1.21	1.14	1.25
Tỷ số thanh toán nhanh	1.20	1.09	1.03	1.03	1.10

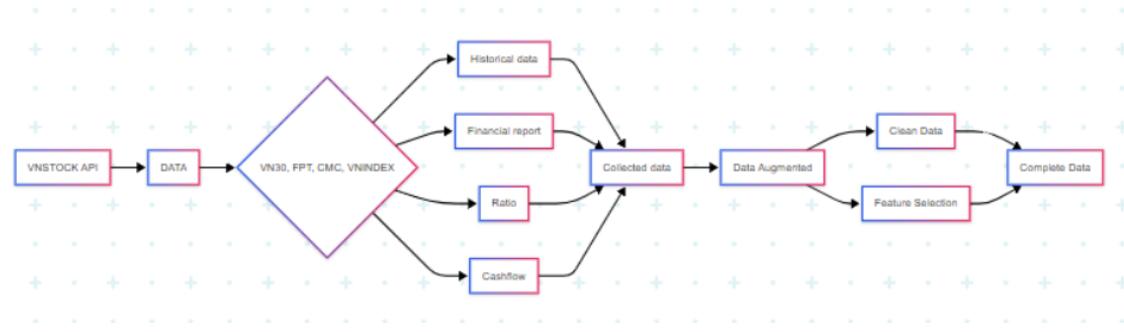
Nhận xét:

- Tỷ số thanh toán hiện hành: Giảm từ 1.45 (2020) xuống 1.14 (2023), sau đó tăng nhẹ lên 1.25 (2024) → Điều này cho thấy khả năng thanh toán nợ ngắn hạn bằng tài sản lưu động của công ty có xu hướng giảm trong giai đoạn đầu, nhưng đã có dấu hiệu cải thiện vào năm 2024.
- Tỷ số thanh toán nhanh: Giảm từ 1.20 (2020) xuống 1.03 (2022–2023), rồi tăng lên 1.10 (2024) → Mặc dù tỷ số này vẫn trên mức 1, phản ánh khả năng thanh toán nợ ngắn hạn mà không cần bán hàng tồn kho, nhưng xu hướng giảm cho thấy công ty cần chú ý đến việc duy trì đủ tài sản dễ chuyển đổi thành tiền.

Chương 3: Xây Dựng Mô Hình Dự Đoán Giá Cổ Phiếu

3.1 Phân tích dữ liệu

3.1.1. Thu thập dữ liệu



- Trên đây là pipeline cho quá trình thu thập dữ liệu. Chúng tôi sử dụng VNSTOCKS API để lấy dữ liệu về lịch sử giá cả, báo cáo tài chính và các chỉ số thị trường của hai công ty FPT và CMC. Bên cạnh đó chúng tôi sử dụng thêm dữ liệu từ chỉ số VN30 - đại diện cho 30 công ty có giá trị vốn hóa và thanh khoản đầu tiên trên sàn HOSE, và chỉ số VNINDEX để phản ánh biến động giá toàn bộ cổ phiếu đại diện cho thị trường.

- Thời gian: Từ năm 2017 - Đầu tháng 4 /2025.
- Tính chính xác và minh bạch: Đảm bảo do API gọi tới các trang web public và có uy tín như VietStock.vn, SSI.com, CafeF.vn, Trading View,...

3.1.2. Mô tả dữ liệu

Nhằm tăng cường số lượng điểm dữ liệu để huấn luyện, ta sẽ tính toán và tăng cường thêm các đặc trưng là các chỉ số kỹ thuật cộng thêm các chỉ số về thị trường. Sau đó chúng ta sẽ kết hợp từ các đặc trưng được lấy từ dòng tiền và báo cáo tài chính và các chỉ số, chi tiết xem tại bảng ở **[Phụ lục 1]**. Cuối cùng ta sẽ được một trận đầy đủ toàn bộ đặc trưng được đưa vào huấn luyện.

Đặc trưng của FPT hoặc CMC

Đặc trưng của VN30

Đặc trưng của VNINDEX

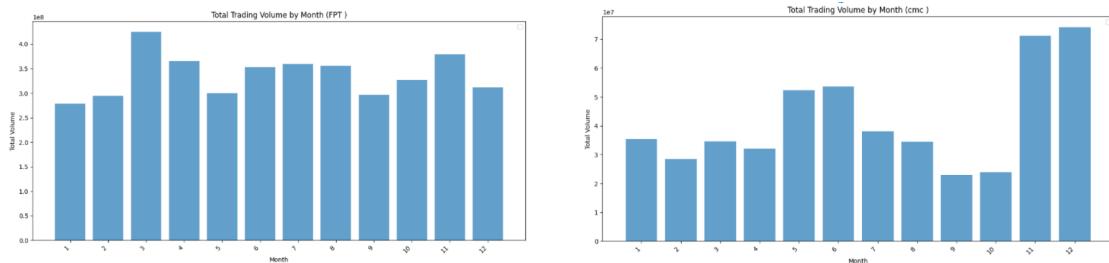
3.1.3. Tiền xử lý

- Do một số đặc trưng là các chỉ số tính theo chu kỳ tần xuất nên sẽ thừa và thiếu nhiều điểm dữ liệu nên chúng ta sẽ fill chúng bằng Linear Interpolation (nội suy tuyến tính)
- Với 210 đặc trưng của FPT và CMC, nguy cơ nhiều và dư thừa cao, khi huấn luyện có thể dẫn tới hiện tượng overfitting và giảm độ chính xác. Chúng ta sẽ trích chọn đặc trưng.
- Phương pháp trích chọn đặc trưng: Dựa vào tương quan tuyến tính và ý nghĩa thống kê (p-value). Cụ thể sẽ lọc các đặc trưng mà :
 - + Hệ số tương quan với close price có trị đối <0.1
 - + p-value > 0.05
 - + Ngoài ra các đặc trưng liên quan tới domain knowledge như các chỉ số tài chính, thị trường sẽ được ép buộc giữ lại để đảm bảo tính nhất quán.
- Sau khi lọc thì sẽ còn **164** đặc trưng.
- Nhận thấy các đặc trưng có vùng giá trị khác biệt, chẳng hạn như close price sẽ có vùng giá trị >30 trong khi đặc trưng Fixed Asset có vùng giá trị thuộc $[0,1]$. Vì thế chúng tôi sẽ lựa chọn scale toàn bộ các giá trị bằng Min-Max Scaler kết hợp với Power Transformation.

3.1.4. Trực quan hóa dữ liệu

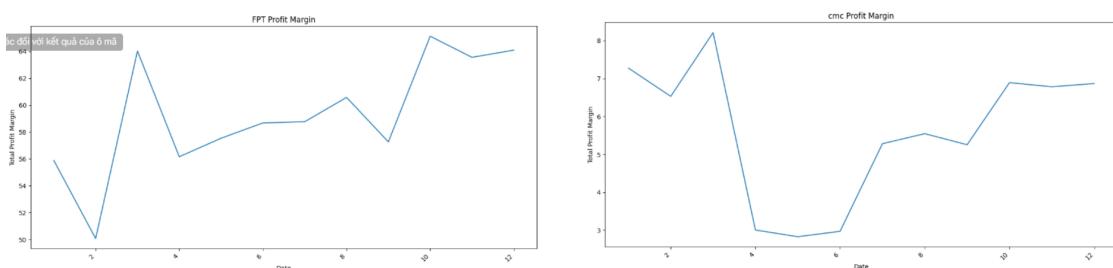
- Trực quan hóa dữ liệu của 2 công ty:
 - + Tổng khối lượng giao dịch theo tháng trong 7 năm của 2 công ty.
 - + Biên lợi nhuận theo tháng trong 7 năm của 2 công ty.

- + Phân phối giá đóng cửa của 2 mã cổ phiếu.
- + Phân phối của hệ số tương quan giữa giá đóng cửa 2 mã cổ phiếu và các yếu tố khác.



Hình 5: Biểu đồ thê hiện tổng khối lượng giao dịch theo tháng

- **Nhận xét:** Dữ liệu khối lượng giao dịch của 2 công ty theo các tháng trong vòng 7 năm:
 - + Khối lượng giao dịch theo tháng của FPT cao, ổn định, cao nhất vào tháng 3, thấp nhất vào tháng 1 → giao dịch ổn định, phản ánh tính thanh khoản của cổ phiếu FPT cao.
 - + Khối lượng giao dịch theo tháng của CMC dao động từ 3.5×10^7 → 7.5×10^7 , biến động không đều, tăng đột biến vào tháng 5, 6, 11, 12, giảm mạnh vào tháng 9, 10 → CMC là cổ phiếu có tính thanh khoản trung bình. Biến động khối lượng giao dịch phản ánh rủi ro thời vụ và phụ thuộc tin tức.



Hình 6: Biểu đồ thê hiện biên lợi nhuận theo tháng

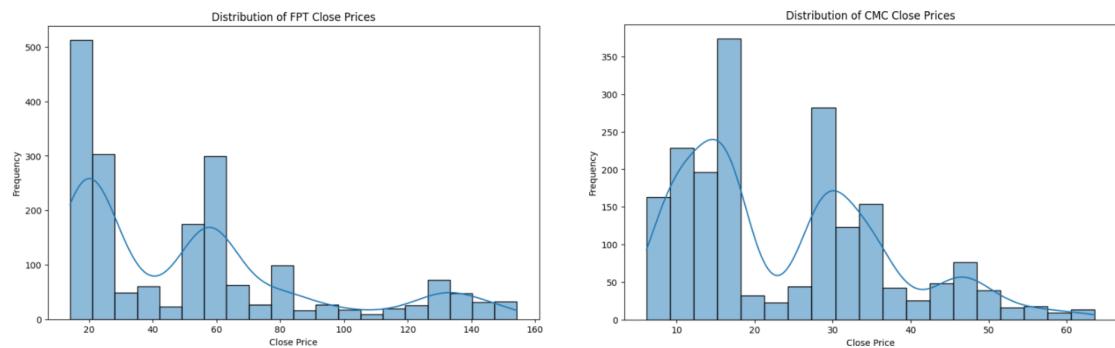
- **Nhận xét:** Dữ liệu biên lợi nhuận của 2 công ty theo các tháng trong vòng 7 năm:
 - + Biên lợi nhuận của FPT biến động giữa các tháng trong năm. Trong đó tháng cao nhất là tháng 3 và tháng 10, tháng thấp nhất là tháng 2.
 - Giá đóng cửa FPT có thể chịu ảnh hưởng rõ rệt bởi biên lợi nhuận theo mùa vụ,

tháng 3 và 10 (biên lợi nhuận cao) thường trùng với các sự kiện quan trọng:

- Tháng 3: Thời điểm công bố báo cáo tài chính năm trước và kế hoạch năm mới.
- Tháng 10: Giai đoạn công bố kết quả kinh doanh quý III.

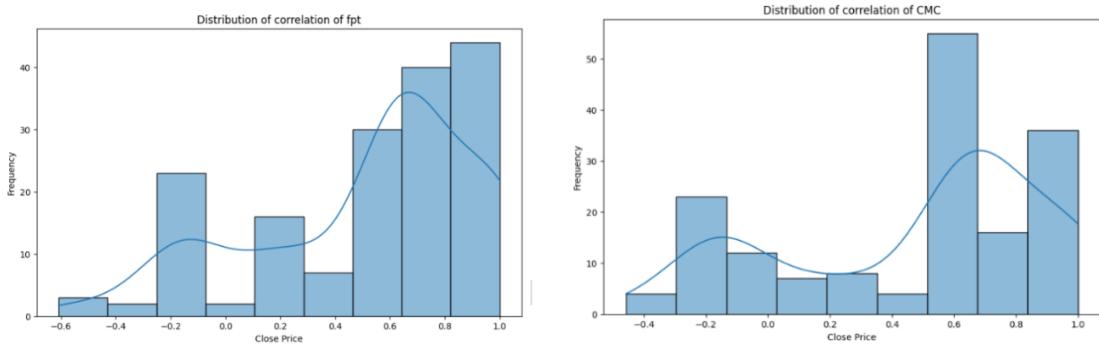
+ Biên lợi nhuận của CMC biến động mạnh giữa các tháng trong năm, giảm mạnh vào khoảng tháng 4 đến tháng 6, tăng cao nhất vào tháng 3, nguyên nhân có thể đến từ:

- Tháng 3: Thời điểm công bố báo cáo tài chính năm trước và kế hoạch năm mới.



Hình 7: Biểu đồ thể hiện sự phân phối giá đóng cửa của 2 công ty

- Giá đóng cửa của cổ phiếu FPT:
 - + Giá đóng cửa dao động chủ yếu trong khoảng 20–160. Tần suất cao nhất (khoảng 300 lần) tập trung ở mức giá 20, cho thấy đây là vùng giá phổ biến.
 - + Xu hướng: phân phối lệch phải → giá đóng cửa chủ yếu ở mức thấp (20-60).
 - Giá đóng cửa của cổ phiếu CMC:
 - + Giá đóng cửa dao động trong khoảng 10-60. Tần suất cao nhất khoảng trên 350 lần tập trung ở mức giá sấp sỉ 18, cho thấy đây là mức giá phổ biến.
 - + Xu hướng lệch phải → giá chủ yếu ở mức 10 - 30, giá trị cổ phiếu ít khi tăng lên đến 40 - 60.
- Như vậy có thể thấy giá cổ phiếu FPT biến động nhiều hơn giá cổ phiếu CMC do khoảng biến động giá rộng hơn. Giá đóng cửa phổ biến của cổ phiếu FPT 20 nghìn VND cao hơn giá đóng cửa cổ phiếu CMC 18 nghìn VND.



Hình 8: Biểu đồ phân phối của hệ số tương quan giữa giá đóng cửa 2 mã cổ phiếu và các yếu tố khác

- Đối với giá đóng cửa của cổ phiếu của FPT:
 - + Hệ số tương quan của các đặc trưng khác với close price của FPT dao động trong khoảng (-0.6;1.0), tần số trong khoảng từ (2;45)
 - + Đồ thị về phân phối tương quan giá đóng cửa của cổ phiếu FPT lệch trái, cho thấy phần lớn các hệ số tương quan dương, gần 1.0 → có nhiều yếu tố tác động dương đối với sự biến động của giá cổ phiếu và tác động có giá trị gần 1.0 cho thấy các yếu tố tác động mạnh đến giá đóng cửa của cổ phiếu FPT.
 - + Hệ số tương quan trong khoảng (0.8;1.0) có tần suất lớn nhất = 45 → có 40 yếu tố tác động đến giá cổ phiếu có hệ số tương quan trong khoảng (0.8;1.0), tương quan trong khoảng (-0.4;-0.2) có tần suất nhỏ nhất sấp sỉ 2.
- Đối với giá đóng cửa của cổ phiếu CMC:
 - + Hệ số tương quan của giá đóng cửa của CMC với các yếu tố bên ngoài dao động trong (-0.4;1.0), tần số trong khoảng từ (3;55).
 - + Đồ thị về phân phối tương quan giá đóng cửa của cổ phiếu CMC cho thấy hệ số tương quan gần 0.6 có tần suất lớn nhất trên 50 lần xuất hiện → có 50 yếu tố tác động đến giá cổ phiếu có hệ số tương quan gần 0.6, hệ số tương quan = 0.4 và -0.4 có tần suất nhỏ nhất sấp sỉ 3.
 - Giá đóng cửa của cả hai mã cổ phiếu đều tương quan dương và mạnh với nhiều yếu tố bên ngoài, trong đó, cổ phiếu của FPT chịu tác động của ít yếu tố hơn CMC nhưng mức độ tương quan mạnh hơn cổ phiếu của CMC.

```
#top 10 correlation với closeprice fpt
top_10_correlation = correlation_matrix['Close_FPT'].abs().nlargest(20)
print(top_10_correlation)

Close_FPT           1.000000
cumulative_return_FPT 1.000000
EMA_short_term3_FPT  0.999867
High_FPT            0.999866
Low_FPT             0.999817
MA_short_term3_FPT  0.999777
SMA_3_FPT           0.999777
Open_FPT            0.999664
EMA_short_term7_FPT 0.999482
TSF_3_FPT           0.999383
TSF_7_FPT           0.999352
MA_short_term7_FPT  0.999268
SMA_7_FPT           0.999268
TSF_21_FPT          0.998777
EMA_mid_term21_FPT 0.998034
MA_mid_term21_FPT  0.997341
SMA_21_FPT          0.997341
Upper_Band_FPT     0.997137
TSF_63_FPT          0.997018
TSF_100_FPT         0.995115
Name: Close_FPT, dtype: float64
```

Hình 9: Những yếu tố có tương quan mạnh nhất đến giá đóng cửa của FPT

- Nhận xét:

+ cumulative_return_FPT có $r = 1.0 \rightarrow$ lợi nhuận tích lũy tương quan hoàn hảo với giá đóng cửa, điều này là hợp lý vì lợi nhuận tích lũy được tính dựa trên giá đóng cửa.

+ Các chỉ số kỹ thuật có tương quan rất cao (>0.99) trong đó:

- Các chỉ số trung bình động ngắn hạn (EMA, MA, SMA với chu kỳ 3 và 7 ngày) đều có $r > 0.999$
- Các chỉ số giá cao nhất (High_FPT) và thấp nhất (Low_FPT) trong ngày cũng có $r > 0.999$

+ Xu hướng giảm dần theo chu kỳ:

- Các chỉ số với chu kỳ dài hơn (21, 63, 100 ngày) có hệ số tương quan thấp hơn so với các chỉ số ngắn hạn: EMA_short_term3 ($r=0.999867$) $>$ EMA_short_term7

($r=0.999482$) > EMA_mid_term21 ($r=0.998034$).

- TSF có tương quan cao với các chu kỳ khác nhau, trong đó TSF_3_FPT ($r=0.999383$) > TSF_7_FPT ($r=0.999352$) > TSF_21_FPT ($r=0.998777$) > TSF_63_FPT ($r=0.997018$).

```
top_10_correlation = correlation_matrix['Close_FPT'].abs().nlargest(10)
print(top_10_correlation)
```

Month	0.012388
RSI_long_term100_VNINDEX	0.104270
CCI_63_VN30	0.105860
ADX_21_VNINDEX	0.109599
MFI_21_FPT	0.115118
Williams_R_100_VN30	0.117578
Williams_R_100_FPT	0.122040
return_week_std_VN30	0.122419
MFI_100_VN30	0.125742
Momentum_100_VNINDEX	0.126557
Name: Close_FPT, dtype: float64	

Hình 10: Những yếu tố có tương quan yếu nhất đến giá đóng cửa của FPT

- **Nhận xét:** + Các chỉ số có tương quan cực thấp, gần như không tác động đến giá đóng cửa của FPT.

- Month ($r=0.012388$): Tháng trong năm gần như không ảnh hưởng đến giá FPT, cho thấy giá cổ phiếu này không theo mô hình mùa vụ rõ ràng.
- Các chỉ số về thị trường (RSI, CCI, ADX, Williams

→ Khi phân tích có thể bỏ qua hoặc ít quan tâm đến: Các chỉ số dao động dài hạn
Yếu tố mùa vụ Chỉ số thị trường chung (trừ khi có biến động cực lớn)

Top features		
Close_CMC		1.000000
cumulative_return_CMC	1.000000	
EMA_short_term3_CMC	0.999523	
High_CMC	0.999422	
Low_CMC	0.999238	
MA_short_term3_CMC	0.999232	
SMA_3_CMC	0.999232	
Open_CMC	0.998772	
EMA_short_term7_CMC	0.998088	
TSF_3_CMC	0.997994	
TSF_7_CMC	0.997756	
MA_short_term7_CMC	0.997187	
SMA_7_CMC	0.997187	
TSF_21_CMC	0.994699	
EMA_mid_term21_CMC	0.993020	
MA_mid_term21_CMC	0.990864	
SMA_21_CMC	0.990864	
Upper_Band_CMC	0.989156	
TSF_63_CMC	0.986168	
Lower_Band_CMC	0.980983	

Hình 11: Những yếu tố có tương quan có tương quan mạnh nhất đến giá đóng cửa của CMC

- Nhận xét:

+ cumulative_return_CMC có $r = 1.0 \rightarrow$ lợi nhuận tích lũy tương quan hoàn hảo với giá đóng cửa, điều này là hợp lý vì lợi nhuận tích lũy được tính dựa trên giá đóng cửa.

+ Các chỉ số kỹ thuật có tương quan rất cao ($r > 0.99$) trong đó:

- Các chỉ số trung bình động ngắn hạn (EMA, MA, SMA với chu kỳ 3 và 7 ngày) đều có $r > 0.99$
- Các chỉ số giá cao nhất (High_CMC) và thấp nhất (Low_CMC) trong ngày cũng có $r > 0.999$

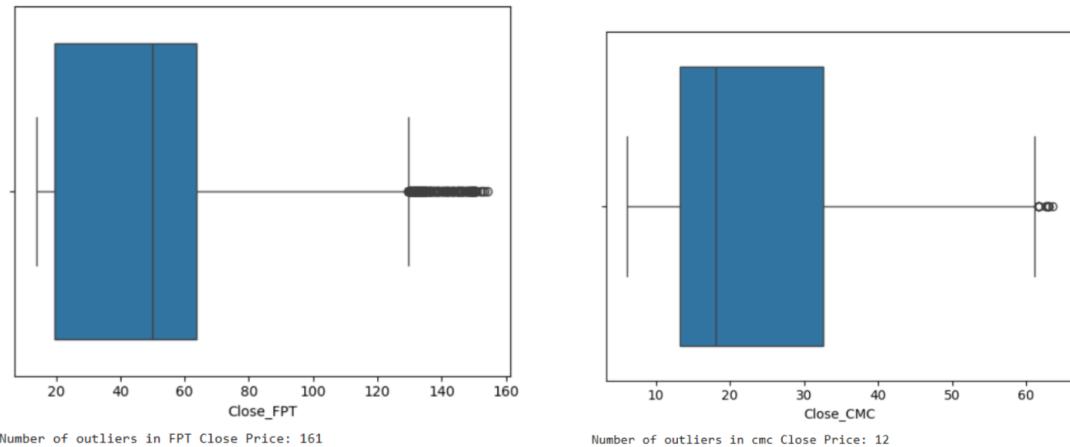
+ Xu hướng giảm dần theo chu kỳ:

- Các chỉ số với chu kỳ dài hơn (63, 100 ngày) có hệ số tương quan thấp hơn so với các chỉ số ngắn hạn: EMA_short_term3 ($r=0.999523$) > EMA_short_term7 ($r=0.998088$) > EMA_mid_term21 ($r=0.993020$).

- TSF có tương quan cao với các chu kỳ khác nhau, trong đó TSF_3_FPT ($r=0.997994$) > TSF_7_FPT ($r=0.997756$) > TSF_21_FPT ($r=0.994699$) > TSF_63_FPT ($r=0.986168$).

Bottom featuresMonth	
MFI_63_VNINDEX	0.101250
volatility_month_CMC	0.105959
RSI_mid_term63_VN30	0.110084
MFI_21_VN30	0.115164
RSI_mid_term63_VNINDEX	0.119548
ROCR_63_VN30	0.121943
Momentum_63_VN30	0.121943
RSI_long_term100_VN30	0.125602
Williams_R_100_CMC	0.126309

Hình 12: Những yếu tố có tương quan có tương quan yếu nhất đến giá đóng cửa của CMC



Hình 13: Số outlier của Close Price của 2 công ty

```
Number of outliers in FPT Close Price (pct_change > 15%): 0
Number of outliers in CMC Close Price (pct_change > 15%): 0
```

Hình 14: Số outlier của Close Price có pct_change > 15%

- **Nhận xét:** Mặc dù có những mức giá cao được ghi nhận, nhưng mức độ biến động giá giữa các phiên vẫn nằm trong ngưỡng ổn định - không có hiện tượng “nhảy giá” cực đoan xảy ra trong ngắn hạn.

→ KHÔNG CÓ OUTLIER THEO LOGIC THỊ TRƯỜNG

3.1.5. Giả thuyết và kiểm định

- Giả thuyết và kết quả kiểm tra giả thuyết chi tiết ở bảng trong [\[Phụ lục 1.1\]](#)
- Kết quả kiểm chứng các giả thuyết thống kê đều hợp lý và có cơ sở khoa học, phù hợp với lý thuyết tài chính và đặc thù của thị trường chứng khoán Việt Nam.

3.2 Mô hình phân loại xu hướng cổ phiếu

- Dữ liệu sử dụng: Lịch sử chứng khoán của hai công ty FPT và CMC từ năm 2017 tới năm 2025 gồm high, low và column của cổ phiếu hai công ty.
- Tiền xử lý dữ liệu:
 - + Chuẩn hóa column bằng cách chia cho 1000000
 - + Bỏ các ngày thị trường đóng cửa
 - + Biến mục tiêu: $(high + low)/2$
 - + Dùng hàm lượng giác để sử dụng chu kỳ thời gian như một biến định lượng cho mô hình.
 - + Tạo các đặc trưng thống kê như MA, EMA, MACD, OBV, Bollinger Bands
 - + Tạo các đặc trưng về mô tả sai số như volatility và pct change
 - + Chia tập train test theo tỉ lệ 8:2
- Ta sẽ xây dựng 5 target dựa trên tiêu chí sau:
 - + strong_up : daily change pct > 1.5
 - + moderate_up: daily change pct thuộc khoảng $[0.5, 1.5]$
 - + sideways: daily change pct thuộc khoảng $[-0.5, 0.5]$
 - + moderate_down: daily change pct thuộc khoảng $[-1.5, -0.5]$
 - + strong down: daily change pct < -1.5
- Xây dựng mô hình : Ta kết hợp XGBoost và Random Forest với trọng số 0.6:0.4 để dự đoán biến động giá.
- Xây dựng chiến lược sử dụng vốn hiệu quả: input gồm prediction và độ tin cậy

cộng với giá cả thị trường, ta sẽ xây dựng chiến lược như sau:

- + Mở vị thế: Dự đoán có độ tin cậy cao và chưa có nhiều tín hiệu.
- + Quản lý vị thế: Chốt lời nếu đạt take profit , cắt lỗ nếu stop loss, đóng lệnh nếu giữ quá lâu.

Take Profit

- Nếu bạn đang giữ lệnh **Long** (mua vào):

- Take Profit khi:

$$\text{current price} \geq \text{entry price} \times (1 + \text{take profit}\%)$$

- Nếu bạn đang giữ lệnh **Short** (bán khống):

- Take Profit khi:

$$\text{current price} \leq \text{entry price} \times (1 - \text{take profit}\%)$$

Stop Loss

- Nếu bạn đang giữ lệnh **Long** (mua vào):

- Stop Loss khi:

$$\text{current price} \leq \text{entry price} \times (1 - \text{stop loss}\%)$$

- Nếu bạn đang giữ lệnh **Short** (bán khống):

- Stop Loss khi:

$$\text{current price} \geq \text{entry price} \times (1 + \text{stop loss}\%)$$

- Dánh giá mô hình:

- + FPT: Strong up (95.3%) , Moderate Up (76.6%), Sideways (56.2%), Moderate down (73.4%), Strong down (79,7%). Những yếu tố ảnh hưởng nhất: pct của biến động giá, volatility 5,10,30 ngày; bolinger bandwidth.
- + CMC: Strong up (84.4%) , Moderate Up (82.8%), Sideways (64.1%), Moderate down (76.6%), Strong down (79,7%). Những yếu tố ảnh hưởng nhất: pct của biến động giá, tỉ số giữa khối lượng giao dịch và giá đóng trong 10, 30 ngày ,MA , volatility 30 ngày; bolinger bandwidth.

3.3 Mô hình dự đoán giá cổ phiếu

3.3.1. Giải pháp sơ khai

Ban đầu, chúng tôi sử dụng những model ML, DL đơn giản không tinh chỉnh để đánh giá sơ qua về độ nhiễu và độ phức tạp của bài toán.

- Biến mục tiêu: `close_shift_x`: giá đóng cửa sau x ngày nữa (x thuộc $\{3, 7, 21, 63\}$)
- Tại sao sử dụng phạm vi giá trị khác nhau: Các chỉ số kỹ thuật (ví dụ: RSI, MACD), khối lượng giao dịch, và giá cổ phiếu có thể có phạm vi giá trị rất khác nhau. MinMaxScaler và StandardScaler giúp đưa các biến này về cùng một thang đo, ngăn chặn việc một số biến có giá trị lớn hơn chi phối mô hình. Điều này đặc biệt quan trọng đối với các thuật toán nhạy cảm với khoảng cách như SVR.
- Phân phối không chuẩn: Giá chứng khoán và các chỉ số liên quan thường không tuân theo phân phối chuẩn. Điều này có thể làm ảnh hưởng đến hiệu suất của một số mô hình thống kê và học máy. Power Transformer có thể giúp biến đổi dữ liệu để nó gần với phân phối chuẩn hơn, cải thiện tính ổn định của phương sai và khả năng nâng cao hiệu suất của mô hình.

→ Do đó, Scaler: ở đây với những mô hình đơn, ta sẽ sử dụng những phương pháp scale sau:

- + **MinMax Scaler**: Chuyển đổi dữ liệu về một phạm vi nhất định, thường là $[0, 1]$. Nó giữ nguyên hình dạng phân phối ban đầu của dữ liệu.

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

- + **Standard Scaler**: Chuẩn hóa dữ liệu bằng cách loại bỏ giá trị trung bình và chia cho độ lệch chuẩn. Kết quả là dữ liệu có giá trị trung bình gần bằng 0 và độ lệch chuẩn bằng 1.

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

- + **Power Transformer**: Một nhóm các phép biến đổi nhằm ổn định phương sai và làm cho phân phối dữ liệu gần với phân phối Gaussian (chuẩn) hơn. Các phép biến đổi phổ biến bao gồm Box-Cox và Yeo-Johnson.

- Nếu $\lambda \neq 0$:

$$x^{(\lambda)} = \frac{x^\lambda - 1}{\lambda}$$

- Nếu $\lambda = 0$:

$$x^{(0)} = \ln(x)$$

- Ở đây chúng tôi sử dụng những mô hình như sau:

+ **Linear Regression**: label sẽ là `close_shift_x` và đặc trưng sẽ là toàn bộ các đặc trưng còn lại.

+ **SVM**: Scale bằng standard scaler kết hợp với SVR sau đó grid search với các tham số thử sẽ là

```
'svr_C' : [0.1, 1, 10, 100],  
'svr_gamma' : ['scale', 0.1, 0.01],  
'svr_kernel' : ['rbf']
```

+ **XGBOOSTs**: scale bằng standanrd scaler kết hợp power transformer, do bản chất XGBoost là mô hình ensemble nên việc đưa nhiều đặc trưng rất dễ gây nhiễu vì thế ở mô hình đơn chỉ chọn các đặc trưng cơ bản như `date`, `open`, `high`, `low`, `volumn`, `close`. Lookback của sổ trượt sẽ chọn là **150**.

+ **LSTM**: scale bằng standanrd scaler kết hợp power transformer, mô hình đơn chỉ chọn các đặc trưng cơ bản như `date`, `open`, `high`, `low`, `volumn`, `close`. Lookback của sổ trượt sẽ chọn là **30**. Cấu trúc gồm 4 tầng:

- Tầng **Feature Extraction**: thêm nhiều ngẫu nhiên ($std=0.2$) vào input và tầng khác để tăng tính tổng quát; decay để giảm ảnh hưởng của nhiễu. Sau đó giảm chiều input và chuẩn hóa.
- Tầng **Temporal Process**: học dữ liệu thời gian, dropout 30% và vẫn thêm nhiễu vào đầu.
- Tầng **Attention** (gồm 2 tầng con): sử dụng multihead attention kết hợp residual connection.

- Tầng **Mean/max Pooling**: tổng hợp thông tin dữ liệu và chuẩn hóa.
 - + Output layer: dùng hàm kích hoạt mish theo công thức sau kết hợp layernorm và dropout 50% để giảm overfitting.

$$\text{Mish}(x) = x \cdot \tanh(\text{softplus}(x)) = x \cdot \tanh(\ln(1 + e^x))$$

+ **biLSTM-CNN**: scale bằng standanrd scaler kết hợp power transformer, mô hình đơn chỉ chọn các đặc trưng cơ bản như `date`, `open`, `high`, `low`, `volumn`, `close`. Lookback của số trượt sẽ chọn là 150. Cấu trúc của mạng sẽ gồm:

- Tầng **CNN**: 64 kernel, bộ lọc kích thước 3, trượt theo thời gian , dropout 20% tránh overfitting.
- Tầng **BiLstm**: 2 lớp: lớp đầu nhận đầu vào 64 kênh từ CNN, đầu ra 128. Lớp sau gồm 128 từ lớp trước, tăng cường học các đặc trưng.
- Tầng **Attention**: giúp mạng lưới học hiệu quả hơn với kích thước đầu vào là 128 đầu ra 4 và tương thích với dữ liệu.
- Tầng **Dense**: giảm chiều từ 128 → 64 sau đó dropout 20% trước tầng cuối → 1

+ **Gru**: scale bằng standanrd scaler kết hợp power transformer, mô hình đơn chỉ chọn các đặc trưng cơ bản như `date`, `open`, `high`, `low`, `volumn`, `close`. Lookback của số trượt sẽ chọn là 150. 50 epochs.. Cấu trúc của Gru xây dựng gồm 4 tầng:

- Tầng **Gru**: 2 tầng xếp chồng 64 vector ẩn và sử dụng bidirectional
- Tầng **Batchnorm**: Chuẩn hóa dữ liệu để ổn định huấn luyện tầng dropout: 20%
- Tầng **Dense**: dùng hàm kích hoạt ReLu, 2 lớp ẩn, mỗi lớp sẽ giảm một nửa chiều dữ liệu, lớp cuối chỉ có 1 chiều, dropout 20%./

- Tiếp đến, chúng tôi sử dụng những Metric để đánh giá gồm MAE,RMSE, R²:

+ **MAE**: Sai số tuyệt đối trung bình, đo lường độ lớn trung bình của các sai số

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

+ **RMSE**: Sai số căn bậc 2 trung bình, Tính căn bậc hai của trung bình các bình phương sai số giữa giá trị dự đoán và giá trị thực tế. RMSE phạt các lỗi lớn hơn nhiều hơn MAE

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

+ **R²**: tỷ lệ phương sai của biến phụ thuộc có thể được dự đoán từ (các) biến độc lập. Nó cho biết mức độ phù hợp của mô hình

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

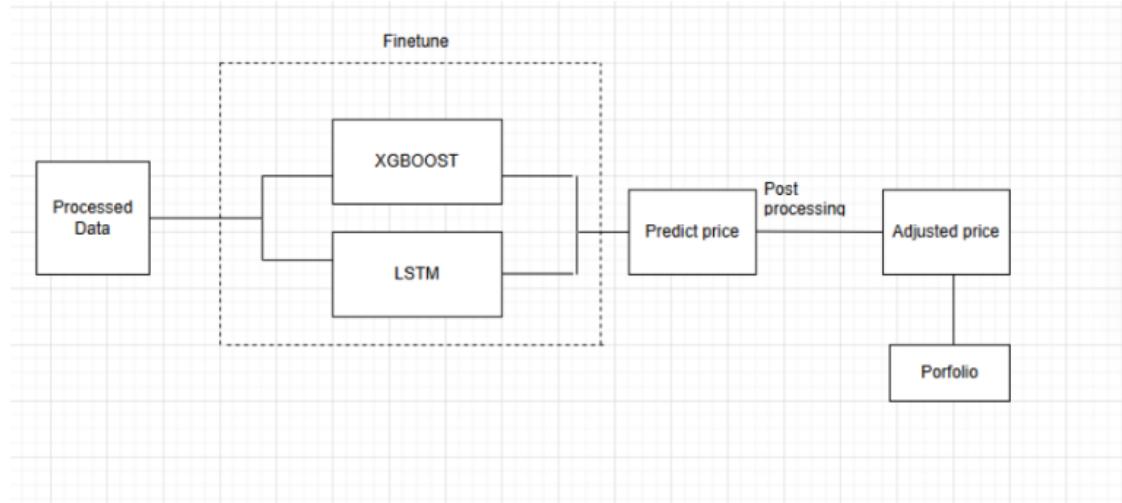
- Sau khi huấn luyện và đánh giá các mô hình, ta có bảng thống kê như sau:(ta sẽ ưu tiên đánh giá FPT trước do data có sự biến thiên phức tạp hơn CMC):

Khung thời gian	Mô hình	MAE	RMSE	R ²
3 ngày	Linear Regression	11.4163	14.7226	0.6494
	SVM	68.0729	76.0082	-9.2154
7 ngày	Linear Regression	11.9460	15.4657	0.6122
	SVM	69.2549	76.6878	-9.4725
21 ngày	Linear Regression	8.3620	10.5235	0.8188
	SVM	68.7887	76.4967	-9.6185
63 ngày	Linear Regression	12.7656	14.8795	0.6300
	SVM	67.3837	74.8636	-9.6358

Khung thời gian	Mô hình	MAE	RMSE	R²
3 ngày	XGBOOST	41.13	44.04	-6.4083
	GRU	53.91	59.64	-13.4376
	biLSTM	82.14	92.13	-10.08
	LSTM	7.24033	14.32269	9.291
7 ngày	XGBOOST	42.59	39.60	-5.9771
	GRU	15.21	16.44	0.07
	biLSTM	53.98	59.06	-3.32
	LSTM	14.71	17.27	5.15
21 ngày	XGBOOST	44.57	41.58	-6.7606
	GRU	18.18	20.00	0.04
	biLSTM	72.15	69.93	-5.5
63 ngày	XGBOOST	42.96	40.14	-6.4866
	GRU	12.85	15.24	0.7117
	biLSTM	74.49	90.91	-8.5735

Sau khi tổng hợp và đánh giá các mô hình đơn, dựa vào mức độ đánh giá và ưu điểm từng mô hình, chúng tôi chọn 2 model đơn để tiến tới bước sau: XGBOOST, LSTM.

3.3.2. Mô hình phức hợp



Hình 15: Pipeline quá trình huấn luyện

Giai đoạn 1 - Tinh chỉnh XGBOOST

- Tiền xử lý dữ liệu: Bắt đầu từ 5 đặc trưng cơ bản của giá cổ phiếu trong lịch sử: `high, low, open, close, volume`; ta tính toán và xây dựng thêm các đặc trưng mới bao gồm:
 - + `Close_shift_n` (n thuộc $3, 7, 21, 63$): biến mục tiêu.
 - + `MA_n, STD_n, EMA_n` (n thuộc $3, 7, 14, 21, 30, 63$).
 - + Chênh lệch giá trong khoảng thời gian n
 - + PCT change.
 - + Volatility: dung lượng trong khoảng thời gian $7, 14, 30, 60$
 - + `ROC_n, EWM_n, ROC_n, RSI_n`
 - + hiệu giữa `high` là `low` và tỉ lệ chinh lệnh.
- + Nhận thấy rằng mô hình XGBOOSTs không hiểu được dữ liệu timeseries(Date), vì thế, chúng tôi sẽ mã hóa dữ liệu thời gian thành các đặc trưng mới, đảm bảo được tính tuần hoàn của ngày tháng năm và thứ tự luân phiên của chúng như sau:
 - chỉ số tháng, $\sin(2\pi \times \text{chỉ số tháng}/12)$, $\cos(2\pi \times \text{chỉ số tháng}/12)$
 - $x = \text{Thứ tự ngày trong tuần}$, $\sin(2\pi \times x/12)$, $\cos(2\pi \times x/12)$
 - $z = \text{chỉ số ngày}$, $\sin(2\pi \times z/12)$, $\cos(2\pi \times z/12)$

- Do đây là dữ liệu thừa nên tương tự mục trên, ta cũng lấp đầy dữ liệu thiếu bằng phương pháp nội suy.
- Chia tỉ lệ train/test: 80/20
- Huấn luyện tinh chỉnh mô hình: Tinh chỉnh tìm ra tham số tốt nhất bằng optuna
 - + Optuna: sử dụng Tree_structured Parzen Estimator và Random Sampling để tìm ra siêu tham số trong không gian truy vấn, mỗi lần thử sẽ chọn các giá trị dựa trên phân phối xác xuất để khai thác vùng “nghi ngờ”. Tree_structured Parzen Estimator xây dựng hai phân phối xác xuất $I(x)$ là các điểm tốt đã thử và $g(x)$ là các điểm còn lại, nhiệm vụ của thuật toán là tìm x sao cho tối đa $I(x)/g(x)$.
 - + Optuna sẽ cắt sớm nếu các lần thử không có triển vọng tốt, vì thế cũng hiệu quả để chống overfitting.
- Vùng truy vấn: ở phụ lục
- 'objective': Mục tiêu là hồi quy, dùng hàm mất mát bình phương sai số
- 'eval_metric': 'rmse' Đánh giá bằng cản sai số bình phương trung bình
- Chọn thuật toán booster: cây quyết định (gbtree), tuyến tính (gblinear), hoặc kết hợp (dart)
 - Tham số điều chỉnh L2 (Ridge): [1e-8, 1.0]
 - Tham số điều chỉnh L1 (Lasso): [1e-8, 1.0]
 - Tỷ lệ cột được chọn ngẫu nhiên cho mỗi cây [0.3, 1.0]
 - Tỷ lệ mẫu được dùng cho mỗi cây [0.4, 1.0]
 - 'learning_rate': [0.001, 0.3]
 - Số lượng cây quyết định [100, 1000]
 - Độ sâu tối đa của cây :[3, 10]
 - Số lượng mẫu tối thiểu để chia một node: [1, 10]
 - Tham số regularization để kiểm soát độ phức tạp mô hình [1e-8, 1.0]
 - Kết quả biểu đồ so sánh giá cổ phiếu dự đoán với số liệu thực.
 - Dự đoán: Tại phần mục lục [**Phụ lục**]
- Đánh giá mô hình:

Bảng 12: Bảng đánh giá sai số của mô hình XGBOOST

Khung thời gian	Công ty	MAE	R ²	RMSE
3 ngày	FPT	4.0563	0.9591	5.0098
	CMC	1,2365	0,9452	2,1432
7 ngày	FPT	4.3349	0.9453	5.7554
	CMC	2.5133	0.7815	3.6653
21 ngày	FPT	6.9880	0.760	9.0345
	CMC	4.2513	0.3239	6.3474
63 ngày	FPT	13.5886	0.4141	15.9232
	CMC	6.7811	-0.8521	9.6708

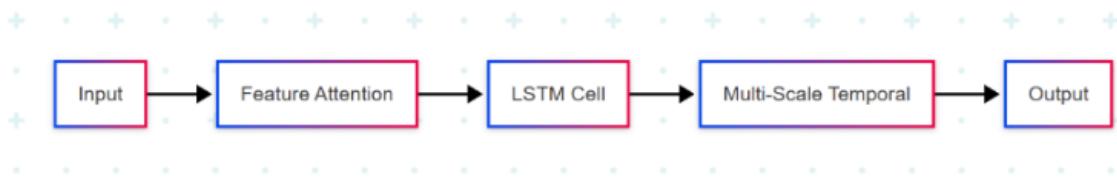
Giai đoạn 2 - Tinh chỉnh LSTM

1. Chuẩn hóa và phân tách dữ liệu

- Dữ liệu tập huấn luyện được chuẩn hóa bằng phương pháp MinMaxScaler với khoảng giá trị đầu ra là (-1, 1). Việc chuẩn hóa theo cách này là một thực tiễn phổ biến trong các bài toán dự báo chuỗi thời gian sử dụng mô hình học sâu, đặc biệt là kiến trúc như LSTM. Nguyên nhân là do hàm kích hoạt mặc định trong các lớp ẩn của LSTM thường là hàm tanh, vốn hoạt động hiệu quả nhất khi dữ liệu vào được phân bố quanh giá trị 0 và nằm trong khoảng (-1, 1). Chuẩn hóa dữ liệu theo khoảng này không chỉ giúp mô hình học nhanh hơn và ổn định hơn trong quá trình huấn luyện, mà còn góp phần giảm thiểu các vấn đề phô biến như hiện tượng mất hoặc bùng nổ gradient. Công thức chuẩn hóa bằng MinMaxScaler với đầu ra (-1, 1)

$$\chi' = -1 + \frac{2(x - x_{\min})}{x_{\max} - x_{\min}}$$

- Chúng tôi áp dụng kỹ thuật chia chuỗi thời gian theo phương pháp sliding window, với các kích thước cửa sổ tương ứng với các khoảng thời gian dự báo cụ thể: 3, 7, 21 và 63 ngày. Việc lựa chọn đa dạng độ dài cửa sổ này nhằm tối ưu hóa hiệu suất của mô hình trong từng khung thời gian dự báo—bao gồm ngắn hạn, trung hạn và dài hạn. Cách tiếp cận này đồng thời tạo tiền đề thuận lợi cho quá trình xây dựng mô



Hình 16: Pipeline

hình tổ hợp (ensemble), giúp cải thiện độ chính xác tổng thể của hệ thống dự báo bằng cách kết hợp thông tin đặc trưng từ nhiều mức độ biến động thời gian khác nhau.

- Tiếp đó, tập dữ liệu được phân tách thành ba bộ với tỉ lệ 80% cho huấn luyện, 10% cho xác thực và 10% cho kiểm thử, nhằm đảm bảo quá trình tối ưu tham số và kết hợp quả mô hình hiệu quả.

2. Xây dựng kiến trúc mô hình

- Tổng quan mô hình

+ Attention LSTM là một kiến trúc học sâu hiện đại, được thiết kế để khai thác hiệu quả các đặc trưng phân cấp trong chuỗi thời gian hoặc dữ liệu có cấu trúc lớp lang. Trong đó, LSTM đóng vai trò trung tâm như một khối học tuần tự mạnh mẽ, cho phép mô hình ghi nhớ các phụ thuộc, kết hợp Attention đánh giá đúng trọng số cần tập trung. Cấu trúc cụ thể được miêu tả như sơ đồ sau:

- + Kết hợp LSTM với hai lớp attention phân cấp:

- Attention đầu tiên hoạt động trên không gian đặc trưng đầu vào, cho phép mô hình xác định những đặc trưng quan trọng nhất tại mỗi bước thời gian.
- Attention thứ hai hoạt động theo trực thời gian, giúp mô hình học được các thời điểm quan trọng trong toàn bộ chuỗi, từ đó tập trung hơn vào những thông tin có tính quyết định cao.
- Điểm khác biệt nổi bật so với các kiến trúc khác là cách LSTM được đặt giữa hai tầng attention, không chỉ đóng vai trò truyền tải thông tin tuần tự mà còn làm nền tảng ổn định để hai tầng attention có thể hoạt động hiệu quả hơn. Nhờ đó, mô hình vừa học tốt quan hệ theo thời gian, vừa tự động xác định được trọng số chú ý ở cả cấp độ thấp (ngắn hạn) và cao (dài hạn).

- Chi tiết mô hình
- + Feature Attention:

- Mục tiêu: Cơ chế attention này được thiết kế nhằm học trọng số cho từng đặc trưng đầu vào, từ đó cho phép mô hình chính ưu tiên xử lý các đặc trưng có độ quan trọng cao hơn và giảm thiểu ảnh hưởng của các đặc trưng nhiễu. Việc này giúp cải thiện khả năng biểu diễn của mô hình và tăng hiệu quả học máy.
- Cấu trúc: Gồm có hai mạng con hai lớp Linear với hàm kích hoạt phi tuyến SiLU và Mish. Lớp tuyến tính đầu tiên thực hiện phép chiếu xuống không gian ẩn có kích thước thấp hơn nhằm nén thông tin đặc trưng. Sau đó, lớp tuyến tính thứ hai khôi phục kích thước ban đầu để tạo ra một mặt nạ attention mềm. Các hàm kích hoạt phi tuyến được chèn giữa và sau hai lớp giúp tăng tính phi tuyến và cải thiện khả năng biểu diễn của mạng con attention.

- + LSTM:

- Mục tiêu: Khai thác tính phụ thuộc thời gian trong chuỗi dữ liệu.
- Cấu trúc: gồm các lớp LSTM, một biến thể của mạng nơ-ron hồi tiếp (RNN), trong đó mỗi tế bào LSTM bao gồm ba cổng điều khiển: cổng vào (input gate), cổng quên (forget gate) và cổng đầu ra (output gate). Những cổng này sử dụng các hàm kích hoạt sigmoid và tanh để điều chỉnh luồng thông tin trong quá trình huấn luyện, từ đó cho phép mạng tự động chọn lọc thông tin quan trọng cần lưu trữ và loại bỏ thông tin không cần thiết qua từng bước thời gian. Công thức lan truyền tiên tiến như sau:

$$\text{Forget gate: } f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

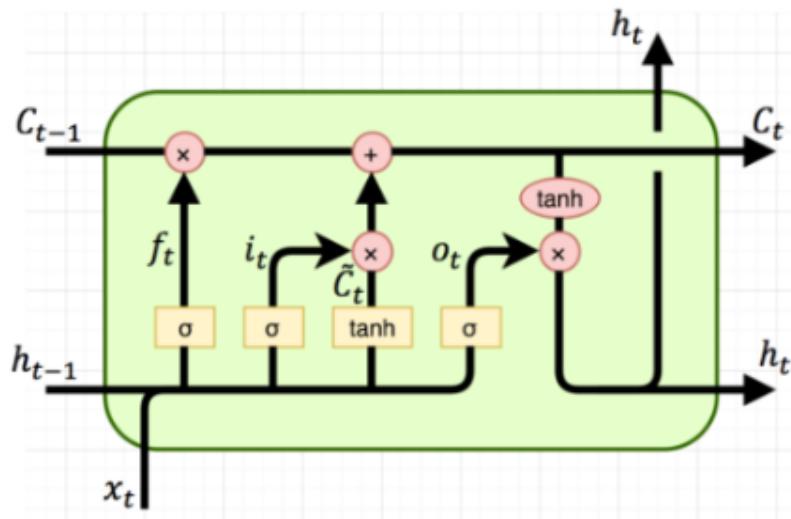
$$\text{Input gate: } i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

$$\text{Output gate: } o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$\text{Cell state update: } c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$\text{Hidden state update: } h_t = o_t \odot \tanh(c_t)$$

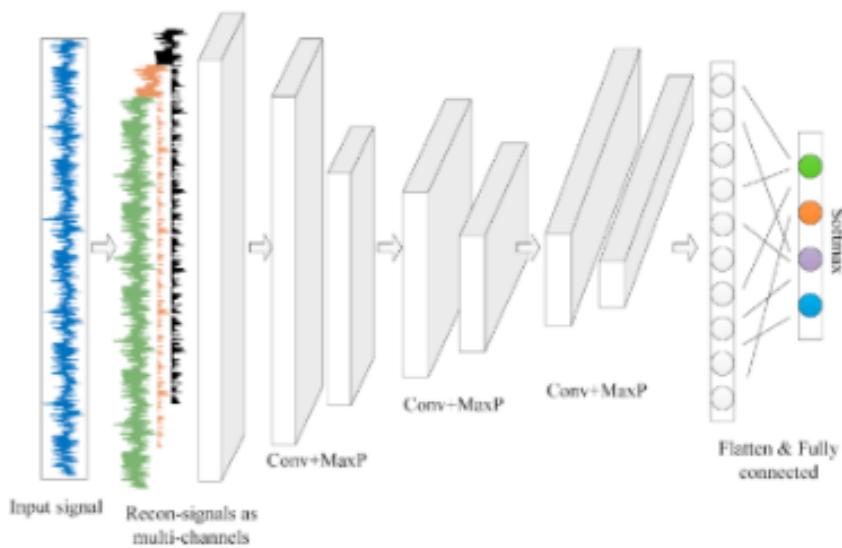


Hình 17: Sơ đồ cấu trúc một phân tử LSTM

+ MultiScale Temporal:

- Mục tiêu: Học các đặc trưng theo tỉ lệ thời gian. Nếu lớp ‘Feature Attention’ dùng để tính toán tầm quan trọng của từng đặc trưng thì lớp này tập trung vào tầm quan trọng của các đặc trưng đó trong khoảng thời gian nhất định.
- Cấu trúc: Sử dụng bộ lọc Convolution 1D di chuyển qua chuỗi dữ liệu đầu vào và xác định khoảng thời gian ảnh hưởng lớn nhất tới giá cổ phiếu tại thời điểm đang xét. Với công thức cập nhật trọng số như sau:

$$\mathbf{y}_i = \sum_{k=1}^m \mathbf{x}_{i+k-1} \mathbf{w}_k$$



Hình 18: Cấu trúc bộ lọc Conv1D

+ Output: Tầng đầu ra được thiết kế như một chuỗi các lớp tuyến tính có kèm theo các kỹ thuật phi tuyến. Tầng đầu tiên mở rộng không gian đặc trưng để tăng khả năng biểu diễn, tầng thứ hai lan truyền thông tin và giảm chiều dữ liệu trước khi đưa vào lớp tuyến tính cuối cùng để đưa ra dự báo với sai số thấp nhất.

- Các kỹ thuật kết hợp

+ Gaussian Noise: Các lớp nhiễu gauss được đặt trước khi đưa vào từng cấu trúc lớn của mô hình để tổng quát hóa, giúp mô hình tránh học các giá trị quá giống nhau dẫn đến overfitting. Cùng với đó chúng tôi giảm cường độ nhiễu qua mỗi epoch huấn luyện, giúp cho mô hình không thêm nhiễu một cách ‘khô cứng’. Công thức của nhiễu Gauss, với giá trị độ lệch chuẩn khởi tạo được khởi tạo nằm trong khoảng từ 0.01 và 0.05, sẽ biến đổi giá trị độ lệch chuẩn theo công thức sau:

$$x' = x + \mathcal{N}(0, \sigma^2)$$

+ Residual Connection kết hợp hệ số tùy chỉnh: Do độ sâu của kiến trúc, chúng tôi áp dụng skip connection có trọng số học được giữa các tầng, giúp thông tin không bị mất mát và tăng khả năng lan truyền gradient trong quá trình huấn luyện. Phương pháp này được sử dụng sau MultiScale Temporal, kết hợp trọng số đầu ra của chính

nó và LSTM với hệ số alpha tùy chỉnh. Công thức như sau:

$$\text{Output} = \text{Multiscale} + \alpha \cdot \text{LSTM}$$

- Tham số alpha được khởi tạo với phương pháp Xavier Initialization, giúp duy trì sự cân bằng giữa các gradient trong quá trình huấn luyện. Công thức khởi tạo tham số như sau:

$$W \sim \mathcal{N}(0, \sigma^2)$$

với $\sigma =$

$$\sigma = \sqrt{\frac{2}{\text{input dimension} + \text{output dimension}}}$$

- + LayerNorm: Giúp ổn định gradient, tăng tốc độ hội tụ, giảm sự phụ thuộc vào việc chọn học suất bằng cách chuẩn hóa đầu vào của mỗi lớp để các giá trị có phân phối chuẩn hơn. Công thức chuẩn hóa:

$$x_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

Sau đó,

$$y_i = \gamma x_i + \beta$$

- + Dropout: Tắt một tỉ lệ nhất định các neuron trong quá trình huấn luyện, giúp giảm overfitting mô hình.

3. Huấn luyện mô hình

- Tối ưu hàm mất mát: Hàm mất mát của chúng tôi được xây dựng bởi 3 thành phần chính:

- + Value Loss: Sử dụng hàm Huber Loss để đo lường sai số giữa giá trị dự đoán và thực tế. Huber Loss có khả năng kết hợp ưu điểm của Mean Squared Error và Mean Absolute Error, đặc biệt hiệu quả trong việc giảm ảnh hưởng của các ngoại lai, từ đó làm ổn định quá trình học. Với a là sai số dự đoán, công thức Huber Loss như sau:

$$L_\sigma(a) = \begin{cases} \frac{1}{2}a^2 & \text{nếu } |a| \leq \sigma \\ \sigma(|a| - \frac{1}{2}\sigma) & \text{ngược lại} \end{cases}$$

+ Direction Loss: Để mô hình không chỉ dự đoán chính xác giá trị mà còn nắm bắt đúng chiều hướng thay đổi (tăng hay giảm), ta sử dụng một hàm mất mát định hướng (direction-aware loss). Hàm mất mát này được thiết kế nhằm tối ưu hóa đồng thời:

- Đầu tiên, chúng tôi gán nhãn chiều hướng, $y = 1$ với giá tăng và $y = 0$ khi giá giảm. Sau đó, chúng tôi áp dụng BCE giữa chiều hướng dự đoán và nhãn đã làm mượt.

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

- Sau đó, chúng tôi thêm hàm ‘phạt’ các giá trị ngược về chiều.

$$L_{penalty} = \frac{1}{N} \gamma \sum_{i=1}^N \mathbb{I}[\Delta p_i^\wedge \Delta p_i < 0] L_{BCE}^{(i)}$$

- Vậy nên, sai số tổng quát sẽ là

$$L_{direction} = L_{BCE} + \gamma L_{penalty}$$

+ Correlation Loss: Đo lường mức độ tương quan giữa chuỗi dự đoán và chuỗi thực tế thông qua hệ số tương quan Pearson. Thành phần này thúc đẩy mô hình học các xu hướng dài hạn và mối quan hệ rộng hơn giữa các chuỗi thời gian, từ đó giảm thiểu hiện tượng overfitting và cải thiện khả năng dự báo trong dài hạn.

$$\rho_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

- Chiến lược huấn luyện mô hình: Để đảm bảo mô hình hội tụ nhanh và ổn định, đồng thời đạt được hiệu quả dự báo cao, chúng tôi áp dụng các kỹ thuật tối ưu hiện đại sau:

+ OneCycleLR Scheduler: Sử dụng lịch điều chỉnh học suất OneCycle, được xem là một trong những chiến lược hiệu quả nhất cho các bài toán học sâu hiện nay. Scheduler này giúp mô hình đạt tốc độ hội tụ nhanh hơn, tránh rơi vào các cực trị

địa phương không mong muốn.

+ Gradient Clipping: Áp dụng cắt gradient nhằm hạn chế hiện tượng exploding gradients, đặc biệt quan trọng trong các mô hình tuần tự như LSTM khi xử lý các chuỗi thời gian có biến động lớn. Việc này giúp quá trình cập nhật trọng số trở nên ổn định hơn và tránh phát sinh lỗi số học (NaN gradients). Nếu gradient vượt quá một khoảng cho trước, nó sẽ bị cắt về khoảng tối đa

+ Tối ưu siêu tham số bằng Optuna: Chúng tôi sử dụng thư viện Optuna, một thư viện sử dụng các thuật toán tối ưu hóa TPE và Random Search, giúp tự động hóa quá trình tìm kiếm các siêu tham số tối ưu cho mô hình và các tham số huấn luyện. Điều này đảm bảo rằng mô hình huấn luyện với tập hợp tham số tốt nhất được xác định từ không gian tìm kiếm đã định nghĩa, giúp nâng cao hiệu suất dự đoán và khả năng tổng quát hóa.

3. Dự đoán giá đóng cửa

Chi tiết xem tại phần [Phụ lục](#)

4. Dánh giá hiệu quả mô hình

- Kết quả biểu đồ so sánh giá cổ phiếu dự đoán với số liệu thực:

Bảng 13: Bảng đánh giá sai số của mô hình LSTM

Khung thời gian	Công ty	MAE	R ²	RMSE
3 ngày	FPT	6.42	0.73	8.68
	CMC	2.74	0.81	4.06
7 ngày	FPT	3	0.95	3.87
	CMC	4.64	0.81	4.06
21 ngày	FPT	4.94	0.85	6.25
	CMC	3.09	0.7	5.23
63 ngày	FPT	3.62	0.88	4.8
	CMC	4.68	0.6	6.4

Giai đoạn 3 -Tinh chỉnh và đánh giá chuyên sâu

1. Điều chỉnh giá bán và đánh giá các chỉ số đánh giá tài chính đầu tư

- Để tính toán chỉ số đánh giá tài chính Porfolio, ta sẽ điều chỉnh giá đóng cửa thành giá trị mới để tiện tính toán.
- Log return: Là tỉ suất sinh lời logarit đo lường mức sinh lời trong việc đầu tư chứng khoán trong khoảng thời gian (3,7,21,63 ngày). Điểm đặc biệt là log return không bị ảnh hưởng bởi giá gốc nên có thể so sánh giữa các công ty, từ đó dẫn tới quyết định đầu tư hợp lý.
- Công thức điều chỉnh giá với x là giá đóng cửa tại thời điểm

$$x_t = \ln \left(\frac{x_t}{x_{t-n}} \right) = \ln(x_t) - \ln(P_{x-n}) (n \in 3, 7, 21, 63)$$

Sau đó, ta sẽ dựa vào giá trị thực tế và giá trị dự đoán giá đóng cửa để đưa ra đánh giá với các chỉ số đánh giá tài chính đầu tư như sau. Gọi P là giá đóng cửa dự đoán:

- + Porfolio Actual Return: Lợi nhuận thực tế của danh mục đầu tư. Đây là chỉ số đo mức lợi nhuận (hoặc lỗ) của cổ phiếu hoặc danh mục đầu tư sau một khoảng thời gian, giúp nhà đầu tư đánh giá hiệu quả đầu tư trong khoảng thời gian yêu cầu, với công thức:

$$R_{\text{actual}} = \frac{P_{\text{end}} - P_{\text{start}}}{P_{\text{start}}} \quad (1)$$

- + Sharpe Ratio: tỉ số đo hiệu quả điều chỉnh theo rủi ro, mỗi đơn vị rủi ro mang bao nhiêu lợi nhuận. Tỉ số càng cao, khoản đầu tư càng hấp dẫn. Công thức của:

$$\text{Sharpe Ratio} = \frac{\mathbb{E}[r_t]}{\sigma_{r_t}} \quad (2)$$

- + Predict return: Lợi nhuận dự đoán, giúp nhà đầu tư đánh giá tiềm năng và ra quyết định mua bán. Công thức:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}} \quad (3)$$

- + Actual Variance: Phương sai lợi nhuận dự đoán, phương sai càng lớn thì độ rủi ro càng cao. Công thức:

$$\text{Var}(r_t) = \frac{1}{n} \sum_{i=1}^n (r_i - \bar{r})^2 \quad (4)$$

- Ta sẽ triển khai đánh giá hai công ty FPT và CMC với nhiệm vụ là dự đoán giá đóng cửa 3, 7, 21, 63 ngày tiếp theo.

Bảng 14: Bảng đánh giá tiềm năng đầu tư theo LSTM

Khung thời gian	Công ty	Portfolio Actual Return	Sharpe Ratio	Actual Variance	Predict Return
3 ngày	FPT	-0.07	-0.95	0.0013	0.02
	CMC	-0.13	-4.84	0	-0.02
7 ngày	FPT	-0.13	-0.66	0.0012	-0.02
	CMC	-0.33	-0.69	0.0008	-0.2
21 ngày	FPT	-0.23	-0.5	0.007	-0.097
	CMC	-0.33	-0.69	0.008	-0.2
63 ngày	FPT	-0.29	-0.28	0.0004	-0.42
	CMC	-0.42	-0.39	0.0005	-0.38

Bảng 15: Bảng đánh giá tiềm năng đầu tư theo XGBoost

Khung thời gian	Công ty	Portfolio Actual Return	Sharpe Ratio	Actual Variance	Predict Return
3 ngày	FPT	-0.07	-0.95	0.0013	0
	CMC	-0.13	-4.84	0	
7 ngày	FPT	-0.13	-0.66	0.0012	0
	CMC	-0.24	-1.23	0.0013	
21 ngày	FPT	-0.23	-0.5	0.0007	0
	CMC	-0.33	-0.69	0.008	
63 ngày	FPT	-0.29	-0.28	0.0004	
	CMC	-0.42	-0.39	0.0005	

- Đánh giá:

- + Công ty FPT: Công ty FPT hiện không hấp dẫn để đầu tư ngắn/ trung hạn do lợi nhuận thực tế luôn âm và áp lực bán suy giảm liên tục, rủi ro đầu tư ngắn hạn thấp hơn dài hạn, dù độ ổn định trong dài hạn tốt hơn. Đặc biệt là xu hướng trung hạn giảm rất mạnh trong 21 ngày, vì thế không nên đầu tư trong khoảng thời gian này. Điều này là chính xác vì dữ liệu chúng tôi thu thập là tới 9/4/2025, sau khi có sắc lệnh áp thuế của tổng thống Mỹ Donald Trump.
- + Công ty CMC: rất tiêu cực cho nhà đầu tư giữ dài hạn, ngoài ra rủi ro giảm dần theo thời gian cho thấy sẽ hợp lý hơn nếu đầu tư dài hạn nhưng vẫn không khuyến khích đầu tư, và độ tin cậy mô hình cao.
- + Về hai công ty, trong ngắn hạn thì FPT đáng để đầu tư và có hiệu suất tốt hơn, CMC thì mô hình học tốt hơn nhưng vẫn không nên đầu tư ngắn hạn và tuyệt đối không nên đầu tư dài hạn. Công ty FPT thì ngắn hạn có thể cân nhắc nhưng càng về lâu thì càng theo dõi thêm.

Chương 4: Triển Khai RAG

4.1 Thu thập dữ liệu

- Nguồn dữ liệu của dự án này bao gồm cả dữ liệu công khai và dữ liệu từ API:
- + Dữ liệu công khai:
 - Báo cáo tài chính: Dữ liệu được thu thập từ website của FPT và CMC, bao gồm các báo cáo tài chính dạng PDF/HTML.
 - Tin tức cổ phiếu: Dữ liệu từ các trang tin tức tài chính như CafeF, Vietstock, và RSS feeds.
 - Biên bản họp cổ đông: Được tải từ cổng thông tin doanh nghiệp (ví dụ: HOSE), cung cấp thông tin trực tiếp từ các cuộc họp cổ đông của FPT và CMC.
- + Dữ liệu từ API:
 - Google News API: Lọc tin tức liên quan đến FPT và CMC bằng các từ khóa "FPT cổ phiếu", "CMC chứng khoán".

- VNstock Python: Sử dụng thư viện VNstock để truy vấn dữ liệu thị trường thời gian thực.

- Phương pháp thu thập dữ liệu:

Dữ liệu được thu thập tự động thông qua các phương thức sau:

- + Tải về và lưu trữ tự động: Các báo cáo tài chính và biên bản họp cổ đông được tải định kỳ từ các website chính thức của FPT và CMC, đảm bảo cập nhật đầy đủ.
- + API lấy dữ liệu thời gian thực: Google News API và VNstock python cho phép thu thập nhanh chóng các tin tức và dữ liệu thị trường mới nhất.
- + RSS Feeds: Được sử dụng để cập nhật tin tức từ CafeF và Vietstock liên tục, giúp hệ thống luôn tiếp nhận các sự kiện mới.

4.2 Phương pháp thực hiện

4.2.1. Tiền xử lý

Do dữ liệu văn bản đến từ nhiều nguồn khác nhau với định dạng và cấu trúc đa dạng, việc làm sạch và chuẩn hóa dữ liệu là bước quan trọng để đảm bảo độ chính xác của mô hình RAG.

- Làm sạch văn bản:

- + Loại bỏ các HTML tags, ký tự đặc biệt, số điện thoại, email để đảm bảo văn bản chỉ chứa nội dung thông tin cần thiết.
- + Chuẩn hóa tiếng Việt: Xử lý dấu câu và chuyển toàn bộ văn bản về lowercase để giảm thiểu sự khác biệt không cần thiết giữa các từ.

- Chunking:

- + Văn bản được chia thành các đoạn nhỏ (500 tokens) bằng phương pháp LangChain Text Splitter, đảm bảo kích thước tối ưu cho mô hình LLM xử lý.
- + Phương pháp chia đoạn sử dụng tiêu chí ngữ nghĩa, giúp các đoạn văn bản vẫn giữ được tính kết nối logic

- Chuẩn hóa ký tự và định dạng:

- + Loại bỏ hoàn toàn các HTML, URL, và ký tự đặc biệt không cần thiết để tránh gây nhiễu cho mô hình.
- + Chuyển mã Unicode NFC và gộp các khoảng trắng để đảm bảo tính nhất quán

giữa các văn bản.

- Mapping trường dữ liệu:

+ **date**: Được chuẩn hóa theo định dạng ISO 8601, đảm bảo tính nhất quán trong toàn bộ dữ liệu.

+ **ticker**: Được xác định thông qua regex \b(FPT|CMG)\b, giúp phân biệt nhanh các thông tin liên quan đến FPT và CMC.

+ **source**: Được xác định cố định theo các nguồn cafef, dividend, shareholder, internal, giúp phân loại nguồn dữ liệu.

+ **text**: Ghép title và summary thành một đoạn văn bản hoàn chỉnh để tối ưu hóa ngữ nghĩa.

+ **Dữ liệu bảng**: Các bảng thông tin (cổ tức, DHDCD, giao dịch) được chuyển đổi thành câu hoàn chỉnh, với các giá trị trống được đánh dấu là “UNKNOWN” để tránh nhiễu thông tin.

- Xử lý giá trị thiếu và ngoại lệ:

+ Các trường dữ liệu bị thiếu sẽ được điền bằng “UNKNOWN” để tránh tình trạng mô hình xử lý không đồng nhất.

+ Các giá trị bất thường hoặc sai định dạng sẽ bị loại bỏ trong quá trình tiền xử lý để đảm bảo chất lượng dữ liệu đầu vào.

```

      record_date          date ticker \
0   2025-03-12 00:00:00 2025-03-12 00:00:00    FPT
1   2025-03-11 00:00:00 2025-03-11 00:00:00    FPT
2   2025-03-11 00:00:00 2025-03-11 00:00:00    FPT
3   2025-03-11 00:00:00 2025-03-11 00:00:00    FPT
4   2025-03-11 00:00:00 2025-03-11 00:00:00    FPT
...
982 2023-05-22 00:00:00 2023-06-20 00:00:00    CMG
983 2023-04-26 00:00:00 2023-05-25 00:00:00    CMG
984 2023-03-23 00:00:00 2023-04-21 00:00:00    CMG
985 2023-03-15 00:00:00 2023-04-13 00:00:00    CMG
986 2023-03-14 00:00:00 2023-03-14 00:00:00    CMG

           text      source
0  Phiên 12/3: Khối ngoại bán chiến biển hơn 900 ...    cafef
1  Chứng minh ngày mai (12-3): VN-Index tiếp tục ...    cafef
2  FPT "Bắt tay" Tỉnh Bắc Giang phát triển toàn d...    cafef
3  CTCK tự doanh không mong đợi trở lại "gom" một...    cafef
4  Chứng chỉ thoát hiểm 'phút 89'. Tóm tắt thị tr...    cafef
...
982 Loại giao dịch: GD của người liên quan. Người ... internal
983 Loại giao dịch: GD của người liên quan. Người ... internal
984 Loại giao dịch: GD của người liên quan. Người ... internal
985 Loại giao dịch: GD của người liên quan. Người ... internal
986 Loại giao dịch: GD CĐ lớn. Người thực hiện: Py... internal

```

4.2.2. Embedding

Trong dự án này chúng tôi sử dụng PhoBERT để embedding dữ liệu văn bản (các báo cáo tài chính, tin tức, biên lai, bản hợp đồng, dataset BTC đưa ra) nhằm tạo ra các vector biểu diễn từ ngữ có ý nghĩa ngữ nghĩa cao. Các vector này sẽ được sử dụng trong hệ thống RAG (Retrieval-Augmented Generation) để cải thiện khả năng tìm kiếm và trả lời câu hỏi.

- Sử dụng mô hình PhoBERT-base với kiến trúc Transformer 12 lớp, 768 chiều, và 12 đầu attention. PhoBERT-base với kích thước nhỏ gọn nhưng vẫn đảm bảo hiệu suất cao, giúp tối ưu tài nguyên tính toán. Mô hình được huấn luyện trên một khối lượng lớn văn bản tiếng Việt, giúp nắm bắt tốt ngữ cảnh ngữ nghĩa của văn bản.
- + Đầu vào là các đoạn văn bản (text) được chuyển đổi thành token ID với tokenizer của PhoBERT.
- + Đầu ra của PhoBERT là một tensor có kích thước [batch_size, seq_length, hidden_size].
- Mean Pooling: Để chuyển tensor này thành vector biểu diễn cố định, Mean Pooling

được áp dụng trên chiều seq_length. Công thức: Chuỗi token $w_{1:L}$. PhoBERT (RoBERTa-base vi) sinh tensor $H \in \mathbb{R}^{L \times 1024}$. Mean-Pooling:

$$\mathbf{v} = \frac{1}{|M|} \sum_{j=1}^L M_j H_j, \quad M_j = \begin{cases} 1 & \text{nếu } w_j \neq [\text{PAD}] \\ 0 & \text{khác} \end{cases}$$

Trong đó:

- $\mathbf{v} \in \mathbb{R}^{1024}$: Vector biểu diễn cuối cùng của toàn bộ đoạn văn bản
 - $\mathbf{H}_j \in \mathbb{R}^{1024}$: Vector biểu diễn của token thứ j trong chuỗi
 - L : Độ dài tối đa của chuỗi token (bao gồm cả padding)
 - $\|\mathbf{M}\|$: Số lượng token thực sự (không tính token [PAD])
- Chuẩn hóa vector: L2 Normalization: Vector biểu diễn được chuẩn hóa về đơn vị (norm = 1) để đảm bảo tính nhất quán và tránh ảnh hưởng của độ dài vector.
+ Công thức: $\hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}$. Vector có tính chất $\|\hat{\mathbf{v}}\|_2 = 1 \rightarrow$ giữ bất biến khoảng cách cosine-L2.

4.2.3. Vector search

Vector Search là một phương pháp tìm kiếm hiệu quả trong các hệ thống RAG (Retrieval-Augmented Generation), cho phép tìm kiếm dựa trên ngữ nghĩa và độ tương tự giữa các vector biểu diễn của văn bản. Trong hệ thống này, chúng tôi triển khai phương pháp Hybrid Search, kết hợp giữa FAISS (Vector Search) và TF-IDF (Lexical Search), cùng với kỹ thuật Metadata Filtering và trích xuất thông tin từ query bằng mô hình LLM Qwen 2.5 7b Instruct.

- Mục tiêu của hệ thống tìm kiếm này là cải thiện hiệu suất tìm kiếm và độ chính xác của các kết quả trả về thông qua:
- + Kết hợp Hybrid Search để cân bằng giữa khả năng tìm kiếm dựa trên từ khóa (TF-IDF) và ngữ nghĩa (Vector Search).
 - + Trích xuất thông tin từ query bằng LLM Qwen 2.5 7b Instruct, đảm bảo hiểu rõ ý nghĩa câu hỏi người dùng.
 - + Sử dụng Metadata Filtering để tối ưu hóa việc phân loại và lọc kết quả tìm kiếm.
 - Quy trình Hybrid Search và Metadata Filtering

1. Tìm kiếm kết hợp (Hybrid Search)

- + Truy vấn được chuyển thành vector nhờ PhoBERT và chuẩn hóa bằng L2 (FAISS).
- + Thực hiện tìm kiếm Vector Search trên FAISS với query vector để lấy danh sách kết quả ban đầu.
- + Thực hiện tìm kiếm TF-IDF (Lexical Search) trên toàn bộ văn bản và tính điểm tương đồng Cosine.
- + Điểm số của mỗi văn bản được tính theo công thức kết hợp Danh sách các kết quả được sắp xếp lại theo điểm số kết hợp.

2. Lọc kết quả bằng Metadata (Metadata Filtering)

- + Với mỗi kết quả được tìm thấy:
 - Lọc theo **ticker** nếu có (FPT, CMG).
 - Lọc theo **source** nếu có (CafeF, Vietstock, Google News, HOSE).
 - Lọc theo **date** nếu trong khoảng thời gian quy định.
 - Lọc theo **record_date** nếu có và nằm trong khoảng thời gian.
- + Nếu không đủ kết quả, hệ thống sẽ bổ sung thêm từ danh sách kết quả đã tìm ban đầu.

3. Trích xuất thông tin từ Query bằng LLM Qwen 2.5 7b Instruct

- + LLM Qwen 2.5 7b Instruct được sử dụng để phân tích query và xác định:
 - Đối tượng tìm kiếm (ticker, công ty).
 - Loại thông tin (báo cáo tài chính, tin tức, biên bản họp).
 - Khoảng thời gian (năm, quý).

+ Đầu ra từ LLM được sử dụng để xây dựng query cho hệ thống Hybrid Search và Metadata Filtering.

→ **Kết luận:** Phương pháp Vector Search với Hybrid Search kết hợp FAISS và TF-IDF (Lexical Search), cùng với việc trích xuất thông tin từ query bằng LLM Qwen 2.5 7b Instruct và Metadata Filtering, mang lại sự cân bằng giữa khả năng tìm kiếm chính xác theo từ khóa và ngữ nghĩa. Phương pháp này giúp nâng cao độ chính xác của hệ thống RAG trong việc tìm kiếm và trả lời câu hỏi.

4.2.4. Quy trình tạo đầu ra của hệ thống RAG

1. Truy xuất thông tin (Retrieve Phase):

- Hệ thống sử dụng mô hình Hybrid Search (FAISS + TF-IDF) để tìm kiếm các đoạn văn bản có độ tương đồng cao với truy vấn người dùng.
- Phương pháp tính điểm tương đồng bao gồm:
 - + Vector Search (FAISS): Sử dụng khoảng cách Cosine giữa vector của truy vấn và vector của các đoạn văn bản.
 - + Lexical Search (TF-IDF): Sử dụng điểm BM25 để đo độ liên quan dựa trên từ khóa.
 - + Điểm tổng hợp được tính theo công thức.:

$$\text{Score}_{\text{hybrid}} = \alpha \cdot \text{Score}_{\text{vector}} + (1 - \alpha) \cdot \text{Score}_{\text{lexical}} \quad (5)$$

- Hệ thống sắp xếp các kết quả theo điểm tổng hợp và chọn ra Top-K kết quả có điểm cao nhất.

2. Xử lý kết quả tìm kiếm:

- Các kết quả tìm kiếm được phân loại và lọc bằng Metadata Filtering:
 - + **ticker**: Lọc theo mã cổ phiếu (ví dụ: FPT, CMG).
 - + **date**: Lọc theo khoảng thời gian yêu cầu.
 - + **source**: Lọc theo nguồn dữ liệu (CafeF, Vietstock, Google News, HOSE).
- Các kết quả được định dạng lại để đảm bảo tính dễ đọc, mỗi kết quả bao gồm các trường thông tin như:
 - + Nguồn gốc thông tin.
 - + Ngày tháng.
 - + Nội dung văn bản tóm tắt.

3. Tạo phản hồi cuối cùng (Generation Phase):

- Mô hình ngôn ngữ lớn (LLM) nhận danh sách Top-K kết quả.
- Mô hình phân tích các kết quả và kết hợp chúng thành một phản hồi mạch lạc.
- Phản hồi có thể ở dạng:
 - + Tóm tắt tổng hợp từ nhiều kết quả.
 - + Đoạn văn bản trả lời câu hỏi của người dùng.

4.3 Kết quả phân tích và định hướng

- Hệ thống được chia làm ba thành phần chính:
- + Retriever (Truy xuất ngữ cảnh): Truy vấn đầu vào của người dùng được sử dụng để tìm các đoạn văn bản phù hợp nhất trong kho vector.
- + Generator (Mô hình ngôn ngữ lớn - LLM): Các đoạn văn được truy xuất được kết hợp với câu hỏi và đưa vào mô hình LLM (Qwen, llama) để sinh câu trả lời chính xác, có dẫn chứng.
- + Trình bày kết quả: Câu trả lời được trả lại kèm theo đoạn trích dẫn gốc nhằm đảm bảo tính minh bạch và khả năng kiểm chứng.

4.3.1. Kết quả

- Hệ thống có khả năng truy xuất chính xác các thông tin cổ tức, hiệu quả kinh doanh, và sự kiện liên quan đến ban lãnh đạo.
- Một số ví dụ tiêu biểu:
 - + Câu hỏi: "*FPT có chia cổ tức năm 2024 không?*"
→ Hệ thống trả lời có kèm dẫn chứng từ tập dữ liệu.
 - assistant
Dựa trên thông tin được cung cấp, FPT có lịch chia cổ tức vào ngày 9 tháng 5 năm 2024. Tuy nhiên, thông tin này chưa đề cập đến mức cổ tức cụ thể. Nguồn thông tin từ Cafef.
 - + Câu hỏi: "*Lợi nhuận quý I/2024 của CMG có tăng không?*"
→ Hệ thống trích đúng bảng số liệu từ báo cáo tài chính hợp nhất quý I/2024.
 - Độ chính xác phụ thuộc vào chất lượng vector hóa và độ dài chunk; chunk quá dài làm giảm độ phù hợp truy vấn, chunk quá ngắn làm mất mạch văn bản.

Dựa trên thông tin tham khảo mà bạn cung cấp, tôi thấy rằng có hai giao dịch lớn liên quan đến CMG vào ngày 26/03/2024 và 05/03/2024. Tuy nhiên, thông tin này không cung cấp chi tiết về lợi nhuận quý I/2024 của CMG.
Để dự đoán liệu lợi nhuận quý I/2024 của CMG có tăng hay không, chúng ta cần xem xét thêm các yếu tố như:
 1. **Biên động thị trường**: CMG là một công ty viễn thông di động, doanh thu và lợi nhuận của họ thường phụ thuộc vào tình hình kinh tế chung và thị trường viễn thông.
 2. **Sản phẩm và dịch vụ**: CMG có thể đã ra mắt các sản phẩm mới hoặc cải tiến dịch vụ nào đó, điều này có thể ảnh hưởng đến doanh thu và lợi nhuận.
 3. **Chi phí**: Chi phí vận hành, chi phí đầu tư vào mạng lưới di động, chi phí marketing, v.v., cũng có thể ảnh hưởng đến lợi nhuận.
 4. **Số lượng cổ phiếu**: Giao dịch lớn vào ngày 26/03/2024 có thể cho thấy có sự thay đổi trong sở hữu cổ phiếu, nhưng không nhất thiết là doanh thu hoặc lợi nhuận tăng.
 5. **Báo cáo tài chính**: Thông tin chi tiết về doanh thu, chi phí, và lợi nhuận quý I/2024 sẽ được CMG công bố sau khi kết thúc quý, thường là vào giữa tháng 5/2024.
 Vì vậy, để có cái nhìn chính xác hơn về lợi nhuận quý I/2024 của CMG, chúng ta cần chờ đợi báo cáo tài chính chính thức được công bố.

Hình 19: Phản hồi Output

4.3.2. Huấn luyện tinh chỉnh mô hình Qwen:

*Thu thập dữ liệu

Nhận thấy rằng ưu điểm của các mô hình ngôn ngữ lớn và hiểu được “ngôn ngữ tự nhiên”, tuy nhiên dữ liệu mà ban tổ chức cung cấp không ở dạng lý tưởng (**instruction-style conversation**) mà chỉ là các văn bản đơn lẻ không có câu hỏi-trả lời để có thể sử dụng kỹ thuật train SFT (Supervised Fine-Tuning), nên trước hết chúng tôi sẽ dùng LLM để tự động tạo nguồn dữ liệu có **instruction-target**.

- **Bước 1:** Thu thập dữ liệu văn bản (dữ liệu BTC cung cấp, báo cáo tài chính FPT/CMC,...)
- **Bước 2:** Dùng kỹ thuật gọi là **self-instruct** hoặc prompt + LLM để tạo bộ dữ liệu huấn luyện lý tưởng.
- **Bước 3:** Lưu dữ liệu và lọc các dữ liệu gây nhiễu.

* Định Dạng Dữ Liệu Mỗi mục dữ liệu bao gồm ba trường:

- **instruction:** Câu hỏi hoặc nhiệm vụ.
- **input:** Bối cảnh bổ sung cho nhiệm vụ (có thể để trống).
- **output:** Câu trả lời hoặc phản hồi mong đợi.

* Xử lý dữ liệu Bộ dữ liệu được định dạng lại theo cấu trúc prompt-response sử dụng mẫu Alpaca.

Các prompt được thêm ký tự EOS (*End of Sentence*) để tránh sinh văn bản vô hạn trong quá trình suy luận.

* Lựa chọn mô hình và huấn luyện

Mô hình gốc được lựa chọn để tinh chỉnh là `unslloth/Qwen2.5-7B-Instruct`, phiên bản 7 tỷ tham số của mô hình Qwen.

Kỹ Thuật Tối Ưu Hóa

- Mô hình được tải với lượng tử hóa 4-bit (bnb-4bit) để tối ưu hóa bộ nhớ.
- Sử dụng gradient checkpointing (`use_gradient_checkpointing = "unslloth"`) để tối ưu hóa bộ nhớ trong quá trình huấn luyện.

- Quá trình tinh chỉnh sử dụng PEFT (Parameter Efficient Fine-Tuning) với LoRA (Low-Rank Adaptation) áp dụng vào các lớp:
 - Các module mục tiêu: ["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"]
 - LoRA Rank: 16
 - LoRA Alpha: 16

Cấu hình huấn luyện Huấn luyện được thực hiện bằng SFTTrainer từ thư viện trl.

Cấu hình huấn luyện:

- Độ dài tối đa: 2048
- Learning Rate: 2e-4
- Kích thước Batch: 2 (mỗi thiết bị), với Gradient Accumulation Steps là 4.
- Optimizer: AdamW với độ chính xác 8-bit (adamw_8bit)
- FP16/BF16 Mixed Precision tự động tùy thuộc vào GPU.
- Số bước huấn luyện: 100.

Kết quả

- Mô hình được tinh chỉnh hiệu quả mà không gặp lỗi tràn bộ nhớ, nhờ vào lượng tử hóa 4-bit và gradient checkpointing.
- Đánh giá ban đầu cho thấy mô hình trả lời tốt hơn đối với các câu hỏi liên quan đến dữ liệu thị trường chứng khoán Việt Nam.
- Đánh giá mô hình bằng các chỉ số như Training Loss, Validation Loss, Accuracy, BLEU Score, ROUGE-2 Score, và Avg Inference Time.



Hình 20: Đánh giá mô hình RAG

4.3.3. Định hướng phát triển

- Kết hợp mô hình embedding Việt hóa (như BGE-small-vi) để cải thiện chất lượng semantic search.
- Triển khai vector DB dạng API-based như Qdrant Cloud để hỗ trợ mở rộng hệ thống.
- Phân loại nội dung và gán nhãn theo chủ đề: cổ tức, tài chính, nhân sự, pháp lý,... để hỗ trợ lọc thông minh trước khi sinh câu trả lời.

Chương 5: Huấn Luyện Tinh Chỉnh Mô Hình Ngôn Ngữ Lớn

5.1 Thu thập dữ liệu

Nhận thấy rằng ưu điểm của các mô hình ngôn ngữ lớn và hiểu được “ngôn ngữ tự nhiên”, tuy nhiên dữ liệu mà ban tổ chức cung cấp chỉ có từ năm 2023-2024 mà dữ liệu chúng tôi sử dụng từ năm 2017 nên chúng tôi sẽ thu thập lại dữ liệu tin tức.

- Bước 1: Thu thập link các bài báo
 - + Chúng ta sẽ sử dụng thư viện Beautiful soup kết hợp với đường dẫn rss của Google News API với format như sau để tìm kiếm đường link bài báo:

rss_url = f"https://news.google.com/rss/search?q=

+ Ta sẽ thu thập các bài báo với từ khóa tìm kiếm như sau:

- FPT: từ khóa tìm kiếm: FPT, Tài chính, chứng khoán, đầu tư, kinh tế
- CMC: từ khóa tìm kiếm: CMC, Tài chính, chứng khoán, đầu tư, kinh tế

+ Sau đó, chúng ta sẽ cào từ năm 2017 tới năm 2025, kết quả thu thập được: với FPT: 4414 đường links, với CMC: 4134 đường links.

- Bước 2: Cào headline và nội dung bài báo

+ Sau khi thu thập link bài báo, chúng ta sẽ sử dụng thư viện selenium để cào nội dung của từng bài báo một. Vấn đề ở đây là do đường link mà chúng ta thu thập được ở bước 1 là đường link rss nên Selenium không thể đọc được.

+ Vì thế chúng ta sẽ có thêm bước đưa đường link rss về đường link html như sau: bỏ đuôi rss ở cuối đường link, và dùng hàm driver để đưa về link html.

+ Sau đó chúng ta sẽ cào sạch nội dung của bài bằng cách lấy hết mục content của url html của từng đường link.

Bước 3: lọc dữ liệu: Nhận thấy khi cào như này, điểm mạnh là cào được hết nội dung, điểm yếu là nhiều dữ liệu nhiễu do cào cả những header footer, ta sẽ lọc những dòng có $>=10$ từ để đảm bảo đó là nội dung bài báo.

Chi tiết dữ liệu xem ở phần phụ lục.

5.2 Gán nhãn cảm xúc của dữ liệu

- Vector hóa tin tức dữ liệu sử dụng TF-IDF , giữ lại 1000 từ có tần suất xuất hiện cao nhất. Sau đó, sẽ sử dụng thuật toán để tạo ra ma trận thừa số học của dữ liệu tin tức.

- Sau đó ta sử dụng Sentiment Intensity Analyzer kết hợp với từ điển cảm xúc vader lexicon của nltk để đưa ra chỉ số cảm xúc cho từng tin tức.

- Encode chúng như sau:

+ score thuộc $[0,5;1]$ sẽ mã hóa thành 1 - đại diện cho cảm xúc tích cực.

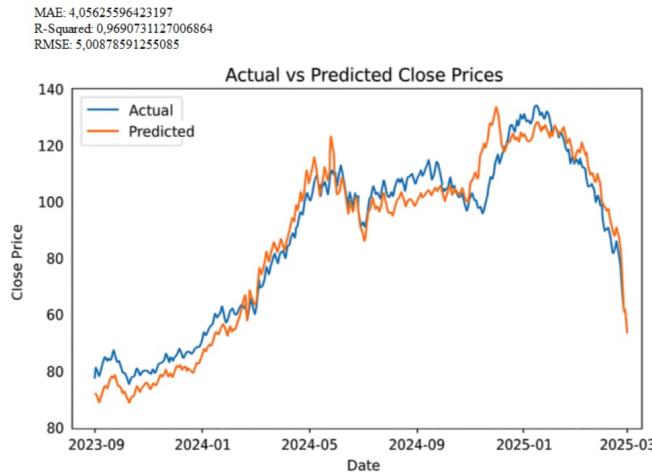
+ score thuộc $(-0,5; 0,5)$ sẽ mã hóa thành 0 - đại diện cho cảm xúc trung lập.

+ score thuộc $(-1,-0,5]$ sẽ mã hóa thành -1 - đại diện cho cảm xúc tiêu cực.

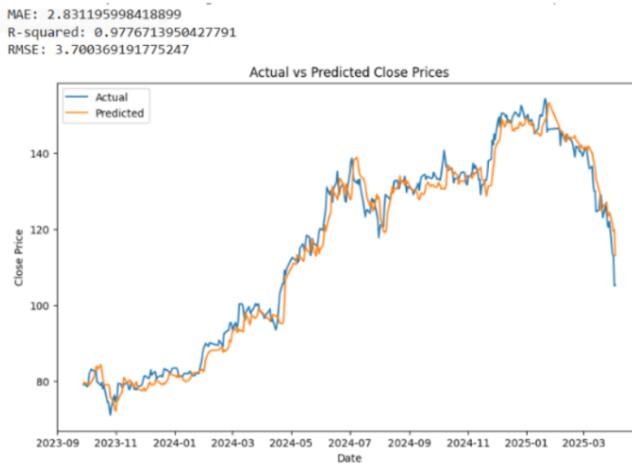
- Để chứng minh cho tính đúng đắn khi xem xét các đặc trưng định tính như tin

tức, ta sẽ cùng đến với thử nghiệm sau: ta sẽ tính sai số của model XGBOOST tinh chỉnh 15 lần thử bằng optuna để làm 2 nhiệm vụ mà là thế mạnh của XGBOOST như sau:

- + Dự đoán giá đóng cửa sau 3 ngày trước khi và sau khi thêm đặc trưng cảm xúc tin tức:



Hình 21: Sai số của mô hình trước khi thêm đặc trưng cảm xúc tin tức



Hình 22: Sai số của mô hình sau khi thêm đặc trưng cảm xúc tin tức

- + Dự đoán giá đóng cửa sau 7 ngày trước khi và sau khi thêm đặc trưng cảm xúc tin tức:



Hình 23: Sai số của mô hình trước khi thêm đặc trưng cảm xúc tin tức



Hình 24: Sai số của mô hình trước khi thêm đặc trưng cảm xúc tin tức

Nhận xét: Có thể thấy sau khi thêm đặc trưng định tính là dữ liệu tin tức, sai số mô hình thu về được giảm đi nhiều, điều đó cho thấy tầm quan trọng của đặc trưng này, cũng như củng cố thêm tính hợp lý khi đưa dữ liệu để tinh chỉnh mô hình ngôn ngữ lớn.

5.3 Xử lý dữ liệu

- Bộ dữ liệu gốc sử dụng: những thông tin về lịch sử cổ phiếu của hai công ty FPT và CMC bao gồm date, giá high, low, open, close, volume. Thời gian của cổ phiếu từ đầu năm 2017 tới 9/4/2025. Ngoài ra ta cũng thêm cột ‘content’ là nội dung bài báo tại thời điểm đó.

- Ta sẽ chuyển dữ liệu từ dạng dataframe sang dạng json chuyên dùng cho LLM, tức dữ liệu sẽ được chuyển hết về ngôn ngữ tự nhiên theo cú pháp sau:
 - + Prompt: "Asset: A (A is FPT or CMC)\nPredict the close price of FPT after 3 days given that.\nRecent Data:\n Date_x High=, Low=, Open=, Close=, Vol=\n.[News is] \n Predict A prices for the next n days:" với x là số từ 1 tới 10; A là FPT hoặc CMC; n là số thuộc {3, 7, 21, 63}
 - + Label: close price after 3 ngày.

5.4 Kiểm thử mô hình

- Trước khi đến với bước tinh chỉnh mô hình LLM, chúng tôi sẽ kiểm thử trước khi huấn luyện mô hình không tinh chỉnh để xem xét tính hợp lý khi sử dụng LLM trong việc dự đoán giá cổ phiếu.
- Mô hình lựa chọn: Llama 7 tỉ tham số.
- Kết quả dự đoán: [Phụ lục](#)
- Đánh giá:

Bảng 16: Dánh giá mô hình dự đoán giá cổ phiếu

Cổ phiếu	Thời gian (ngày)	R2	RMSE	MAE
FPT	3	-2.7371	7.1099	5.9000
CMC	3	-6.0030	4.8085	4.8000
FPT	7	-0.4054	7.8964	5.7857
CMC	7	-0.9949	5.0841	4.5714
FPT	21	-3.7341	19.1336	17.5095
CMC	21	-3.2218	7.6920	7.0190
FPT	63	-7.7472	34.1332	32.2000
CMC	63	-9.4983	12.4898	11.9714

Nhận xét: Sai số của mô hình chấp nhận được với một mô hình dự đoán chưa tinh chỉnh, có thể thấy mô hình rất khó để dự đoán sự chuyển biến của giá đóng với những task dài hạn như 21 hay 63 ngày. Vì thế, có thể nói là hợp lý để có thể sử dụng mô hình ngôn ngữ lớn trong việc dự đoán giá đóng của cổ phiếu.

5.5 Tinh chỉnh mô hình ngôn ngữ lớn

Chúng tôi sẽ lựa chọn các mô hình tiền huấn luyện khác nhau để tinh chỉnh sau đó sẽ đánh giá và chọn ra mô hình hợp lý nhất xét trên phương diện cân bằng được tính chính xác và chi phí huấn luyện. Ở đây chúng tôi sẽ tinh chỉnh với nhiệm vụ là dự đoán giá đóng cửa 3 ngày sắp tới của công ty FPT và CMC. Trước tiên, ta sẽ ưu tiên thử nghiệm cả năm mô hình dưới đây với dữ liệu của công ty CMC với nhiệm vụ dự đoán giá đóng cửa 3 ngày:

5.5.1. Phi-3 3,8B, Llama guard 8B, Llama 3 8B Instruct, Llama 3 13B hf

- Chính lại định dạng dữ liệu theo mẫu sau để tokenize hiệu quả hơn: "<s>[INST] prompt [/INST] label </s>"
- Loại bỏ kí tự /n không cần thiết.
- Độ dài tối đa: 512 ký tự và padding token ở bên phải để toàn bộ đều cùng độ dài là 512
- Quantization: mục tiêu là giảm bộ nhớ khi tải mô hình, ta sẽ tải mô hình với trọng số 4-bit, quantization 4-bit phi tuyến, loại số được sử dụng là float 16 và nén trong hai giai đoạn: nén xuống 8 bit và nén từ 8 bit xuống 4 bit. Lợi ích của chúng là giảm rất nhiều footprint bộ nhớ mà vẫn được hiệu năng và độ chính xác tốt và ổn định.
- Tải mô hình: Dạng casual language model (sinh văn bản), và tự động chọn class đúng và tokenizer theo tên mô hình. Đặc biệt là với các model như LLaMa thì cần đặt tham số cấu hình song song bằng 1, tức là không chia tensor
- Xây dựng cấu hình cho LoRA: hạng ma trận low-rank LORA là 16, chỉ học các ma trận nhỏ, hệ số scaling là 32. Các layer thì đủ các layer trong Transformers (attention và feed-forward). Ngoài ra để tránh over fitting thì dropout 5%. - Tham số khởi tạo : batch size=2, tích lũy 4 steps gradients, learning_rate= 2e-4, gradient clipping norm = 0,3 , 3% bước đầu dùng learning rate nhỏ và không thay đổi loại LR theo thời gian

5.5.2. Gemma 2 2,1B phi

- Format data để fintune mô hình này có một chút khác những mô hình trên:

"<start_of_turn>usernexample['prompt']<end_of_turn>n<start_of_turn>modelnexample['label']"

- Do mô hình này có số lượng tham số nhỏ nên sẽ không quantization để bảo toàn độ chính xác.

Toàn bộ dự đoán sử dụng tinh chỉnh ngôn ngữ lớn xem tại phần phụ lục [Phụ lục](#)

5.6 Đánh giá mô hình

5.6.1. Đánh giá sai số dự đoán

Bảng 17: Đánh giá hiệu suất các mô hình dự đoán giá cổ phiếu

Mô hình	Cổ phiếu	Kỳ (ngày)	R2	RMSE	MAE
Microsoft Phi 3 mini 4k instruct (3.8B)	CMC	3	-1.4659	2.8534	2.2000
		7	-87771.7552	1066.4352	575.7143
		21	-214612.5693	1734.2665	1516.9286
		63	-243406.8515	1901.7982	1826.2746
LLama 2 (7B)	CMC	3	-1.4659	2.8534	2.2000
		7	-2.7784	6.9969	6.0000
		21	-7.4933	10.9100	10.2476
		63	-14.1124	14.9852	14.4810
Gemma (2.1B phi)	CMC	3	-6.0030	4.8085	4.8000
		7	-0.9949	5.0841	4.5714
		21	-2.7546	7.2539	6.2952
		63	-7.1950	11.0350	10.1952
LLama 2 (13B hf)	CMC	3	-1.4659	2.8534	2.2000
		7	-2.7036	6.9274	5.9529
		21	-6.9391	10.5481	9.8400
LLama 3 instruct (8B)	CMC	3	-0.2804	2.0561	1.6833
		7	-973.3301	7.4668	7.4667
		21	-1047.5063	11.7496	11.7333
		63	-1272.4376	17.9809	17.9667

5.6.2. Đánh giá trên các khía cạnh khác

Bảng 18: Đánh giá so sánh các mô hình ngôn ngữ lớn (Tóm tắt)

Tiêu chí	Phi-3 Mini (3.8B Instr.)	Llama 2 7B (Chat/Instr.)	Llama 3 8B (Instruct)	Llama 3 13B (Custom FT)	Gemma 2B (Instruct)
Nhà phát triển	Microsoft	Meta	Meta	Meta (Fine-tuned)	Google
Kích thước	3.8B	7B	8B	13B	2B (hoặc 9B)
Mục đích chính	Da nhiệm, hiệu quả, thiết bị biên	Da nhiệm, chatbot, sinh văn bản	Trợ lý ảo, chatbot, sinh văn bản (hướng dẫn)	Da nhiệm, hiệu quả cao hơn bản 8B	Da nhiệm, hiệu quả, thiết bị biên/di động
MMLU (5-shot, cao hơn là tốt)	66.1	64.3	66.4	Ước tính > Llama 3 8B	52.89
HellaSwag (0/8-shot)	63.6 (0-shot, CoT)	23.3 (0-shot) / 56.8 (8-shot)	80.6 (6-shot)	Ước tính > Llama 3 8B	24.72 (5-shot)
HumanEval (0-shot)	57.9	12.8	60.4 / 72.6	Ước tính > Llama 3 8B	20.12
Hiệu năng	Rất nhẹ: 4-8GB	Trung bình: 8-16GB	Trung bình: 8-16GB	Cao: 16-24GB	Rất nhẹ: 4-8GB
	Rất nhanh	Nhanh	Nhanh	Trung bình (chậm hơn 8B)	Rất nhanh
Dộ tự nhiên	Rất tốt/kích thước, mượt, lý luận	Tốt, tự nhiên. Có thể lặp lại/kém sáng tạo (phức tạp)	Rất cao, tự nhiên, linh hoạt, hiểu ngữ cảnh	Rất cao, sâu sắc hơn bản 8B	Tốt/kích thước, mạch lạc. Kém hơn mô hình lớn (sáng tạo)
Khả năng chuyên biệt	Tốt: tóm tắt, Q&A, email, code cơ bản	Tốt: hội thoại cơ bản, tóm tắt, Q&A	Tuyệt vời: hội thoại, code, sáng tạo, tóm tắt, lý luận	Xuất sắc: tác vụ phức tạp, chi tiết cao	Tốt: tóm tắt, Q&A, code đơn giản
Tối ưu ngôn ngữ	Chủ yếu Tiếng Anh	Đa ngôn ngữ (Anh chính)	Đa ngôn ngữ (Anh chính)	Đa ngôn ngữ (Anh chính)	Chủ yếu Tiếng Anh, hỗ trợ đa ngôn ngữ cơ bản
Giấy phép	MIT	Llama 2 Community (TM có giới hạn)	Llama 3 Community (TM)	Llama 3 Community (TM)	Gemma Terms of Use (khá tự do)

5.7 Kết luận

Sau khi cân nhắc kỹ lưỡng về độ chính xác, số lượng tham số cùng các khía cạnh khác, chúng tôi đánh giá **LLaMA 3 8B** là hợp lý nhất. Dưới đây là sai số khi tinh chỉnh mô hình **LLaMA 3 8B**:

- FPT (3 ngày):

- $R^2 = -2.7371$
- $RMSE = 7.1099$
- $MAE = 5.9000$

- CMC (3 ngày):

- $R^2 = -0.2804$
- $RMSE = 2.0561$
- $MAE = 1.6833$

Kết quả dự đoán:

- FPT: [122.0, 113.5, 105.6]
- CMC: [30.5, 31.05, 31.05]

Chương 6: Hồi Kết

6.1 Đánh giá

Trong quá trình nghiên cứu, chúng tôi đã triển khai và đánh giá một chuỗi các mô hình học sâu dự báo giá cổ phiếu, bao gồm LSTM, XGBOOSTS trên các khoảng dự báo (horizon) 3, 7, 21 và 63 ngày, sau đó truy xuất thông tin tài chính và sử dụng ngôn ngữ lớn trong nhiệm vụ dự đoán giá cổ phiếu. Cụ thể, quy trình tổng quan gồm các bước chính sau:

6.1.1. Xử lý dữ liệu

Dữ liệu lịch sử được làm sạch, tạo đặc trưng (feature engineering), chia nhỏ (train-test split) và chuẩn hóa. Các mô hình (LSTM, XGBOOST) được huấn luyện và tinh chỉnh, sau đó dự báo giá cho từng cổ phiếu ở các mốc thời gian tương lai khác nhau (3, 7, 21 và 63 ngày).

6.1.2. Sử dụng học máy, học sâu dự đoán giá đóng cửa ngắn, trung và dài hạn

Sai số tạm chấp nhận được, nhưng vẫn cần cải thiện trong tương lai. Kết quả thực nghiệm cho thấy việc kết hợp các mô hình dự báo chuỗi thời gian hiện đại cùng phương pháp truyền thống (LSTM, XGBoost) với quy trình đánh giá mức độ đầu tư sẽ giúp nâng cao hiệu quả đầu tư, hạn chế rủi ro, cũng như mở ra hướng nghiên cứu và phát triển các chiến lược đầu tư tự động trong tương lai, đặc biệt là bài viết cũng chỉ ra được điểm mạnh của XGBoost trong dự đoán ngắn hạn và LSTM trong dự đoán trung hạn và dài hạn.

6.1.3. RAG truy vết thông tin cổ phiếu và tài chính

Chúng tôi thấy rằng lời văn sinh ra khá tự nhiên và đầy đủ nội dung muốn tìm kiếm. Chúng tôi trong tương lai sẽ cố gắng xây dựng thêm câu hỏi ở cuối phần trả lời để một phần đánh giá mô hình chủ quan bằng reinforcement learning và một phần tạo một cuộc hội thoại đa chiều.

6.1.4. Tinh chỉnh mô hình ngôn ngữ lớn

Bài viết cũng đã chỉ ra được tác dụng của mô hình ngôn ngữ lớn trong việc dự đoán giá cổ phiếu tài chính. Tuy nhiên, thực nghiệm chỉ ra rằng với các dự đoán trung và dài hạn thì các mô hình ngôn ngữ nhỏ có vẻ sẽ không chỉ ra được sự biến động của giá cổ phiếu. Ngoài ra, trong tương lai, chúng tôi sẽ dự định phát triển NLP Reasoning kết hợp với RAG tự động để đưa ra giải thích cho các nhà đầu tư lý do thay đổi, mặc dù hơi chủ quan và có thể dẫn tới ảo giác.

6.2 Hướng phát triển trong tương lai

6.2.1. Nâng cấp mô hình dự báo

- Kết hợp dữ liệu phi cấu trúc (tin tức, sentiment từ mạng xã hội) bằng NLP để bổ sung ngữ cảnh.
- Thử nghiệm và đánh giá dài hạn hơn: 90 ngày, 180 ngày,...

6.2.2. Cải tiến RAG

- Xây dựng cơ chế đánh giá tự động bằng Reinforcement Learning với human-in-the-loop (phản hồi từ nhà đầu tư).
- Mở rộng cơ sở dữ liệu sang nguồn đa ngôn ngữ và dữ liệu vĩ mô (lãi suất, GDP).

6.2.3. Tối ưu LLM

- Fine-tuning LLM trên dữ liệu tài chính chuyên sâu hơn để giảm ảo giác.
- Phát triển NLP Reasoning tự động giải thích biến động giá dựa trên mối quan hệ nhân quả (ví dụ: "Cổ phiếu A giảm do báo cáo lợi nhuận quý thấp hơn kỳ vọng").

6.2.4. Hệ thống đầu tư tự động thông minh

- Xây dựng pipeline end-to-end kết hợp dự báo giá, quản lý rủi ro (Value-at-Risk), và tối ưu danh mục.
- Triển khai ứng dụng thực tế với giao diện hỏi-đáp tự nhiên, hỗ trợ ra quyết định cho nhà đầu tư cá nhân.

6.2.5. Đánh giá khách quan hơn

- Benchmark đa tiêu chí: So sánh với mô hình baseline (ARIMA, Prophet) và chỉ số thị trường.

- Kiểm tra backtesting trên nhiều giai đoạn thị trường (bull/bear market) để đánh giá tính ổn định.

Phụ Lục

Dữ liệu huấn luyện cho dự đoán

Bảng 19: Bảng mô tả dữ liệu

Chỉ số	Tên tiếng Việt	Mô tả	Công thức
open	Giá mở cửa	Giá tại thời điểm bắt đầu phiên giao dịch	thu thập
close	Giá đóng cửa	Giá tại thời điểm kết thúc phiên giao dịch	thu thập
high	Giá cao nhất	Mức giá cao nhất trong phiên giao dịch	thu thập
low	Giá thấp nhất	Mức giá thấp nhất trong phiên giao dịch	thu thập
volume	Khối lượng giao dịch	Tổng số cổ phiếu/hợp đồng được giao dịch trong phiên	
{x}return	Tỷ suất lợi nhuận theo thời gian	Tỷ lệ thay đổi giá theo thời gian	$\frac{G_i(x)}{G_i(x-n)} - 1$
SMA _n	Trung bình cộng n ngày gần nhất	Trung bình cộng giá đóng cửa của n ngày gần nhất	$\frac{\sum_{i=0}^{n-1} G_{close}(t-i)}{n}$
RSI _n	Chỉ số sức mạnh tương đối	Tính mức tăng và giảm trung bình theo n phiên để đo lường động lượng giá	$100 - \frac{100}{1 + \frac{\text{trung bình tăng}}{\text{trung bình giảm}}}$
EMA _n	Dường trung bình động hàm mũ	Trung bình có trọng số, các giá gần hơn có trọng số lớn hơn	Giá hôm nay $\times \frac{2}{n+1} + EMA_{\text{hôm qua}} \times \left(1 - \frac{2}{n+1}\right)$
Bollinger Bands	Dải Bollinger	Ba đường bao quanh giá, đo mức biến động	$SMA \pm 2 \times \text{Độ lệch chuẩn}$
OBV _n	Chỉ báo khối lượng cân bằng	Cộng hoặc trừ khối lượng giao dịch dựa trên thay đổi giá	$OBV_{\text{hôm qua}} + \text{khối lượng hôm nay (nếu giá tăng)}$ $OBV_{\text{hôm qua}} - \text{khối lượng hôm nay (nếu giá giảm)}$
ATR _n	Biến động trung bình thực	Do lường mức độ biến động giá	$\max(high - low , high - close_{\text{hôm qua}} , low - close_{\text{hôm qua}})$
MACD	Dường trung bình động hội tụ phân kỳ	Hiệu số giữa EMA 12 kỳ và EMA 26 kỳ, tín hiệu mua bán	$EMA_{12} - EMA_{26}$
MFI	Chỉ báo dòng tiền	Do lường áp lực mua và bán dựa trên giá và khối lượng	$100 - \frac{100}{1 + \text{Tỉ lệ dòng tiền}}$
Momentum _n	Động lượng n kỳ	Do lường tốc độ thay đổi giá trong n kỳ	$\frac{G(t)}{G(t-n)} - 1$
CCI _n	Chỉ báo kênh hàng hóa	Do lường độ lệch của giá so với mức trung bình thống kê	$\frac{\text{Giá diển hình} - SMA_n}{0.015 \times MAD_n}$

Chỉ số	Tên tiếng Việt	Mô tả	Công thức
Williams %R _n	Chỉ báo Williams %R	Do lường mức độ quá mua hoặc quá bán trong n kỳ	$\frac{High_n - Close}{High_n - Low_n} \times (-100)$
ADX _n	Chỉ báo sức mạnh xu hướng	Do lường sức mạnh của xu hướng trong n kỳ	Trung bình động của các giá trị DX trong n kỳ
TRIX _n	Chỉ báo TRIX	Tốc độ thay đổi phần trăm của EMA ba lần làm mịn	$\frac{EMA_3(t) - EMA_3(t-1)}{EMA_3(t-1)}$
TSF _n	Dự báo tuyến tính	Dự báo giá dựa trên đường hồi quy tuyến tính trong n kỳ	Giá trị dự báo từ đường hồi quy tuyến tính
Volatility Day	Biến động trong ngày	Phần trăm thay đổi giá đóng cửa trong một ngày	Phần trăm thay đổi giá đóng cửa
Volatility Week	Biến động trong tuần	Độ lệch chuẩn của phần trăm thay đổi giá đóng cửa trong 7 ngày	Độ lệch chuẩn của phần trăm thay đổi giá đóng cửa (7 ngày)
Volatility Month	Biến động trong tháng	Độ lệch chuẩn của phần trăm thay đổi giá đóng cửa trong 30 ngày	Độ lệch chuẩn của phần trăm thay đổi giá đóng cửa (30 ngày)
High-Close	Chênh lệch giá cao - đóng cửa	Chênh lệch giữa giá cao nhất và giá đóng cửa trong ngày	High - Close
Low-Open	Chênh lệch giá thấp - mở cửa	Chênh lệch giữa giá thấp nhất và giá mở cửa trong ngày	Low - Open
Cumulative Return	Lợi nhuận tích lũy	Tổng lợi nhuận từ đầu kỳ đến hiện tại	Tích lũy của $(1 + \text{phần trăm thay đổi giá đóng cửa})$
SMA _N	Trung bình động đơn giản N ngày	Trung bình động giá đóng cửa của N ngày gần nhất	$\sum_{i=0}^{N-1} \frac{G_{close}(t-i)}{N}$
Return Std	Độ lệch chuẩn lợi nhuận ngày	Độ lệch chuẩn của phần trăm thay đổi giá đóng cửa trong 1 ngày	Độ lệch chuẩn của phần trăm thay đổi giá đóng cửa (1 ngày)
Return Week Std	Độ lệch chuẩn lợi nhuận tuần	Độ lệch chuẩn của phần trăm thay đổi giá đóng cửa trong 7 ngày	Độ lệch chuẩn của phần trăm thay đổi giá đóng cửa (7 ngày)
Return Month Std	Độ lệch chuẩn lợi nhuận tháng	Độ lệch chuẩn của phần trăm thay đổi giá đóng cửa trong 30 ngày	Độ lệch chuẩn của phần trăm thay đổi giá đóng cửa (30 ngày)
Liquidity Day	Thanh khoản trong ngày	Giá trị giao dịch trung bình trong một ngày	Trung bình của $(\text{Khối lượng giao dịch} \times \text{Giá đóng cửa})$ trong 1 ngày
Liquidity Week	Thanh khoản trong tuần	Giá trị giao dịch trung bình trong một tuần	Trung bình của $(\text{Khối lượng giao dịch} \times \text{Giá đóng cửa})$ trong 7 ngày
Liquidity Month	Thanh khoản trong tháng	Giá trị giao dịch trung bình trong một tháng	Trung bình của $(\text{Khối lượng giao dịch} \times \text{Giá đóng cửa})$ trong 30 ngày

Chỉ số	Tên tiếng Việt	Mô tả
year_revenue_growth	Tăng trưởng doanh thu năm	Mức tăng trưởng doanh thu so với cùng kỳ năm trước
quarter_revenue_growth	Tăng trưởng doanh thu quý	Tăng trưởng so với quý liền kề hoặc quý cùng kỳ
year_operation_profit	Tăng trưởng lợi nhuận hoạt động năm	So sánh lợi nhuận hoạt động hiện tại và năm trước
quarter_operation_profit	Tăng trưởng lợi nhuận hoạt động quý	So sánh theo quý gần nhất
year_share_holder_including	Tăng trưởng thu nhập cổ đông	Biến động thu nhập thuộc về cổ đông theo năm
quarter_share_holder	Tăng trưởng thu nhập cổ đông	Biến động thu nhập thuộc về cổ đông theo quý
price_to_earning	P/E (Hệ số giá trên thu nhập)	Định giá theo lợi nhuận (giá / thu nhập mỗi cổ phiếu)
price_to_book	P/B (Hệ số giá trên giá trị sổ sách)	Định giá theo giá trị sổ sách (giá / giá trị sổ sách)
roe	ROE (Lợi nhuận trên vốn chủ sở hữu)	Hiệu quả sử dụng vốn chủ sở hữu
roa	ROA (Lợi nhuận trên tài sản)	Khả năng tạo lợi nhuận từ tài sản
equity_on_total_asset	Vốn chủ / Tổng tài sản	Cấu trúc vốn doanh nghiệp
equity_on_liability	Vốn chủ / Nợ phải trả	Cân đối giữa vốn chủ sở hữu và nợ
eps_growth	Biến động EPS	Tốc độ tăng trưởng thu nhập mỗi cổ phiếu sau thời gian
asset_on_equity	Tài sản / Vốn chủ sở hữu	Hệ số đòn bẩy tài chính
payable_on_equity	Nợ phải trả / Vốn chủ sở hữu	Phản ánh áp lực tài chính của doanh nghiệp
book_value_per_share	Biến động giá trị sổ sách mỗi cổ phiếu	Biến động giá trị sổ sách ròng chia theo cổ phần
invest_k_VND	Chi phí đầu tư	Chi phí doanh nghiệp bỏ ra cho đầu tư
from_invest_k_VND	Luồng tiền từ hoạt động đầu tư	Tiền thu vào từ đầu tư tài sản hoặc cùng kỳ khác
from_financial_k_VND	Luồng tiền từ tài chính	Luồng tiền quan đến vay nợ, phát hành cổ phiếu
from_sale_k_VND	Luồng tiền từ bán hàng	Doanh thu tiền mặt từ hoạt động kinh doanh chính
free_cash_flow_k_VND	Dòng tiền tự do	Tiền còn lại sau đầu tư để trả nợ hoặc chia cổ tức
revenue_k_VND	Doanh thu thuần	Tổng doanh thu từ hoạt động bán hàng
operation_expense_k_VND	Chi phí hoạt động	Tổng chi phí phát sinh trong kỳ kinh doanh
operation_profit_k_VND	Lợi nhuận hoạt động	Lợi nhuận trước thuế và lãi vay
pre_tax_profit_k_VND	Lợi nhuận trước thuế	Lợi nhuận trước khi trừ thuế thu nhập doanh nghiệp
post_tax_profit_k_VND	Lợi nhuận sau thuế	Lợi nhuận ròng sau khi nộp thuế thu nhập doanh nghiệp
share_holder_income_k_VND	Thu nhập cổ đông	Thu nhập cuối cùng thuộc về cổ đông cùng kỳ
earning_per_share_k_VND	EPS (Lợi nhuận trên mỗi cổ phiếu)	Thu nhập trên một đơn vị cổ phần phổ thông
book_value_per_share_k_VND	Giá trị sổ sách/cổ phần	Tổng tài sản ròng chia cho số cổ phần

I. Kiểm chứng giả thuyết

Bảng 20: Bảng kiểm chứng giả thuyết

STT	Giả thuyết thống kê	Lý do chọn	Kiểm chứng	Lý giải kiểm chứng
1	Biến động giá cổ phiếu tương quan dương với khối lượng giao dịch hàng ngày.	Dựa trên lý thuyết "Khối lượng đi trước giá" (Volume precedes price) - Volume tăng thường phản ánh sự quan tâm của thị trường.	Bắc bỏ	Khối lượng giao dịch cao có thể đến từ cả bên mua và bán → Không đủ để kết luận giá tăng.
2	Khối lượng giao dịch tăng bất thường đi kèm với mức độ biến động giá cao hơn.	Áp dụng lý thuyết kỳ vọng phân kỳ - Volume đột biến phản ánh sự tranh chấp mạnh giữa phe mua/bán.	Thừa nhận	Phù hợp với thực tế: Volume tăng đột biến thường đi kèm biến động giá mạnh (ví dụ: tin tốt/xấu).
3	Mức độ biến động giá cổ phiếu ảnh hưởng đến khối lượng giao dịch của nhà đầu tư.	Kiểm tra hiệu ứng tâm lý - Nhà đầu tư có xu hướng giao dịch nhiều hơn khi giá biến động mạnh.	Bắc bỏ	Biến động giá không trực tiếp thúc đẩy volume → Volume phụ thuộc vào yếu tố khác (tin tức, thanh khoản).
4	Giá cổ phiếu có xu hướng giảm mạnh khi khối lượng giao dịch tăng đột biến.	Dựa trên lý thuyết phân phối (Distribution Theory) - Volume cao có thể báo hiệu "bán tháo".	Bắc bỏ	Volume tăng đột biến không luôn dẫn đến giá giảm (có thể do tích lũy trước đợt tăng).
5	Sự thay đổi khối lượng giao dịch dự báo biến động giá cổ phiếu.	Kiểm tra tính dự báo của volume - Nếu volume tăng trước giá, có thể dùng làm tín hiệu.	Bắc bỏ	Volume chỉ phản ánh hiện tại, không đủ để dự báo tương lai.
6	Khối lượng giao dịch ngành khác có tương quan với giá cổ phiếu ngành công nghệ.	Áp dụng lý thuyết luân chuyển dòng tiền - Nhà đầu tư dịch chuyển vốn giữa các ngành.	Thừa nhận	Ví dụ: Dòng tiền từ ngân hàng chảy vào công nghệ khi lãi suất giảm.
7	Tỷ lệ thanh khoản (biến động + khối lượng) ảnh hưởng đến giá.	Dựa trên lý thuyết thanh khoản (Liquidity Theory) - Thanh khoản cao giảm rủi ro, thu hút nhà đầu tư.	Thừa nhận	Cổ phiếu thanh khoản cao thường có biến động giá thấp hơn.
8	Các phiên có thanh khoản cao (khối lượng lớn) thường có mức tăng/giảm giá mạnh hơn trung bình.	Liên quan đến hiệu ứng khuếch đại thanh khoản - Volume lớn khuếch đại xu hướng giá.	Thừa nhận	Phiên volume cao thường đi kèm tin quan trọng (báo cáo tài chính, sự kiện).
9	Biến động giá (volatility) tương quan ngược với giá.	Áp dụng lý thuyết điều chỉnh kỹ thuật - Giá tăng quá nhanh thường kéo theo điều chỉnh.	Thừa nhận	Volatility cao thường xuất hiện ở đỉnh/dáy ngắn hạn.

STT	Giả thuyết thống kê	Lý do chọn	Kiểm chứng	Lý giải kiểm chứng
10	Rủi ro hệ thống (Beta với VNINDEX) ảnh hưởng đến giá.	Kiểm tra mô hình CAPM - Beta đo lường rủi ro thị trường.	Không đủ bằng chứng	Cổ phiếu công nghệ ít phụ thuộc vào VNINDEX (có yếu tố riêng), chúng thường chịu tác động mạnh từ yếu tố ngành hơn là thị trường chung.
11	Biến động cao từ ngành khác ảnh hưởng đến giá công nghệ.	Dựa trên hiệu ứng lan tỏa rủi ro (Contagion Effect) - Rủi ro ngành này lan sang ngành khác.	Thừa nhận	Ví dụ: Biến động ngành năng lượng ảnh hưởng đến tâm lý toàn thị trường.
12	P/E tương quan với giá cổ phiếu quý sau.	Áp dụng lý thuyết định giá cổ phiếu - P/E phản ánh kỳ vọng tăng trưởng.	Thừa nhận	P/E cao thường đi kèm giá tăng trong dài hạn (nếu doanh thu đạt kỳ vọng).
13	P/B tương quan với giá cổ phiếu quý sau.	Dựa trên lý thuyết giá trị nội tại - P/B thấp có thể báo hiệu cổ phiếu bị định giá thấp.	Thừa nhận	P/B < 1 thường hấp dẫn nhà đầu tư giá trị.

Dữ liệu tin tức

Hình 25: Bảng mô tả dữ liệu tin tức

SIT	Title	Date	Content
0	Toyota Altis 2017 - thê-thao-hon-thêm-công-nghệ - Bảo Khanh Hòa	2016-06-10 7:00	<p>Toyota Altis 2017 - thê-thao-hon-thêm-công-nghệ - Bảo Khanh Hòa điện tử</p> <p>Toyota Altis 2017 - thê-thao-hon-thêm-công-nghệ</p> <p>Kỹ niêm 370 năm xây dựng và phát triển tỉnh Khánh Hòa</p> <p>Kỹ niêm 370 năm xây dựng và phát triển tỉnh Khánh Hòa</p> <p>Toyota Altis 2017 - thê-thao-hon-thêm-công-nghệ</p> <p>Mẫu sedan cỡ C tinh chỉnh thiết kế ngoại thất và nội thất theo hướng sắc nét hơn, thêm các công nghệ an toàn hỗ trợ lái.</p> <p>Mẫu sedan cỡ C tinh chỉnh thiết kế ngoại thất và nội thất theo hướng sắc nét hơn, thêm các công nghệ an toàn hỗ trợ lái.</p> <p>Thị trường Thủ Nhĩ Kỳ sẽ là nơi đầu tiên Toyota giới thiệu Corolla (Altis) 2017 vào cuối tháng 6. Hiện hàng xe Nhật chỉ mới tiết lộ một số hình ảnh và thông tin chính thức về mẫu sedan cỡ C phiên bản mới.</p> <p>Altis 2017 nhận những nâng cấp ngoại hình, đặc biệt phần đầu. Cụm đèn pha vuốt sắc nét với công nghệ LED, đèn ban ngày cũng dạng LED. Lưới tản nhiệt trên chỉ còn một thanh ngang thay cho loại hai thanh như trước, lưới tản nhiệt được cải tạo để giảm sức nóng của khung gầm và đèn sương mù tạo cảm giác mạnh.</p> <p>Hiện tại, Nhật chưa trang bị động cơ cho tay lái trợ lực điện, chỉ có hệ thống trợ lực cơ 2 chế độ.</p> <p>Nội thất xe có bọc da, ghế ngồi có khả năng tựa lưng và tựa hông, có khả năng điều chỉnh độ cao và độ nghiêng.</p> <p>Hàng ghế trước có khả năng tựa lưng và tựa hông, có khả năng tựa lưng và tựa hông, có khả năng tựa lưng và tựa hông, có khả năng tựa lưng và tựa hông.</p> <p>Altis mới sẽ nhận nhiều công nghệ an toàn như Cảnh báo va chạm PCS (pre-collision system), cảnh báo chệch làn, đèn pha tự động.</p> <p>Động cơ đường nhám không thay đổi so với trước. Vẫn là tùy chọn 1.8 lit 2ZR-FE hoặc 2 lit 2ZR-FE. Hộp số CVT với 7 cấp số.</p> <p>Sau Thủ Nhĩ Kỳ, có thể Corolla Altis 2017 sẽ tồn tại các thị trường Australia, châu Âu và Bắc Mỹ. Thị trường ASEAN sẽ phải đợi. Toyota chưa tiết lộ mức giá Altis mới.</p>
1	Bảo tàng Công nghệ vũ trụ Việt Nam mở cửa vào năm 2017 - Bảo Khanh Hòa	2016-09-07 7:00	<p>Bảo tàng Công nghệ vũ trụ Việt Nam mở cửa vào năm 2017 - Bảo Khanh Hòa điện tử</p> <p>Bảo tàng Công nghệ vũ trụ Việt Nam mở cửa vào năm 2017</p> <p>Kỹ niêm 370 năm xây dựng và phát triển tỉnh Khánh Hòa</p> <p>Kỹ niêm 370 năm xây dựng và phát triển tỉnh Khánh Hòa</p> <p>Bảo tàng Công nghệ vũ trụ Việt Nam mở cửa vào năm 2017</p> <p>Bảo tàng Công nghệ vũ trụ Việt Nam đặt trong khuôn viên của Trung tâm Công nghệ vũ trụ đặt tại Khu công nghệ cao Hòa Lạc.</p> <p>Bảo tàng Công nghệ vũ trụ Việt Nam đặt trong khuôn viên của Trung tâm Công nghệ vũ trụ đặt tại Khu công nghệ cao Hòa Lạc.</p> <p>Viet Nam sẽ có bảo tàng vũ trụ vào năm 2017. Ánh minh họa</p> <p>Bảo tàng Công nghệ vũ trụ Việt Nam là một hợp phần trong dự án Trung tâm Vũ trụ Việt Nam, được xây dựng tại khu công nghệ cao Hòa Lạc, với tổng diện tích trong nhà 1.675m².</p> <p>Tổng vốn đầu tư công trình khoảng 150 tỷ đến 200 tỷ đồng. Đáng chú ý, chi rieng phần thiết kế không gian trưng bày và các hoạt động tương tác bên trong bảo tàng đã có chi phí lên tới 44 tỷ đồng.</p> <p>Bảo tàng Công nghệ vũ trụ Việt Nam không chỉ trưng bày hiện vật, mà còn có nhiều loại hình tương tác, dẫn dắt người xem từ khám phá thông qua phương tiện truyền thông, dài quan sát thiên văn cũng như khu vực dành cho khách thăm quan sẽ trực tiếp các kỹ sư làm công việc điều khiển vệ tinh.</p> <p>Đây sẽ là điểm du lịch mới nhất của thành phố Hồ Chí Minh, thu hút du khách đến từ ngoài vịnh biển, vốn còn mới mẻ tại Việt Nam.</p> <p>Đợt đầu, cuối năm 2017, Bảo tàng Công nghệ vũ trụ Việt Nam sẽ mở cửa đón khách thăm quan.</p> <p>Nhiệm vụ đón đầu khoa học ứng dụng công nghệ "sóng trọng tài" để xác định cá biển tại Khánh Hòa</p> <p>Giải thưởng Sáng tạo Khoa học và Công nghệ Việt Nam năm 2025 nhận hồ sơ trước ngày 10-10</p> <p>Ra mắt Meta AI - Trợ lý AI hàng ngày tại Việt Nam</p> <p>Cập nhật 150.000 lượt tìm kiếm trên Google mỗi gấp đôi 1 lần</p> <p>Các nhà khoa học Nga tìm ra giải pháp cho cuộc khủng hoảng pin lithium</p> <p>VĂN ĐỀ HỘM NAY: Trò chơi Sa... 50 năm từ hòn là thành động non nhanh đầu Tà mực (79.4-1075 - 79.4-1075)</p>

Dữ liệu huấn luyện cho ngôn ngữ lớn

Hình 26: Bảng mô tả dữ liệu tin tức FPT

"text": "Asset: FPT\nPredict the close price of FPT after 3 days given that.\nRecent Data:\n 2024-02-23 00:00:00: High=91.51, Low=88.59, Open=90.48, Close=89.11, Vol=95.9700"
,
"text": "Asset: FPT\nPredict the close price of FPT after 3 days given that.\nRecent Data:\n 2024-02-26 00:00:00: High=92.63, Low=88.77, Open=89.11, Close=92.63, Vol=100.3400"
,
"text": "Asset: FPT\nPredict the close price of FPT after 3 days given that.\nRecent Data:\n 2024-02-27 00:00:00: High=93.91, Low=92.11, Open=93.05, Close=92.54, Vol=100.3400"
,
"text": "Asset: FPT\nPredict the close price of FPT after 3 days given that.\nRecent Data:\n 2024-02-28 00:00:00: High=93.05, Low=91.17, Open=92.63, Close=93.05, Vol=99.4900"
,
"text": "Asset: FPT\nPredict the close price of FPT after 3 days given that.\nRecent Data:\n 2024-02-29 00:00:00: High=93.83, Low=92.37, Open=92.80, Close=93.48, Vol=96.9100"
,
"text": "Asset: FPT\nPredict the close price of FPT after 3 days given that.\nRecent Data:\n 2024-03-01 00:00:00: High=95.54, Low=93.57, Open=93.65, Close=95.03, Vol=96.0600"
,
"text": "Asset: FPT\nPredict the close price of FPT after 3 days given that.\nRecent Data:\n 2024-03-04 00:00:00: High=95.88, Low=95.03, Open=95.20, Close=95.54, Vol=98.2000"
,
"text": "Asset: FPT\nPredict the close price of FPT after 3 days given that.\nRecent Data:\n 2024-03-05 00:00:00: High=95.54, Low=94.25, Open=95.54, Close=94.94, Vol=99.6600"
,
"text": "Asset: FPT\nPredict the close price of FPT after 3 days given that.\nRecent Data:\n 2024-03-06 00:00:00: High=94.68, Low=92.97, Open=94.34, Close=93.57, Vol=98.5400"

Hình 27: Bảng mô tả dữ liệu tin tức CMC

```
{
  "text": "Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-24: High=14.65, Low=14.47, Open=14.53, Close=14.62, Vol=43382
  \"prompt\": \"Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-24: High=14.65, Low=14.47, Open=14.53, Close=14.62, Vol=433
  \"label\": \"14.7400\""
},
{
  "text": "Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-25: High=14.67, Low=14.56, Open=14.65, Close=14.56, Vol=37457
  \"prompt\": \"Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-25: High=14.67, Low=14.56, Open=14.65, Close=14.56, Vol=374
  \"label\": \"14.9400\""
},
{
  "text": "Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-28: High=14.59, Low=14.48, Open=14.50, Close=14.59, Vol=56914
  \"prompt\": \"Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-28: High=14.59, Low=14.48, Open=14.50, Close=14.59, Vol=569
  \"label\": \"14.8200\""
},
{
  "text": "Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-29: High=14.74, Low=14.54, Open=14.56, Close=14.64, Vol=82809
  \"prompt\": \"Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-29: High=14.74, Low=14.54, Open=14.56, Close=14.64, Vol=828
  \"label\": \"14.7700\""
},
{
  "text": "Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-30: High=14.68, Low=14.57, Open=14.61, Close=14.65, Vol=32214
  \"prompt\": \"Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-30: High=14.68, Low=14.57, Open=14.61, Close=14.65, Vol=322
  \"label\": \"14.8200\""
},
{
  "text": "Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-31: High=14.96, Low=14.65, Open=14.65, Close=14.87, Vol=11150
  \"prompt\": \"Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-08-31: High=14.96, Low=14.65, Open=14.65, Close=14.87, Vol=111
  \"label\": \"14.8400\""
},
{
  "text": "Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-09-01: High=14.91, Low=14.88, Open=14.90, Close=14.90, Vol=26868
  \"prompt\": \"Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-09-01: High=14.91, Low=14.88, Open=14.90, Close=14.90, Vol=268
  \"label\": \"14.9100\""
},
{
  "text": "Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-09-05: High=14.90, Low=14.79, Open=14.90, Close=14.87, Vol=44733
  \"prompt\": \"Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-09-05: High=14.90, Low=14.79, Open=14.90, Close=14.87, Vol=447
  \"label\": \"15.1400\""
},
{
  "text": "Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-09-06: High=14.85, Low=14.65, Open=14.85, Close=14.77, Vol=51059
  \"prompt\": \"Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-09-06: High=14.85, Low=14.65, Open=14.85, Close=14.77, Vol=510
  \"label\": \"15.1600\""
},
{
  "text": "Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-09-07: High=14.80, Low=14.68, Open=14.71, Close=14.68, Vol=47995
  \"prompt\": \"Asset: CMC\nPredict the close price of CMC after 3 days given that.\nRecent Data:\n 2017-09-07: High=14.80, Low=14.68, Open=14.71, Close=14.68, Vol=479
  \"label\": \"14.9700\""
}
}
```

Giá đóng cửa dự đoán với mô hình học máy học sâu

LSTM

FPT

3DAYS Predicted prices for the next 3 days: [107.031975, 112.07983, 112.514885]

7 ngày Predicted prices for the next 7 days: [100.656, 99.6764, 104.417305, 99.24233, 93.11404, 96.1277, 99.34586]

21 ngày Predicted prices for the next 21 days: [111.89914, 123.35149, 124.44448, 114.52951, 112.27009, 117.09893, 116.16199, 117.407486, 98.94289, 106.366035, 106.10076, 116.71834, 117.10671, 111.40764, 100.60892, 101.999985, 122.112366, 105.19594, 111.57098, 109.96381, 123.9079]

63 ngày Predicted prices for the next 63 days: [129.10124, 131.89734, 125.66589, 133.93631, 132.47812, 132.96368, 124.1714, 122.227844, 115.345924, 76.30237, 111.05956, 75.88941, 47.83464, 42.85803, 42.26599, 20.439072, 32.27601, 56.068584, 76.17237, 80.700165, 82.09679, 105.336784, 112.39676, 97.66254, 95.939545, 81.001015, 87.19712, 79.04767, 75.06122, 67.40283, 70.70277, 80.38206, 72.54467, 70.359, 66.012825, 64.40894, 64.48753, 67.02026, 67.5397, 69.01244, 65.69605, 68.91446, 74.62853, 73.68705, 70.09808, 75.06748, 80.257416, 89.00711, 80.71958, 76.31769, 71.65714, 78.17913, 81.27468, 85.10055, 76.27749, 69.10925, 52.943977, 57.055897, 61.50678, 58.381638, 49.776833, 42.37278, 50.406494]

CMC

3 NGÀY Predictions for the next 3 days: [47.85852, 50.851776, 44.338898]

7 ngày Predictions for the next 7 days: [39.392387, 39.969093, 39.248623, 39.184757, 39.828804, 40.882366, 38.619995]

21 ngày Predictions for the next 21 days: [35.259624, 34.929264, 33.708282, 33.703796, 33.21762, 33.348385, 33.196484, 33.01767, 33.254204, 33.765507, 33.750355, 33.86424, 33.828384, 33.746754, 33.457596, 33.2914, 33.35112, 32.887978, 30.150452, 28.25587, 28.227404]

63 ngày Predictions for the next 63 days: [39.936256, 39.94799, 39.847366, 39.68656, 39.528778, 39.390522, 39.239788, 39.14769, 39.29703, 39.42145, 39.501366, 39.60083, 39.64557, 39.55908, 39.514427, 39.343063, 39.36877, 38.917397, 38.969448, 39.003056, 38.930145, 38.676193, 38.263382, 38.265797, 38.364147, 38.424515, 38.35417, 38.26582, 38.316555, 38.175587, 38.273758, 38.036892, 37.6028, 37.74659, 37.76592, 37.99292, 37.86181, 37.864937, 37.574013, 37.484154, 37.419743, 37.475067, 37.288094, 37.056877, 37.11403, 37.40318, 37.45263, 37.588577, 37.74684, 37.89148, 37.90374, 37.93189, 37.895653, 37.768253, 37.48898, 37.489548, 37.795956, 37.717777, 37.68395, 37.869617, 37.01625, 36.015457, 35.878323]

Giá đóng cửa dự đoán với mô hình tinh chỉnh

ngôn ngữ lớn

Bảng 21: Raw Model Prediction Results (Direct Transcription)

1. microsoft Phi 3 mini 4k instruct 3,8B

Predicted Prices for Next 3 Days:	[28.85, 28.85, 28.85]
Evaluating 3 valid predictions out of 3 requested days	
Evaluation (3 days) - R2:	-1.4659, RMSE: 2.8534, MAE: 2.2000
Predicted Prices for Next 7 Days:	[28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85]
Evaluating 7 valid predictions out of 7 requested days	
Evaluation (7 days) - R2:	-87771.7552, RMSE: 1066.4352, MAE: 575.7000
Predicted Prices for Next 21 Days:	[28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 2025.0]
Evaluating 21 valid predictions out of 21 requested days	
Evaluation (21 days) - R2:	-214612.5693, RMSE: 1734.2665, MAE: 1516.0000
Predicted Prices for Next 63 Days:	[28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 2025.0]
Evaluating 63 valid predictions out of 63 requested days	
Evaluation (63 days) - R2:	-243406.8515, RMSE: 1901.7982, MAE: 1826.0000

2. Llama guard2 8B

Predicted Prices for Next 3 Days:	[28.85, 28.85, 28.85]
Evaluating 3 valid predictions out of 3 requested days	
Evaluation (3 days) - R2:	-1.4659, RMSE: 2.8534, MAE: 2.2000
Predicted Prices for Next 7 Days:	[28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85]

Evaluating 7 valid predictions out of 7 requested days	
Evaluation (7 days) - R2:	-2.7784, RMSE: 6.9969, MAE: 6.0000
Predicted Prices for Next 21 Days:	[28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85, 28.85]
Evaluating 21 valid predictions out of 21 requested days	
Evaluation (21 days) - R2:	-7.4933, RMSE: 10.9100, MAE: 10.2476
Predicted Prices for Next 63 Days:	[28.85, 28.85]
Evaluating 63 valid predictions out of 63 requested days	
Evaluation (63 days) - R2:	-14.1124, RMSE: 14.9852, MAE: 14.4810

3. Gemma 2,1B phi

Predicted Prices for Next 3 Days:	[38.45, 35.8, 33.3]
Evaluating 3 valid predictions out of 3 requested days	
Evaluation (3 days) - R2:	-6.0030, RMSE: 4.8085, MAE: 4.8000
Predicted Prices for Next 7 Days:	[38.45, 35.8, 33.3, 31.0, 28.85, 38.45]
Evaluating 7 valid predictions out of 7 requested days	
Evaluation (7 days) - R2:	-0.9949, RMSE: 5.0841, MAE: 4.5714
Predicted Prices for Next 21 Days:	[38.45, 35.8, 33.3, 31.0, 28.85, 38.45, 38.45, 35.8, 33.3, 31.0, 28.85, 38.45, 38.45]
Evaluating 21 valid predictions out of 21 requested days	
Evaluation (21 days) - R2:	-2.7546, RMSE: 7.2539, MAE: 6.2952
Predicted Prices for Next 63 Days:	[38.45, 35.8, 33.3, 31.0, 28.85, 38.45, 38.45, 35.8, 33.3, 31.0, 28.85, 38.45, 38.45, 35.8, 33.3, 31.0, 28.85, 38.45, 38.45, 35.8, 33.3, 31.0, 28.85, 38.45, 38.45, 35.8, 33.3, 31.0, 28.85, 38.45, 38.45, 35.8, 33.3]

Evaluating 63 valid predictions out of 63 requested days

Evaluation (63 days) - R2:

-7.1950, RMSE: 11.0350, MAE: 10.1952

4. LLama 2 13B hf

Predicted CMC Prices for Next 3 Days:

[27.95, 27.56, 27.56]

CMC (3 days) - R2:

-1.4659262998485616, RMSE: 2.85336059177

Predicted CMC Prices for Next 7 Days:

[28.85, 28.85, 28.85, 28.85, 28.85, 28.85]

CMC (7 days) - R2:

-2.7036152149944863, RMSE: 6.92735674182

Predicted CMC Prices for Next 21 Days (first entry):

[28.85, 28.85, 28.85, 28.85, 28.85, 28.85]

29.86, 29.86, 30.01, 30.01, 30.26, 30.26]

31.26, 31.01, 31.01]

Predicted CMC Prices for Next 21 Days (second entry, used for eval):

[28.85, 28.85, 28.85, 28.85, 28.85, 28.85]

29.36, 29.66, 29.66, 29.66, 29.66, 29.66]

29.9, 30.04, 30.23]

CMC (21 days) - R2:

-6.939076104104134, RMSE: 10.54805104995

5. LLama 3 8B instruct (model pretrain lựa chọn)

CMC:

Predicted CMC Prices for Next 3 Days (generated):

[29.1, 29.1, 29.1]

Predicted CMC Prices for Next 7 Days (generated):

[28.6, 28.85, 29.0, 29.0, 28.85, 28.9]

Predicted CMC Prices for Next 21 Days (generated):

[28.85, 29.1, 30.5, 31.05, 31.05, 31.05]

33.74, 33.94, 33.39, 33.39, 33.19, 33.19]

33.69, 33.59, 33.39]

Predicted CMC Prices for Next 63 Days (generated):

[29.1, 29.1, 29.1, 29.1, 29.1, 29.1, 29.1]

29.68, 29.68, 29.38, 29.68, 30.39, 30.39]

31.94, 32.08, 31.94, 32.08, 32.38, 32.38]

33.92, 33.92, 34.45, 34.45, 33.55, 33.55]

33.05, 33.59, 33.59, 33.59, 33.05, 33.05]

33.05, 33.25, 33.5, 33.5, 33.8, 33.8, 33.8,

33.5, 33.5, 33.8, 33.8, 33.8, 33.93, 33.93]

Predicted CMC Prices for Next 3 Days (used for eval):

[30.5, 31.05, 31.05]

CMC (3 days) - R2:

-0.2804, RMSE: 2.0561, MAE: 1.6833

Predicted CMC Prices for Next 7 Days (used for eval):

[30.5, 31.05, 31.05]

CMC (7 days) - R2:

-973.3301, RMSE: 7.4668, MAE: 7.4667

Predicted CMC Prices for Next 21 Days (used for eval):

[30.5, 31.05, 31.05]

CMC (21 days) - R2:

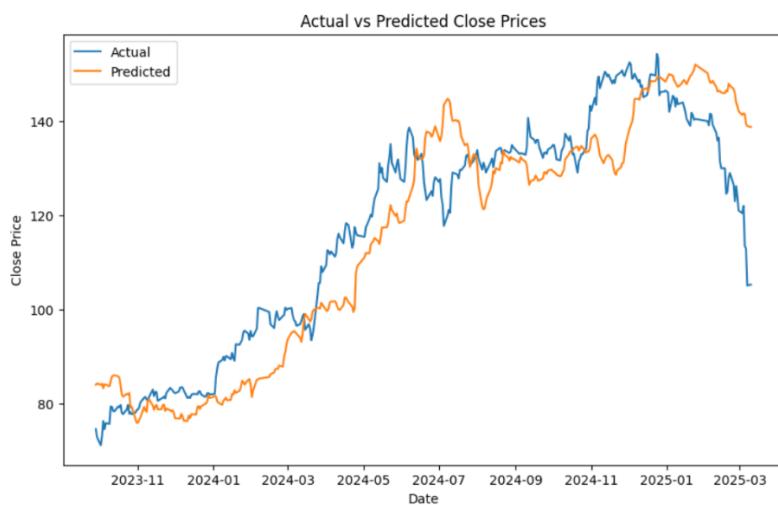
-1047.5063, RMSE: 11.7496, MAE: 11.7333

Predicted CMC Prices for Next 63 Days (used for eval):	[30.5, 31.05, 31.05]
CMC (63 days) - R2:	-1272.4376, RMSE: 17.9809, MAE: 17.9667
<i>FPT:</i>	
Predicted FPT Prices for Next 3 Days:	[122.0, 113.5, 105.6]
FPT (3 days) - R2:	-2.737062592410053, RMSE: 7.109852319141
Predicted FPT Prices for Next 7 Days:	[122.0, 113.5, 105.6, 105.6, 105.6, 105.6, 105.6]
FPT (7 days) - R2:	-0.4053513274824754, RMSE: 7.89638253524
Predicted FPT Prices for Next 21 Days:	[122.0, 113.5, 105.6, 105.6, 105.6, 105.6, 105.6, 105.6, 105.6, 105.6]
FPT (21 days) - R2:	-3.734106146299812, RMSE: 19.13360295147
Predicted FPT Prices for Next 63 Days:	[122.0, 113.5, 105.6]
FPT (63 days) - R2:	-7.747234659499245, RMSE: 34.13315197124

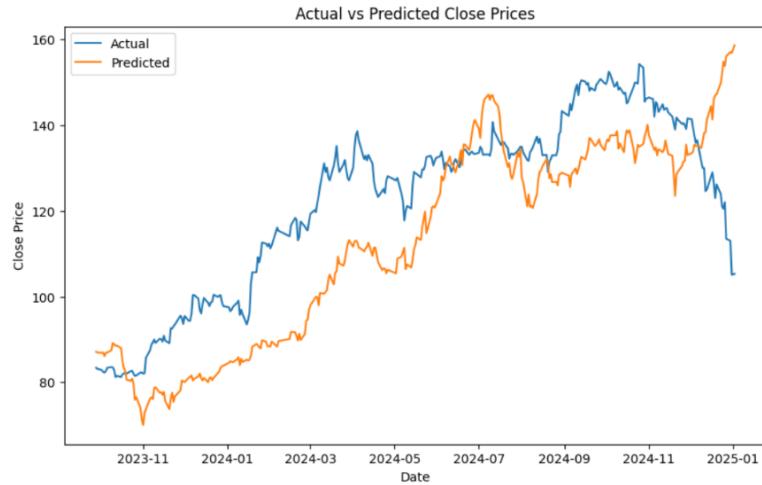
Một số hình ảnh bổ sung



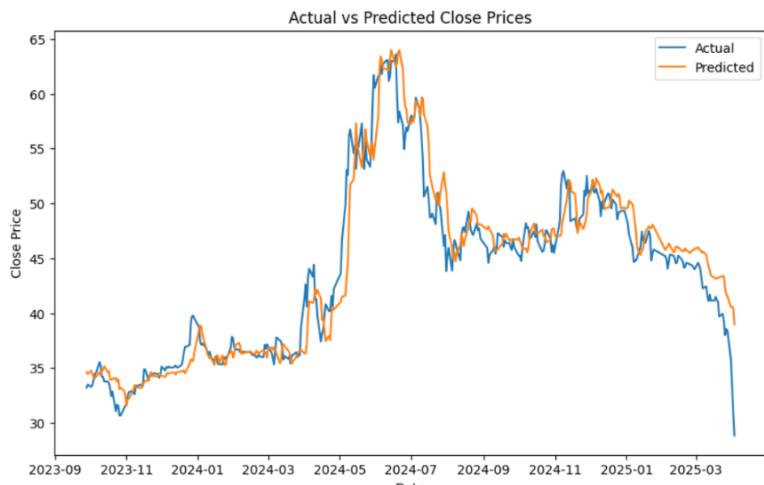
Hình 28: XGBoost FPT 7 ngày



Hình 29: XGBoost FPT 21 ngày



Hình 30: XGBoost FPT 63 ngày



Hình 31: XGBoost CMC 3 ngày

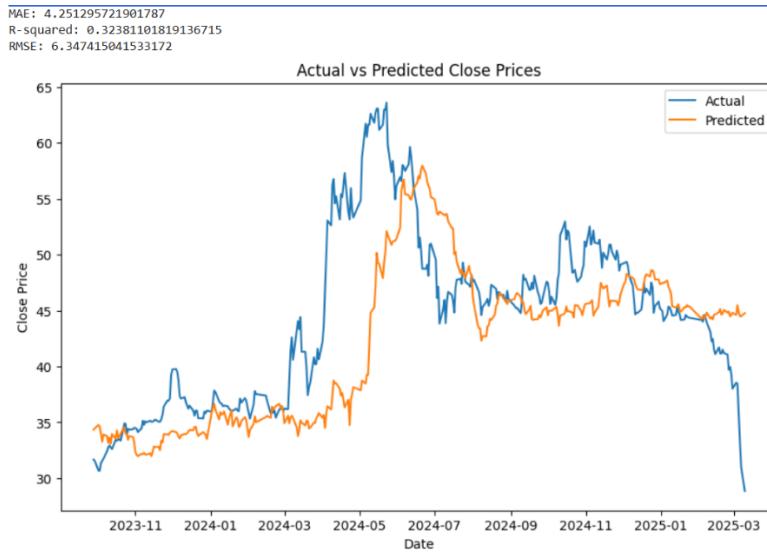


Hình 32: XGBoost CMC 7 ngày

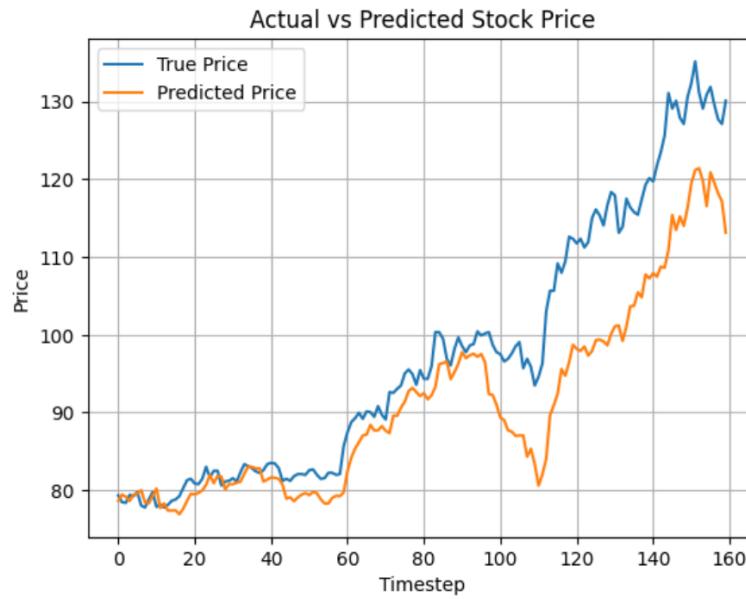
MAE: 4.251295721901787
R-squared: 0.32381101819136715
RMSE: 6.347415041533172



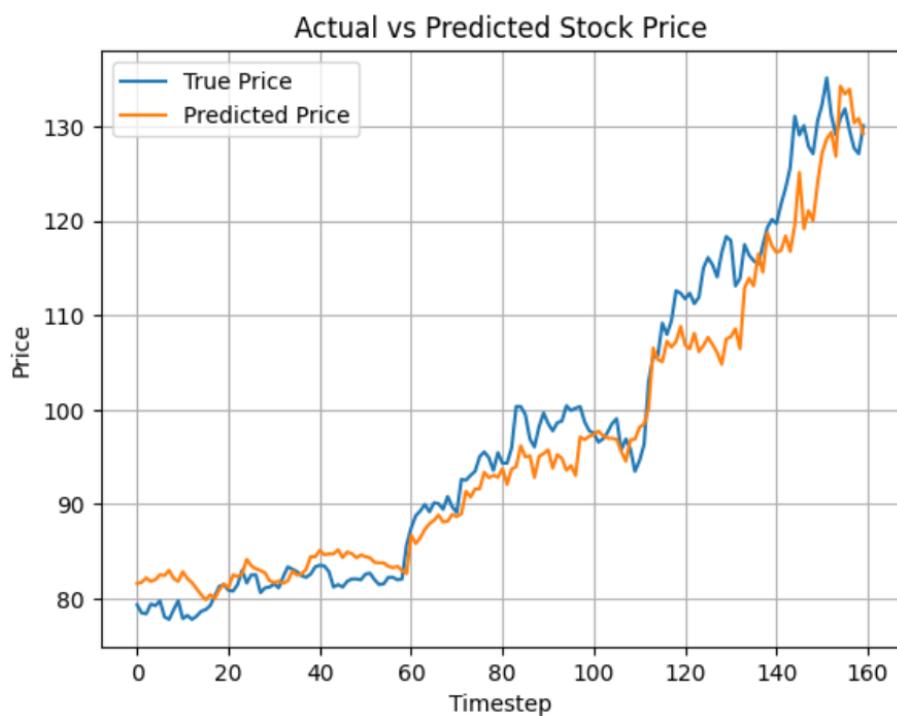
Hình 33: XGBoost CMC 21 ngày



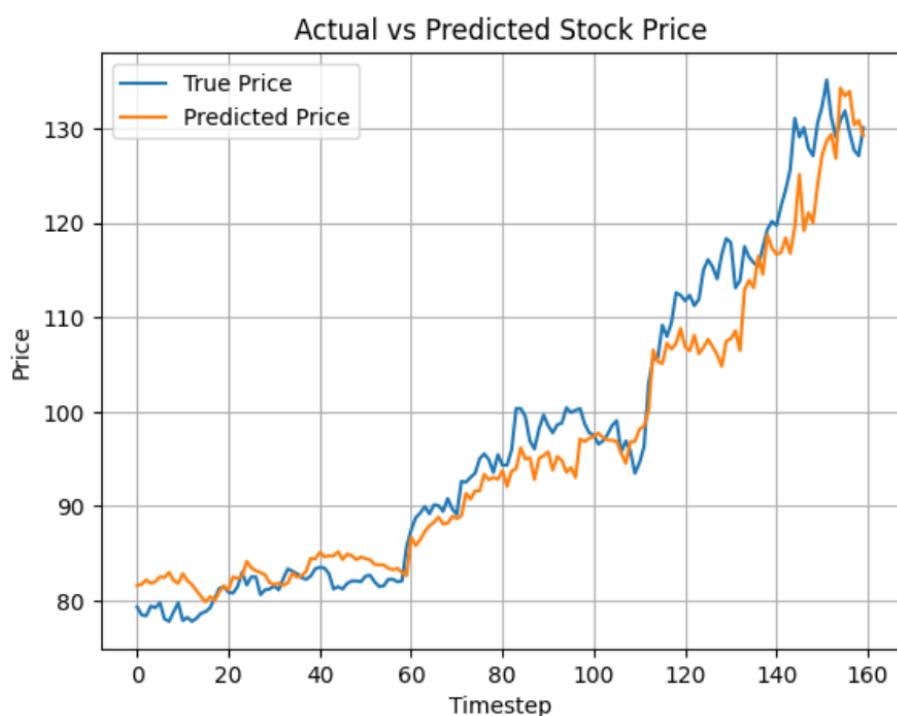
Hình 34: XGBoost CMC 63 ngày



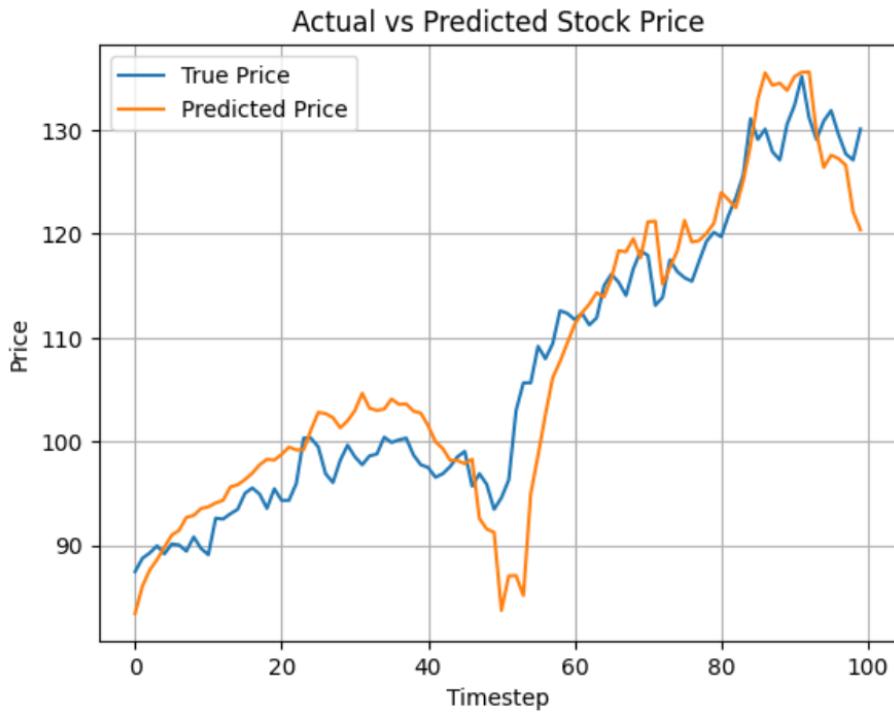
Hình 35: LSTM FPT 3 ngày



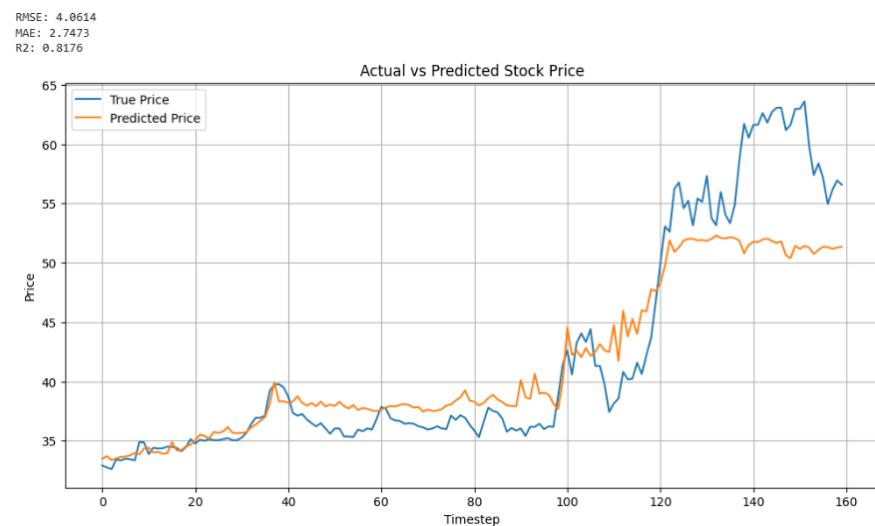
Hình 36: LSTM FPT 7 ngày



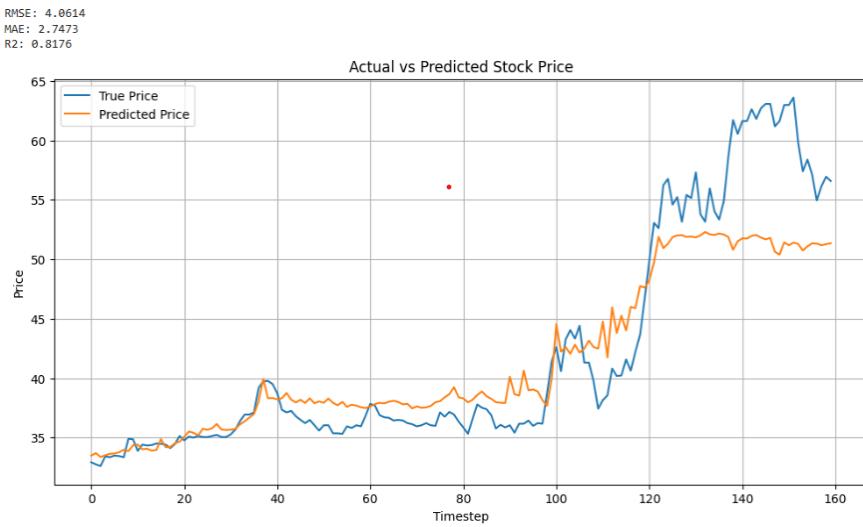
Hình 37: LSTM FPT 21 ngày



Hình 38: LSTM FPT 63 ngày



Hình 39: LSTM CMC 3 ngày

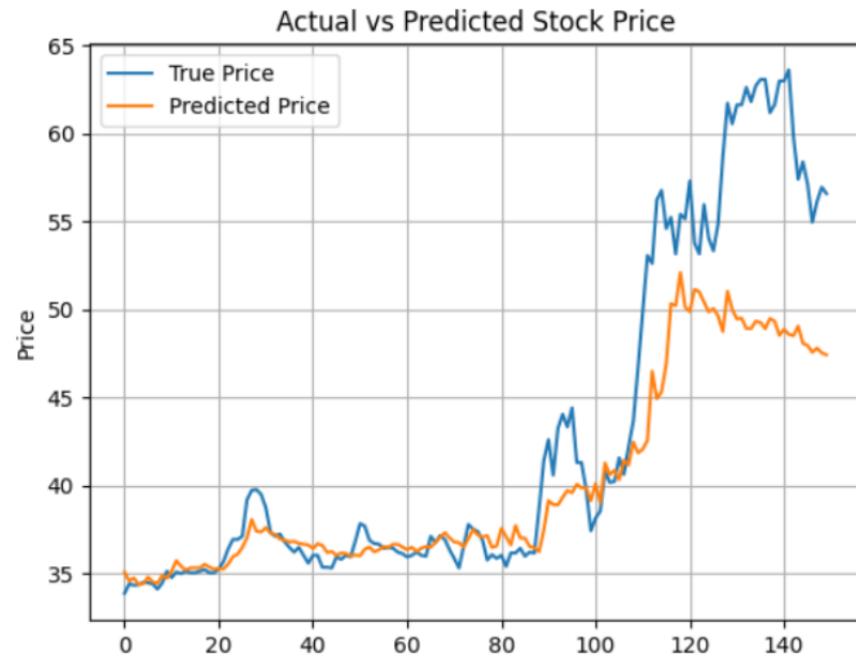


Hình 40: LSTM CMC 7 ngày

RMSE: 5.2286

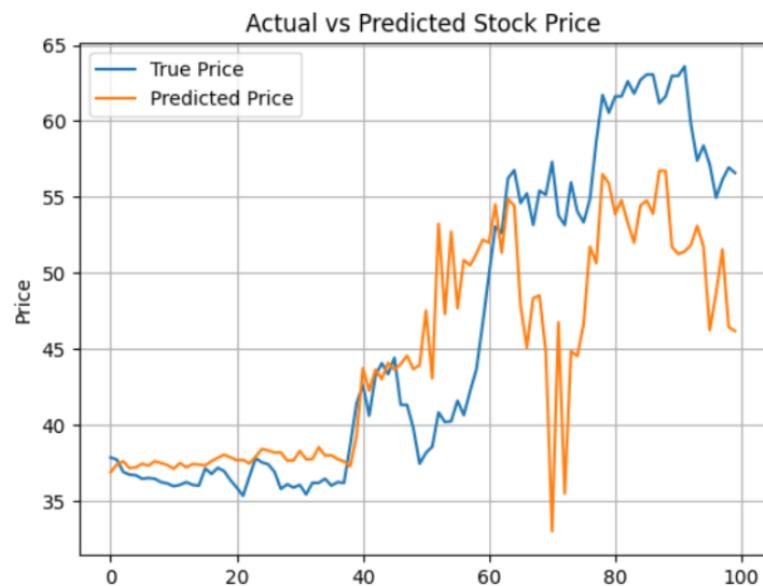
MAE: 3.0904

R2: 0.7000



Hình 41: LSTM CMC 21 ngày

RMSE: 6.4021
MAE: 4.6745
R2: 0.5962



Hình 42: LSTM CMC 63 ngày