TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Công trình Nghiên cứu Khoa học Sinh viên

năm học 2024 - 2025

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

ĐƠN VỊ KHOA CÔNG NGHỆ THÔNG TIN

GIẢNG VIÊN HƯỚNG DẪN:

TIẾN SĨ TRỊNH HÙNG CƯỜNG

NHÓM SINH VIÊN THỰC HIỆN:

1. NGUYỄN QUANG HUY

2. NGUYỄN TRẦN NHẬT AN

TP. Hồ Chí Minh, tháng 7 năm 2025

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

CÔNG TRÌNH NGHIÊN CỨU KHOA HỌC SINH VIÊN NĂM HỌC 2024-2025

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

Giảng viên hướng dẫn: **TS. TRỊNH HÙNG CƯỜNG**

Nhóm sinh viên thực hiện: **NGUYỄN QUANG HUY - 523H0140**

NGUYỄN TRẦN NHẬT AN - 523H0115

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025

LỜI CẢM ƠN

Lời đầu tiên, chúng em xin gửi lời cảm ơn chân thành nhất đến Ban giám hiệu trường Đại Học Tôn Đức Thắng vì đã tạo điều kiện cho chúng em hoàn thành công trình nghiên cứu khoa học này thông qua hệ thống thư viện đa dạng tài liệu hay và bổ ích.

Chúng em xin chân thành cảm ơn **Thầy T.S. Trịnh Hùng Cường** đã giảng dạy và truyền đạt kiến thức một cách tận tình, chi tiết đồng thời cung cấp tài liệu tham khảo giúp chúng em đủ nền tảng để vận dụng vào việc viết bài báo cáo này.

Trong quá trình làm bài báo cáo nghiên cứu, việc chúng em khó tránh khỏi thiếu sót là điều chắc chắn, chúng em rất mong nhận được những ý kiến đóng góp quý báu của quý thầy cô để kiến thức của chúng em được hoàn thiện hơn.

Sau cùng em xin chân thành cảm ơn và kính chúc quý thầy cô trong khoa Công Nghệ Thông Tin luôn khỏe mạnh và thành công trong sự nghiệp giảng dạy.

LỜI CAM KẾT

Chúng em xin cam đoan đây là sản phẩm nghiên cứu của riêng chúng em và được sự hướng dẫn của TS. Trịnh Hùng Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét được chính chúng em thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong nghiên cứu còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào chúng em xin hoàn toàn chịu trách nhiệm về nội dung nghiên cứu của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng em gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 8 tháng 8 năm 2025

Tác giả

(ký tên và ghi rõ họ tên)

Nguyễn Quang Huy

Nguyễn Trần Nhật An

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

TÓM TẮT

Nghiên cứu này trình bày việc xây dựng một hệ thống hỗ trợ đầu tư chứng khoán thông minh, với mục tiêu cốt lõi là nâng cao hiệu quả ra quyết định cho nhà đầu tư tại thị trường chứng khoán Việt Nam. Điểm đột phá của đề tài nằm ở việc kết hợp các mô hình trí tuệ nhân tạo (AI) tiên tiến để phân tích đồng thời hai nguồn dữ liệu quan trọng: dữ liệu định lượng (giá lịch sử) và dữ liệu định tính (văn bản tài chính).

Cụ thể, hệ thống tích hợp sức mạnh của mạng LSTM (Long Short-Term Memory) trong việc dự báo xu hướng giá dựa trên chuỗi dữ liệu giá lịch sử, cùng với khả năng phân tích và lượng hóa cảm tính thị trường từ các nguồn tin tức tài chính của mô hình ngôn ngữ tự nhiên (NLP) PhoBERT. Bằng cách tiếp cận đa chiều này, nghiên cứu hướng tới việc cung cấp một công cụ phân tích sâu sắc, giúp nhà đầu tư giảm thiểu các quyết định cảm tính và đưa ra chiến lược đầu tư dựa trên số liệu dự đoán.

Đề tài nghiên cứu gồm bốn chương:

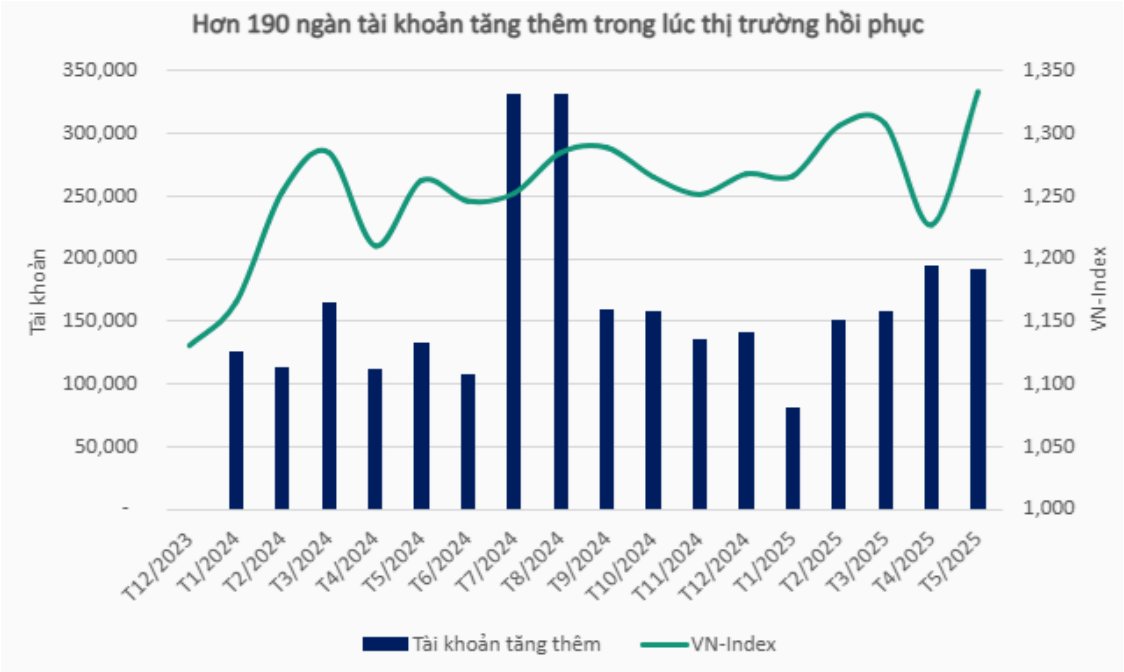
- **Chương 1 - Mở đầu:** Giới thiệu tổng quan, xác định vấn đề nghiên cứu, mục tiêu, phạm vi và làm rõ tính cấp thiết của đề tài trong bối cảnh thị trường chứng khoán Việt Nam hiện nay.
- **Chương 2 - Cơ sở lý thuyết:** Trình bày các nền tảng kiến thức về thị trường chứng khoán, phân tích tài chính và các mô hình học máy trọng tâm được sử dụng như Hồi quy Logistic, LSTM và PhoBERT.
- **Chương 3 - Phương pháp đề xuất và Xây dựng hệ thống:** Mô tả chi tiết kiến trúc hệ thống, quy trình thu thập, tiền xử lý dữ liệu, phương pháp gán nhãn dữ liệu và phương pháp xây dựng, tích hợp các mô hình dự báo.
- **Chương 4 - Kết quả và Thảo luận:** Trình bày kết quả thực nghiệm, đánh giá hiệu quả của mô hình thông qua các chỉ số đo lường và kiểm thử ngược (backtesting), từ đó rút ra kết luận và đề xuất các hướng phát triển trong tương lai.

CHƯƠNG 1. MỞ ĐẦU

1.1 Đặt vấn đề

Thị trường chứng khoán luôn được coi là một trong những kênh đầu tư hấp dẫn nhưng cũng tiềm ẩn rất nhiều rủi ro. Trong bối cảnh kinh tế toàn cầu biến động phức tạp và khó lường, việc đầu tư vào thị trường này ngày càng đòi hỏi nhà đầu tư phải có kỹ năng phân tích và dự báo với độ chính xác cao để đưa ra những quyết định đúng đắn. Tuy nhiên, thực tế cho thấy phần lớn các nhà đầu tư cá nhân vẫn gặp rất nhiều khó khăn do thiếu kinh nghiệm, thông tin và khả năng phân tích chuyên sâu.

Tại Việt Nam, thị trường chứng khoán đang trong giai đoạn phát triển mạnh mẽ, thu hút một lượng lớn nhà đầu tư mới tham gia. Theo số liệu từ Tổng công ty Lưu ký và Bù trừ chứng khoán Việt Nam (VSDC), trong những năm gần đây, hàng tháng có từ 90.000 cho tới hơn 300.000 tài khoản được mở mới. Cùng với đó, chỉ số VN-Index cũng có sự tăng trưởng ấn tượng, có thời điểm vượt mốc 1.300 điểm với giá trị giao dịch mỗi phiên lên đến hàng tỷ đô la Mỹ, cho thấy sức hấp dẫn không thể phủ nhận của thị trường chứng khoán Việt Nam.



Hình 1.1.1. Số lượng tài khoản giao dịch được đăng ký và chỉ số VN-Index hàng tháng.

Tuy nhiên, sự gia tăng nhanh chóng của các nhà đầu tư mới, mà phần lớn chưa được trang bị đầy đủ kiến thức và kinh nghiệm, thường dẫn đến các quyết định đầu tư

cảm tính, thiếu cơ sở khoa học. Chính điều này không chỉ ảnh hưởng trực tiếp tới kết quả đầu tư của mỗi cá nhân mà còn góp phần tạo ra những biến động giá mạnh, khó lường và gây rủi ro cho sự ổn định chung của toàn thị trường.

Trong bối cảnh đó, trí tuệ nhân tạo (AI) nổi lên như một công cụ hỗ trợ đắc lực, có khả năng xử lý những bộ dữ liệu tài chính khổng lồ, nhận diện các mẫu hình giá phức tạp, phân tích cảm tính thị trường từ hàng triệu nguồn tin tức và dự báo xu hướng với độ chính xác cao hơn so với các phương pháp truyền thống. Việc tích hợp AI vào các hoạt động phân tích sẽ giúp nhà đầu tư, đặc biệt là người mới, có thể tiếp cận thị trường một cách khoa học hơn, giảm thiểu rủi ro và nâng cao hiệu quả đầu tư.

1.2 Lý do chọn đề tài

Mặc dù trí tuệ nhân tạo (AI) đang dần trở thành một công cụ mạnh mẽ trong lĩnh vực tài chính toàn cầu. Song, việc ứng dụng công nghệ này vào đầu tư chứng khoán tại Việt Nam vẫn còn khá sơ khai và tồn tại nhiều hạn chế. Đa phần các công cụ hiện có trên thị trường chỉ tập trung vào phân tích kỹ thuật đơn thuần, dựa trên dữ liệu giá lịch sử, mà chưa khai thác được chiều sâu từ các yếu tố phân tích cơ bản – vốn là nền tảng định giá doanh nghiệp.

Nhận thấy khoảng trống này, đề tài được thực hiện với mục tiêu xây dựng một giải pháp toàn diện hơn, kết hợp hài hòa sức mạnh dự báo của các mô hình trí tuệ nhân tạo (AI) với chiều sâu của phân tích cơ bản. Bằng cách này, nghiên cứu không chỉ cung cấp một công cụ phân tích đa chiều mà còn hướng tới việc giúp nhà đầu tư, đặc biệt là nhà đầu tư cá nhân, có một cơ sở vững chắc hơn để ra quyết định, từ đó tối ưu hóa lợi nhuận và quản trị rủi ro hiệu quả.

1.3 Đối tượng nghiên cứu

Trong đề tài này, đối tượng nghiên cứu bao gồm:

- Các mô hình trí tuệ nhân tạo đang được áp dụng trong dự báo và phân tích xu hướng thị trường chứng khoán.
- Phân tích các bài báo, báo cáo tài chính và nhận định từ các chuyên gia ảnh hưởng tới xu hướng ra quyết định của các nhà đầu tư.

1.4 Phạm vi nghiên cứu

Nghiên cứu được giới hạn trong các phạm vi cụ thể sau:

- Không gian: Tập trung vào Thị trường chứng khoán Việt Nam.
- Đối tượng: Các mã cổ phiếu trong rổ VN30, là nhóm cổ phiếu có giá trị vốn hóa lớn, tính thanh khoản cao và đại diện cho xu hướng chung của toàn thị trường chứng khoán Việt Nam.
- Thời gian: Dữ liệu lịch sử sẽ được thu thập và phân tích trong giai đoạn từ năm 2010 đến nay, nhằm đảm bảo tập dữ liệu đủ lớn, bao quát nhiều chu kỳ biến động của thị trường và đảm bảo tính cập nhật cho mô hình.

1.5 Những mô hình trí tuệ nhân tạo (AI) hiện nay

Trong khuôn khổ nghiên cứu, đề tài sẽ xem xét và áp dụng một số mô hình trí tuệ nhân tạo tiêu biểu, từ các mô hình cổ điển đến các kiến trúc hiện đại, để giải quyết bài toán dự báo và phân tích.

1.5.1 Hồi quy Logistic (Logistic Regression)

Hồi quy Logistic là một mô hình thống kê kinh điển, hoạt động dựa trên giả định về mối quan hệ tuyến tính giữa các đặc trưng đầu vào và logit của biến mục tiêu (ví dụ: xác suất giá tăng/giảm).

- Ưu điểm: Ưu điểm lớn nhất của mô hình này là sự đơn giản và dễ diễn giải. Các trọng số của mô hình cho thấy rõ mức độ tác động của từng yếu tố đầu vào lên kết quả dự báo, giúp nó trở thành một công cụ hiệu quả để sàng lọc và đánh giá nhanh các biến quan trọng.
- Hạn chế: Khả năng mô hình hóa các mối quan hệ phi tuyến tính phức tạp là rất kém, điều này trở thành một nhược điểm lớn khi áp dụng vào dữ liệu tài chính vốn biến động và đa chiều. Do đó, Hồi quy Logistic thường được sử dụng như một mô hình cơ sở (baseline) để so sánh hiệu quả với các thuật toán phức tạp hơn.

1.5.2 LSTM (Long Short-Term Memory)

LSTM là một kiến trúc mạng nơ-ron hồi quy (RNN) cải tiến, được thiết kế đặc biệt để khắc phục vấn đề "suy giảm/ bùng nổ gradient" (vanishing/exploding gradient) trong các mạng RNN truyền thống. Điều này cho phép nó học và ghi nhớ các phụ thuộc trong chuỗi dữ liệu qua những khoảng thời gian dài.

- Ưu điểm: Với cấu trúc cổng (gate) gồm: cổng đầu vào (input), cổng quên (forget) và cổng đầu ra (output), LSTM có khả năng vượt trội trong việc mô hình hóa dữ liệu chuỗi thời gian (time-series) như dữ liệu giá cổ phiếu. Nó có thể nắm bắt được các quy luật và mẫu hình ẩn mà các phương pháp tuyến tính không thể nhận ra.
- Hạn chế: Việc huấn luyện LSTM đòi hỏi một tập dữ liệu lớn và tốn nhiều tài nguyên tính toán. Hơn nữa, mô hình có nhiều siêu tham số cần được tinh chỉnh cẩn thận để đạt hiệu suất tối ưu.

1.5.3 PhoBERT

PhoBERT là một mô hình ngôn ngữ tiên tiến, được xây dựng dựa trên kiến trúc BERT của Google và được tiền huấn luyện (pre-trained) với một kho dữ liệu khổng lồ của tiếng Việt (hơn 20GB). Đây là mô hình chuyên biệt và hiệu quả hàng đầu cho các tác vụ xử lý ngôn ngữ tự nhiên (NLP) trong tiếng Việt.

- Ưu điểm:
    - Tối ưu cho tiếng Việt: Do được huấn luyện chuyên sâu trên kho ngữ liệu tiếng Việt, PhoBERT có khả năng hiểu sâu sắc các sắc thái, ngữ cảnh và đặc thù của ngôn ngữ, giúp phân tích cảm xúc chính xác hơn nhiều so với các mô hình đa ngôn ngữ thế hệ cũ.
    - Hiệu suất cao: PhoBERT được xem là một mô hình mạnh mẽ, đạt hiệu suất hàng đầu trong nhiều bài toán NLP tiếng Việt như phân loại văn bản, nhận dạng thực thể và phân tích cảm xúc.

- Nhược điểm:
    - Không chuyên biệt về lĩnh vực tài chính: PhoBERT được huấn luyện trên dữ liệu tiếng Việt tổng quát (báo chí, Wikipedia), không phải dữ liệu chuyên ngành tài chính. Do đó, mô hình có thể gặp khó khăn trong việc hiểu các thuật ngữ, biệt ngữ và sắc thái tinh vi đặc thù của lĩnh vực tài chính - chứng khoán.
    - Yêu cầu tiền xử lý: Văn bản đầu vào cần được tách từ (word segmentation) bằng các công cụ chuyên dụng cho tiếng Việt (ví dụ: VnCoreNLP) trước khi đưa vào mô hình, làm tăng thêm một bước trong quy trình xử lý.

1.5.4 FinBERT

FinBERT là một mô hình ngôn ngữ dựa trên kiến trúc BERT nhưng được tiền huấn luyện chuyên biệt trên một kho dữ liệu tài chính khổng lồ bằng tiếng Anh, chẳng hạn như báo cáo tài chính (10-K, 10-Q), bản ghi cuộc gọi thu nhập và tin tức tài chính. Mục tiêu của FinBERT là để "hiểu" sâu sắc ngôn ngữ và bối cảnh của ngành tài chính, một nhiệm vụ mà các mô hình ngôn ngữ tổng quát thường gặp khó khăn.

- Ưu điểm:
    - Hiểu biết chuyên sâu về lĩnh vực tài chính: FinBERT vượt trội trong việc phân tích văn bản tài chính vì nó nhận diện được các thuật ngữ và ngữ cảnh đặc thù của ngành, giúp phân loại cảm xúc (tích cực, tiêu cực, trung tính) với độ chính xác rất cao.
    - Hiệu quả với dữ liệu giới hạn: Do đã được học trước từ một lượng lớn dữ liệu tài chính chưa gán nhãn, FinBERT có thể được tinh chỉnh (fine-tune) hiệu quả trên các tập dữ liệu gán nhãn nhỏ hơn mà vẫn cho kết quả tốt.

- Nhược điểm:
    - Hạn chế về ngôn ngữ: Đây là hạn chế lớn nhất trong phạm vi đề tài này. FinBERT được phát triển và huấn luyện chủ yếu cho tiếng Anh. Hiện chưa có một phiên bản FinBERT được tiền huấn luyện chính thức và mạnh mẽ cho tiếng Việt, khiến việc áp dụng trực tiếp là không khả thi.
    - Yêu cầu tài nguyên lớn: Giống như các mô hình Transformer lớn khác, việc huấn luyện và triển khai FinBERT đòi hỏi tài nguyên tính toán (GPU) đáng kể.

1.6 Những vấn đề hạn chế trong phát triển trí tuệ nhân tạo

Dự báo thị trường chứng khoán là một bài toán vô cùng phức tạp do bản chất nhiễu (noisy) và "liên tục/không dừng" (non-stationary) của dữ liệu. Giá cổ phiếu không chỉ bị tác động bởi dữ liệu quá khứ mà còn chịu ảnh hưởng của vô số yếu tố đa chiều, từ kinh tế vĩ mô (lãi suất, lạm phát), tin tức tài chính, sự kiện chính trị, cho đến tình hình tài chính của doanh nghiệp. Chính sự phức tạp này tạo ra những thách thức lớn cho việc ứng dụng các mô hình trí tuệ nhân tạo (AI).

1.6.1 Thách thức trong việc hiểu và lượng hóa bối cảnh thị trường

- **Vấn đề:** Các mô hình dự báo truyền thống chủ yếu dựa trên dữ liệu định lượng (giá, khối lượng giao dịch), nhưng lại thường "bỏ qua" các yếu tin tức, báo cáo tài chính, hay sự kiện bất ngờ. Những thông tin phi cấu trúc này lại chứa đựng các tín hiệu quan trọng về tâm lý nhà đầu tư và có thể gây ra những biến động đột ngột mà mô hình thuần số liệu không thể lường trước.
- **Hướng nghiên cứu của đề tài:** Để giải quyết vấn đề này, nghiên cứu sẽ tích hợp các nguồn dữ liệu phi cấu trúc bằng cách sử dụng các mô hình xử lý ngôn ngữ tự nhiên (NLP) tiên tiến như PhoBERT. Mô hình sẽ được dùng để thực hiện phân tích cảm xúc (sentiment analysis) từ tin tức, từ đó chuyển hóa dữ liệu định tính thành một đặc trưng (feature) định lượng.

1.6.2 Thách thức về tính ổn định của mô hình trước biến động thị trường

- Vấn đề: Bản chất của thị trường là luôn thay đổi, các quy luật trong quá khứ có thể không còn đúng trong tương lai. Điều này khiến các mô hình AI, đặc biệt là những mô hình phức tạp như mạng nơ-ron, có nguy cơ bị overfitting vào dữ liệu cũ và cho ra kết quả dự báo kém ổn định, thiếu tin cậy khi thị trường bước vào một giai đoạn biến động mới.
- Hướng nghiên cứu: Áp dụng LSTM kết hợp với các đặc trưng cảm xúc nhằm tối ưu hóa khả năng tổng hợp và điều chỉnh dự báo. Cập nhật dữ liệu hàng tuần sẽ giúp duy trì sự nhất quán trong điều kiện thị trường mới.

1.6.3 Thách thức về xử lý dữ liệu và tốc độ phản hồi

- **Vấn đề:** Thị trường chứng khoán tạo ra một khối lượng dữ liệu khổng lồ theo thời gian thực. Để các quyết định đầu tư có giá trị, việc thu thập, xử lý dữ liệu và đưa ra dự báo phải diễn ra gần như tức thì. Các mô hình phức tạp, nếu không được tối ưu, sẽ có độ trễ lớn, làm mất đi tính kịp thời của dự báo.
- **Hướng nghiên cứu:** Nghiên cứu sẽ tập trung vào việc xây dựng một quy trình xử lý dữ liệu (data pipeline) hiệu quả, từ thu thập, xử lý đến đưa vào mô hình. Đồng thời, nghiên cứu sẽ xem xét các phương pháp tối ưu hóa kiến trúc mô hình và hạ tầng tính toán để đảm bảo tốc độ phản hồi nhanh, đáp ứng yêu cầu của một hệ thống hỗ trợ ra quyết định trong thực tế.

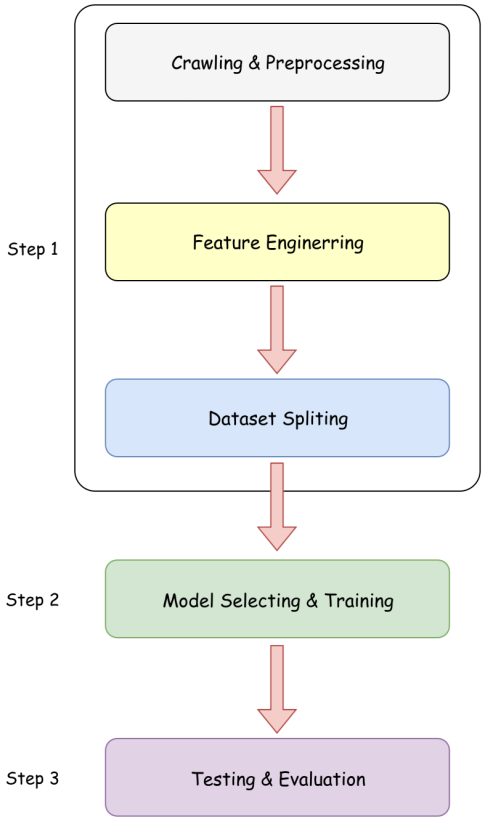
1.6.4 Kết luận

Dù các mô hình như Hồi quy Logistic, LSTM, FinBERT và PhoBERT đã chứng tỏ tiềm năng, chúng vẫn còn những hạn chế cố hữu. Hướng đi của đề tài là giải quyết các thách thức này thông qua việc tích hợp đa dạng nguồn dữ liệu (cả số và văn bản) và kết hợp nhiều mô hình (ensemble) để xây dựng một hệ thống dự báo toàn diện, ổn định và hiệu quả hơn.

1.7 Phương pháp nghiên cứu

Phương pháp nghiên cứu được chia thành 3 giai đoạn:

- **Giai đoạn 1** tập trung vào thu thập và tiền xử lý dữ liệu định lượng (giá cổ phiếu VN30 từ vnstock) và định tính (tin tức tài chính từ CafeF, Báo Đầu Tư). Dữ liệu sau đó được làm sạch, chuẩn hóa và tạo đặc trưng (chỉ báo kỹ thuật; gán nhãn cảm xúc từ tin tức dựa trên hệ số Alpha).
- **Giai đoạn 2** tiến hành xây dựng và huấn luyện mô hình. Dữ liệu được phân chia (cảm xúc ngẫu nhiên; chuỗi thời gian theo thứ tự). Mô hình Hồi quy Logistic được dùng làm cơ sở, sau đó mô hình LSTM tích hợp đặc trưng cảm xúc từ PhoBERT được xây dựng và huấn luyện để dự báo xu hướng giá. Cuối cùng,
- **Giai đoạn 3** đánh giá và kiểm thử mô hình bằng các chỉ số Accuracy, Precision, Recall, F1-Score (phân loại) hoặc MSE, RMSE (dự đoán giá) để đánh giá hiệu quả thực tế.



Hình 1.7.1. Quy trình nghiên cứu

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 Phân tích cảm xúc (Sentiment Analysis) trong lĩnh vực tài chính

Phân tích cảm xúc, còn được biết đến với các thuật ngữ như khai phá ý kiến (opinion mining), là một lĩnh vực của xử lý ngôn ngữ tự nhiên (NLP) tập trung vào việc sử dụng các công cụ tính toán để xác định và phân loại thái độ, cảm xúc hoặc quan điểm được thể hiện trong một đoạn văn bản. Trong bối cảnh tài chính, nơi thông tin là tài sản quý giá và có thể tác động trực tiếp đến quyết định đầu tư, việc tự động hóa quá trình "đọc hiểu" hàng ngàn tin tức, báo cáo và bình luận mỗi ngày đã trở thành một công cụ chiến lược quan trọng. Các doanh nghiệp và nhà đầu tư sử dụng phân tích cảm xúc để nắm bắt nhanh chóng phản ứng của thị trường trước các sự kiện, đánh giá quan điểm của cộng đồng về một cổ phiếu cụ thể, và từ đó đưa ra các quyết định giao dịch kịp thời hơn.

Tuy nhiên, một trong những thách thức lớn nhất khi xây dựng các mô hình phân tích cảm xúc hiệu quả, đặc biệt là cho các ngôn ngữ ít phổ biến hơn như tiếng Việt, là sự thiếu hụt các bộ dữ liệu được gán nhãn chất lượng cao. Việc gán nhãn thủ công bởi các chuyên gia vừa tốn kém về thời gian và chi phí, vừa có thể mang tính chủ quan. Trong khi đó, các phương pháp dựa trên từ điển (lexicon-based) thường gặp khó khăn trong việc bao phủ hết các sắc thái ngữ nghĩa phức tạp và đặc thù của ngành tài chính, nơi một từ có thể mang ý nghĩa khác nhau tùy thuộc vào ngữ cảnh. Thực tế này đã thúc đẩy sự phát triển của các phương pháp luận tiên tiến hơn, cho phép gán nhãn dữ liệu một cách tự động và khách quan, chẳng hạn như dựa trên phản ứng của thị trường hoặc các kỹ thuật giám sát yếu (weak supervision), vốn là trọng tâm của đề tài này.

2.2 Phương pháp luận gán nhãn dữ liệu dựa trên hiệu suất thị trường

2.2.1 Lý thuyết nghiên cứu sự kiện (Event Study)

a. Tổng quan

Nghiên cứu sự kiện là một phương pháp kinh tế lượng được sử dụng rộng rãi trong lĩnh vực tài chính để đo lường tác động của một sự kiện cụ thể lên giá trị của một công ty. Sự kiện ở đây có thể là bất cứ điều gì từ việc công

bố báo cáo tài chính, thông báo sáp nhập, thay đổi nhân sự cấp cao, cho đến việc ra mắt một sản phẩm mới. Trong khuôn khổ của đề tài này, "sự kiện" chính là việc công bố một bản tin tức liên quan đến một cổ phiếu cụ thể. Giả định cơ bản của phương pháp này là nếu thị trường hoạt động hiệu quả, tác động của sự kiện sẽ được phản ánh gần như ngay lập tức vào giá cổ phiếu.

b. Mục tiêu và ứng dụng:

Mục tiêu chính của việc áp dụng nghiên cứu sự kiện trong đề tài này là để xác định một cách định lượng xem một tin tức có tạo ra "lợi nhuận bất thường" (abnormal return) hay không. Lợi nhuận bất thường là phần lợi nhuận không thể giải thích được bởi các biến động chung của toàn thị trường. Bằng cách cô lập phần lợi nhuận này, chúng ta có thể đo lường chính xác hơn tác động riêng lẻ của tin tức, từ đó làm cơ sở vững chắc cho việc gán nhãn dữ liệu để huấn luyện mô hình AI.

2.2.2 Phương pháp gán nhãn dựa trên biến động giá trị tuyệt đối

Đây là phương pháp tiếp cận trực quan và đơn giản nhất để gán nhãn cảm xúc cho tin tức. Logic cơ bản là một tin tức tích cực sẽ khiến giá cổ phiếu tăng và ngược lại, một tin tức tiêu cực sẽ khiến giá giảm. Dựa trên logic này, một bộ quy tắc có thể được thiết lập như sau:

- Tích cực: Nếu tỷ lệ thay đổi giá đóng cửa của cổ phiếu vào ngày sau khi tin ra (T+1) so với ngày tin tức được tung ra (T) vượt qua một ngưỡng nhất định (Trong đề tài là 1%).
- Tiêu cực Nếu tỷ lệ giảm xuống dưới một ngưỡng âm (Trong đề tài là -1%).
- Trung tính: Các trường hợp còn lại.

Tuy nhiên, phương pháp này tồn tại một nhược điểm nghiêm trọng: nó không thể phân biệt được tác động của tin tức riêng lẻ khỏi xu hướng chung của toàn thị trường. Ví dụ, trong một ngày thị trường tăng trưởng mạnh mẽ (bull market), chỉ số VN-Index tăng 3%, một cổ phiếu chỉ tăng 1% thực chất đã có hiệu suất kém hơn so với thị trường. Mặc dù vậy, phương pháp này vẫn sẽ gán

nhãn "Tích cực" cho các tin tức liên quan đến cổ phiếu đó. Sự nhiễu loạn này (market noise) tạo ra các nhãn dữ liệu kém chất lượng, có thể khiến mô hình học máy học sai các mối tương quan. Điều này cho thấy sự cần thiết phải có một phương pháp tinh vi hơn, có khả năng điều chỉnh theo rủi ro thị trường.

2.2.3 Phương pháp gán nhãn dựa vào lợi nhuận bất thường

Để khắc phục hạn chế của phương pháp trên, đề tài áp dụng một cách tiếp cận có nền tảng tài chính vững chắc hơn, dựa trên việc đo lường lợi nhuận bất thường (abnormal return).

a. Tổng quan về mô hình định giá tài sản vốn (CAPM) và hệ số alpha

Mô hình Định giá Tài sản Vốn (Capital Asset Pricing Model - CAPM) là một trong những lý thuyết nền tảng của tài chính hiện đại, mô tả mối quan hệ giữa rủi ro và lợi nhuận kỳ vọng. Theo CAPM, lợi nhuận kỳ vọng của một tài sản được quyết định bởi tỷ suất sinh lời phi rủi ro và một khoản bù rủi ro dựa trên rủi ro hệ thống của tài sản đó. Rủi ro hệ thống này được đo lường bằng hệ số Beta (β), cho biết mức độ biến động của một cổ phiếu so với toàn bộ thị trường.

Lợi nhuận bất thường chính là phần chênh lệch giữa lợi nhuận thực tế mà một cổ phiếu đạt được và lợi nhuận kỳ vọng mà mô hình CAPM dự báo. Phần lợi nhuận này được cho là phản ánh tác động của các thông tin đặc thù của công ty, chẳng hạn như một bản tin vừa được công bố. Trong tài chính, hệ số Alpha (α) được sử dụng như một thước đo trực tiếp cho lợi nhuận bất thường này. Một Alpha dương cho thấy cổ phiếu hoạt động tốt hơn kỳ vọng sau khi đã điều chỉnh rủi ro, và ngược lại.

b. Công thức tính toán

Dựa trên mô hình CAPM, hệ số Alpha được tính như sau:

$$\alpha = R_p - R_f - \beta \times (R_m - R_f)$$

Trong đó:

- R_p: tỉ suất sinh lời thực tế của cổ phiếu.
- R_m: tỉ suất sinh lời của thị trường (Ví dụ: chỉ số VN-30).

- R_f: tỉ suất sinh lời phi rủi ro (Thường được lấy theo lãi suất của trái phiếu chính phủ dài hạn).
- β: hệ số Beta được tính bằng cách hồi quy tuyến tính lợi nhuận của cổ phiếu so với lợi nhuận của thị trường.

c. Ưu điểm

Phương pháp này vượt trội hơn hẳn so với biến động giá tuyệt đối vì nó đã lọc bỏ được nhiễu từ thị trường, giúp cô lập chính xác hơn tác động của tin tức. Điều này tạo ra một bộ nhãn "sạch" và đáng tin cậy hơn.

2.3 Phương pháp luận gán nhãn cho Giám sát Yếu (Weak Supervision)

2.3.1 Giới thiệu tổng quan về Giám sát yếu và Snorkel

Mặc dù phương pháp gán nhãn dựa trên hệ số Alpha đã giải quyết được vấn đề nhiễu từ thị trường, nó vẫn là một heuristic đơn lẻ, dựa trên một bộ giả định nhất định (ví dụ: mô hình CAPM mô tả chính xác lợi nhuận kỳ vọng). Trong thực tế, không phải lúc nào cũng có một quy tắc duy nhất đủ mạnh để bao quát mọi trường hợp. Để giải quyết thách thức này, một mô hình (paradigm) mới trong học máy đã ra đời, được gọi là Giám sát Yếu (Weak Supervision).

Khái niệm cốt lõi của Giám sát Yếu là thay vì tìm kiếm một quy tắc gán nhãn "hoàn hảo" duy nhất, chúng ta sẽ kết hợp nhiều nguồn tín hiệu "yếu" - tức là các quy tắc, heuristic không hoàn hảo, có nhiễu và thậm chí có thể mâu thuẫn với nhau - để gán nhãn cho dữ liệu một cách tự động và có quy mô lớn. Ý tưởng là bằng cách quan sát các mẫu đồng thuận và bất đồng giữa nhiều nguồn tín hiệu yếu, hệ thống có thể học cách tổng hợp chúng một cách thông minh để tạo ra các nhãn huấn luyện chất lượng cao, vượt qua độ chính xác của bất kỳ nguồn tín hiệu đơn lẻ nào.

Để triển khai phương pháp luận này một cách thực tế, Snorkel đã nổi lên như một framework mã nguồn mở phổ biến và mạnh mẽ. Snorkel cung cấp một bộ công cụ cho phép người dùng định nghĩa các nguồn tín hiệu yếu dưới dạng các hàm Python đơn giản, sau đó sử dụng một mô hình xác suất để tự

động khử nhiễu, cân bằng và kết hợp các tín hiệu này nhằm tạo ra một bộ dữ liệu huấn luyện lớn.

2.3.2 Các hàm gán nhãn

Trong Snorkel, các nguồn tín hiệu yếu được hiện thực hóa thông qua các Hàm Gán nhãn (Labeling Functions - LFs). Về cơ bản, một LF là một hàm Python nhận một điểm dữ liệu (ví dụ: một bài báo) làm đầu vào và "gán" cho một nhãn (Tích cực, Tiêu cực, Trung tính) hoặc gán nhãn trắng (Abstain) nếu quy tắc không được áp dụng. Việc phát triển một tập hợp đa dạng các LFs là chìa khóa của phương pháp này. Đối với bài toán phân tích cảm xúc tin tức tài chính, các loại LFs sau có thể được xây dựng:

- LFs dựa trên từ khóa: Đây là loại LF đơn giản nhất, sử dụng các bộ từ điển được xây dựng sẵn để tìm kiếm sự xuất hiện của các từ mang cảm xúc. Ví dụ, một LF có thể gán nhãn "Tích cực" nếu bài báo chứa các từ như "lợi nhuận", "tăng trưởng", "vượt kế hoạch", "kỷ lục", và bỏ phiếu "Tiêu cực" nếu chứa các từ như "thua lỗ", "suy giảm", "vi phạm", "thâm hụt".
- LFs dựa trên mẫu (Pattern-based): Các LF này sử dụng biểu thức chính quy (regular expressions) để phát hiện các mẫu câu hoặc cấu trúc ngữ pháp cụ thể mang ý nghĩa cảm xúc. Ví dụ, một LF có thể tìm kiếm mẫu "lợi nhuận [tăng/đạt] X tỷ đồng" để gán nhãn "Tích cực", hoặc "lỗ lũy kế Y tỷ đồng" để gán nhãn "Tiêu cực".
- LFs bao bọc Heuristic: Đây là một kỹ thuật quan trọng cho thấy khả năng tích hợp của Giám sát Yếu. Các phương pháp luận đã trình bày ở mục 2.2 (Biến động giá và Hệ số Alpha) có thể được "gói" lại thành các LFs riêng biệt. Mỗi heuristic sẽ trở thành một nguồn tín hiệu, một "chuyên gia" trong hệ thống, gán nhãn dựa trên kết quả tính toán của nó.
- LFs Giám sát Từ xa (Distant Supervision): Phương pháp này tận dụng các cơ sở kiến thức hoặc cơ sở dữ liệu có cấu trúc bên ngoài để gán nhãn. Ví dụ, chúng ta có thể xây dựng một cơ sở dữ liệu về

các sự kiện của công ty (ví dụ: danh sách các công ty bị Ủy ban Chứng khoán xử phạt, danh sách các công ty nhận giải thưởng "Doanh nghiệp của năm"). Một LF có thể tự động gán nhãn "Tiêu cực" cho các tin tức về một công ty vào ngày công ty đó bị xử phạt, hoặc "Tích cực" vào ngày nhận giải thưởng.

2.3.3 Mô hình nhãn

Sức mạnh thực sự của Snorkel nằm ở Mô hình Nhãn (Label Model). Sau khi áp dụng tất cả các LFs lên bộ dữ liệu chưa có nhãn, chúng ta thu được một ma trận các phiếu bầu. Label Model là một mô hình xác suất có khả năng học được độ chính xác và sự tương quan của từng LF mà không cần bất kỳ nhãn "thật" (ground truth) nào.

Mô hình này phân tích các mẫu đồng thuận và bất đồng giữa các LFs để suy ra LF nào đáng tin cậy hơn, LF nào có xu hướng bỏ phiếu trùng lặp với nhau. Dựa trên các thông tin học được này, Label Model sẽ tổng hợp các phiếu bầu một cách thông minh, có trọng số, để tạo ra một nhãn huấn luyện duy nhất dưới dạng xác suất (ví dụ: 85% Tích cực, 10% Trung tính, 5% Tiêu cực) cho mỗi điểm dữ liệu. Những nhãn xác suất này sau đó sẽ được sử dụng để huấn luyện một mô hình học sâu cuối cùng (như PhoBERT), giúp mô hình này học được các đặc trưng phức tạp và tổng quát hóa tốt hơn so với việc chỉ học từ các quy tắc rời rạc.

2.4 Lựa chọn và ứng dụng trong đề tài

Qua việc phân tích các phương pháp luận trên, có thể tóm tắt và so sánh chúng như sau:

- Biến động giá: Đơn giản nhất để triển khai nhưng tín hiệu bị nhiễu nặng bởi xu hướng chung của thị trường.
- Hệ số Alpha: Tốt hơn đáng kể, cung cấp một tín hiệu đã được lọc nhiễu và có cơ sở lý thuyết tài chính vững chắc.
- Giám sát Yếu: Mạnh mẽ và linh hoạt nhất, cho phép tổng hợp nhiều nguồn tín hiệu đa dạng để tạo ra các nhãn chất lượng cao ở quy mô lớn.

Dựa trên phân tích này, đề tài sẽ không chỉ lựa chọn một phương pháp duy nhất mà sẽ áp dụng một cách tiếp cận kết hợp để tận dụng thế mạnh của nhiều phương pháp. Cụ thể, phương pháp Giám sát Yếu sẽ được sử dụng làm khung phương pháp luận chính. Trong đó, hệ số Alpha sẽ được triển khai như một trong những Hàm Gán Nhãn (LF) cốt lõi và có trọng số cao, bên cạnh các LFs khác dựa trên từ khóa và mẫu câu. Cách tiếp cận này cho phép chúng ta kết hợp tín hiệu định lượng khách quan từ thị trường (Alpha) với các tín hiệu ngữ nghĩa từ nội dung văn bản, nhằm tạo ra một bộ dữ liệu huấn luyện toàn diện và chất lượng nhất để tinh chỉnh mô hình PhoBERT.

Quy tắc gán nhãn dựa trên ngưỡng Alpha sẽ được hiện thực hóa thành một LF như sau:

- Nếu alpha > 0.01: LF gán nhãn "Tích cực" (positive).
- Nếu alpha < -0.01: LF gán nhãn "Tiêu cực" (negative).
- Nếu -0.01 <= alpha <= 0.01: LF gán nhãn "Trung tính" (neutral).
- Nếu không tính được Alpha (do thiếu dữ liệu): Hàm gán nhãn sẽ gán nhãn trắng (Abstain), cho phép các hàm khác vẫn có thể gán nhãn cho mẫu dữ liệu này.

2.5 Ứng dụng các mô hình trí tuệ nhân tạo (AI) trong đề tài

Trong khuôn khổ của đề tài, nghiên cứu này tập trung vào việc tìm hiểu và ứng dụng sức mạnh tổng hợp của các loại mô hình tiên tiến:

- Mạng nơ-ron LSTM, chuyên dụng cho việc phân tích và dự báo dữ liệu chuỗi thời gian như giá cổ phiếu.
- Mô hình hồi quy Logistic (Logistic Regression), được sử dụng như một mô hình cơ sở để so sánh hiệu quả.
- Mô hình ngôn ngữ lớn PhoBERT, được sử dụng để khai thác thông tin chiều sâu từ dữ liệu văn bản phi cấu trúc như tin tức tài chính.

Sự kết hợp này nhằm mục đích xây dựng một phương pháp phân tích toàn diện, có khả năng nắm bắt cả các yếu tố định lượng và định tính, từ đó cung cấp một góc nhìn đa chiều và sâu sắc hơn về động lực của thị trường.

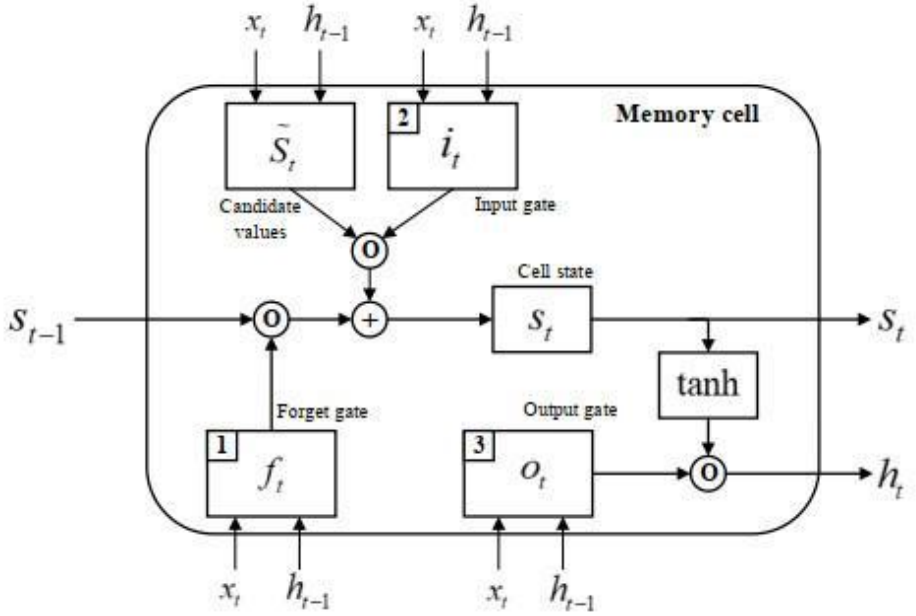
2.5.1 Mạng nơ-ron LSTM (Long Short-Term Memory)

a. Tổng quan

Mạng nơ-ron Bộ nhớ dài-ngắn hạn (Long Short-Term Memory - LSTM) là một phiên bản cải tiến và mở rộng của Mạng nơ-ron hồi quy (Recurrent Neural Network - RNN). LSTM được thiết kế đặc biệt để giải quyết một trong những thách thức lớn nhất của RNN truyền thống: vấn đề phụ thuộc xa (long-term dependencies) hay tiêu biến/bùng nổ gradient (vanishing/exploding gradient). Trong khi RNN gặp khó khăn trong việc ghi nhớ thông tin từ các bước thời gian ở quá xa, LSTM với kiến trúc độc đáo của mình có khả năng học và duy trì thông tin quan trọng trong một khoảng thời gian dài. Điều này làm cho LSTM trở thành một công cụ cực kỳ mạnh mẽ để xử lý và dự báo các loại dữ liệu có dạng chuỗi tuần tự như chuỗi thời gian giá cổ phiếu.

b. Cấu trúc của một ô LSTM

Sức mạnh của LSTM đến từ cấu trúc phức tạp bên trong mỗi ô nhớ (cell). Thay vì chỉ có một tầng mạng nơ-ron đơn giản như RNN, một ô LSTM bao gồm các thành phần tương tác với nhau một cách thông minh. Ý tưởng cốt lõi là trạng thái tế bào (cell state), thường được ví như một "băng chuyền" chạy dọc theo toàn bộ chuỗi, cho phép thông tin được truyền đi một cách gần như nguyên vẹn. Dòng thông tin này được điều khiển bởi ba cấu trúc gọi là **cổng (gates)**.



Hình 2.5.1. Mô hình LSTM

- **Cổng Quên (Forget Gate):** Đây là cổng đầu tiên, có nhiệm vụ quyết định thông tin nào từ trạng thái tế bào trước đó (C_{t-1}) cần được loại bỏ. Nó nhận đầu vào là h_{t-1} (đầu ra của ô trước) và x_t (đầu vào hiện tại), sau đó đưa qua một hàm sigmoid để tạo ra một số từ 0 đến 1 cho mỗi phần tử trong trạng thái tế bào. Giá trị 1 có nghĩa là "giữ lại hoàn toàn", trong khi 0 có nghĩa là "loại bỏ hoàn toàn".
- **Cổng Đầu vào (Input Gate):** Cổng này quyết định thông tin mới nào sẽ được lưu trữ vào trạng thái tế bào. Quá trình này gồm hai bước. Đầu tiên, một tầng sigmoid (cổng đầu vào) quyết định giá trị nào sẽ được cập nhật. Tiếp theo, một tầng tanh tạo ra một vector các giá trị mới ($\sim C_t$), có thể được thêm vào trạng thái. Hai kết quả này sau đó được kết hợp để tạo ra bản cập nhật cho trạng thái tế bào.
- **Cổng Đầu ra (Output Gate):** Cuối cùng, cổng này quyết định đầu ra của ô LSTM sẽ là gì. Đầu ra này sẽ dựa trên trạng thái tế bào đã được lọc. Đầu tiên, một tầng sigmoid quyết định phần nào của trạng thái tế bào sẽ được xuất ra. Sau đó, trạng thái tế bào được đưa qua hàm tanh (để đưa giá trị về khoảng từ -1 đến 1) và nhân với đầu ra

của tầng sigmoid để chỉ xuất ra những phần thông tin đã được quyết định.

c. Ứng dụng trong đề tài

Trong bối cảnh dự báo giá cổ phiếu, LSTM là một lựa chọn tự nhiên và hiệu quả. Dữ liệu giá cổ phiếu là một chuỗi thời gian, nơi giá của ngày hôm nay phụ thuộc vào giá của những ngày trước đó. Khả năng ghi nhớ các phụ thuộc dài hạn của LSTM cho phép nó không chỉ nhìn vào biến động của vài ngày gần nhất mà còn có thể nắm bắt các xu hướng và chu kỳ kéo dài hàng tuần hoặc hàng tháng, điều mà các mô hình đơn giản hơn có thể bỏ lỡ.

Nhiều nghiên cứu đã chứng minh sự phù hợp của LSTM cho thị trường chứng khoán Việt Nam. Các nghiên cứu điển hình về dự báo chỉ số VN-Index và các cổ phiếu trong nhóm VN-30 đã báo cáo độ chính xác cao, lên tới 93%, khi sử dụng LSTM kết hợp với các chỉ báo phân tích kỹ thuật. Trong đề tài này, LSTM sẽ được sử dụng để phân tích chuỗi dữ liệu định lượng (giá đóng cửa lịch sử, khối lượng giao dịch, các chỉ báo kỹ thuật như SMA, RSI) nhằm dự đoán xu hướng giá trong tương lai.

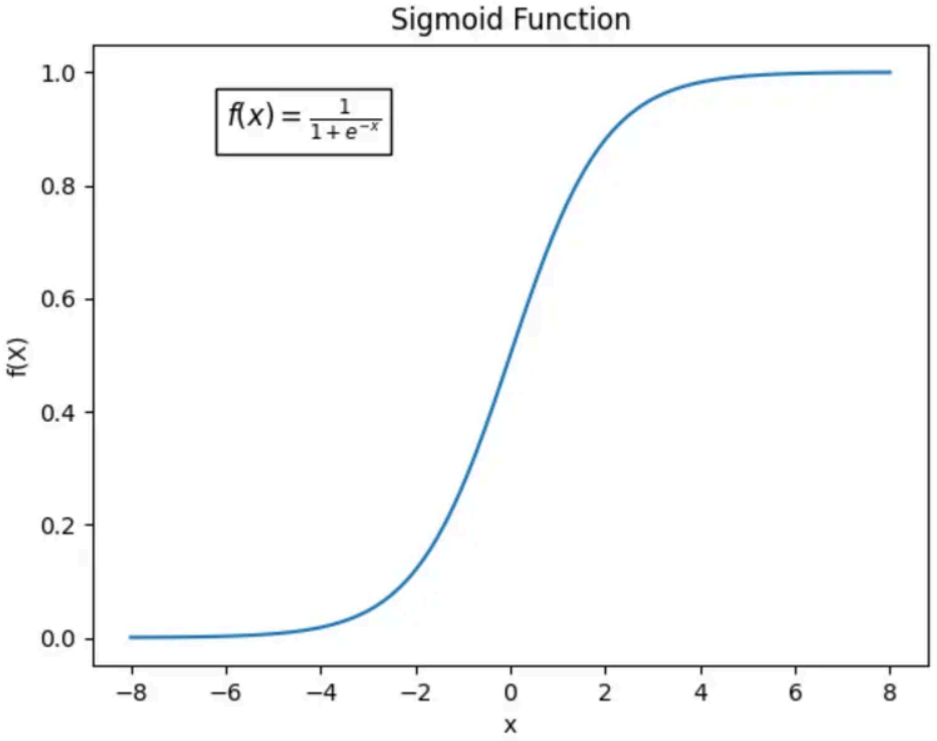
2.5.2 Mô hình hồi quy Logistic (Logistic Regression)

a. Tổng quan

Hồi quy Logistic là một mô hình thống kê và là một thuật toán học máy có giám sát cơ bản, được sử dụng chủ yếu cho các bài toán phân loại nhị phân (binary classification). Mục tiêu của nó là dự đoán xác suất một đầu ra rời rạc (ví dụ: Có/Không, Tăng/Giảm, 1/0) dựa trên một tập hợp các biến đầu vào độc lập. Mặc dù có tên gọi "hồi quy", nó lại được dùng cho bài toán phân loại. Điểm khác biệt chính so với Hồi quy Tuyến tính (Linear Regression) là Hồi quy Logistic cho ra một giá trị xác suất bị chặn trong khoảng , trong khi Hồi quy Tuyến tính dự đoán một giá trị liên tục có thể nằm ngoài khoảng này.

b. Kiến trúc

Kiến trúc của Hồi quy Logistic tương đối đơn giản và minh bạch. Cốt lõi của mô hình là hàm Sigmoid (còn gọi là hàm logistic), một hàm toán học có dạng đường cong hình chữ S.



Hình 2.5.2. Hàm sigmoid

Hàm Sigmoid nhận một giá trị đầu vào là tổ hợp tuyến tính của các biến độc lập (tương tự như trong hồi quy tuyến tính: $z=w^Tx+b$) và chuyển đổi nó thành một giá trị đầu ra trong khoảng (0, 1). Giá trị đầu ra này được diễn giải là xác suất để biến phụ thuộc nhận giá trị là 1 (ví dụ: xác suất cổ phiếu tăng giá).

Phương trình của mô hình có thể được viết như sau:

$$P(y = 1|x) = \sigma(z) = \frac{1}{1+e^{-z}}$$

Trong đó $z = w_1x_1 + w_2x_2 + ... + w_nx_n + b$

Các tham số w (trọng số) và b (hệ số chặn) được học từ dữ liệu huấn luyện thông qua các phương pháp tối ưu hóa như Maximum Likelihood Estimation. Dựa trên xác suất tính được, một ngưỡng quyết định (thường là 0.5)

được sử dụng để phân loại đầu ra cuối cùng. Nếu $P(y = 1|x) > 0.5$, dự đoán là 1; ngược lại là 1, dự đoán là 0.

c. Ứng dụng trong đề tài

Trong lĩnh vực tài chính, Hồi quy Logistic đã được áp dụng để dự báo các sự kiện có tính chất nhị phân như khả năng kiệt quệ tài chính của doanh nghiệp hay rủi ro tín dụng.

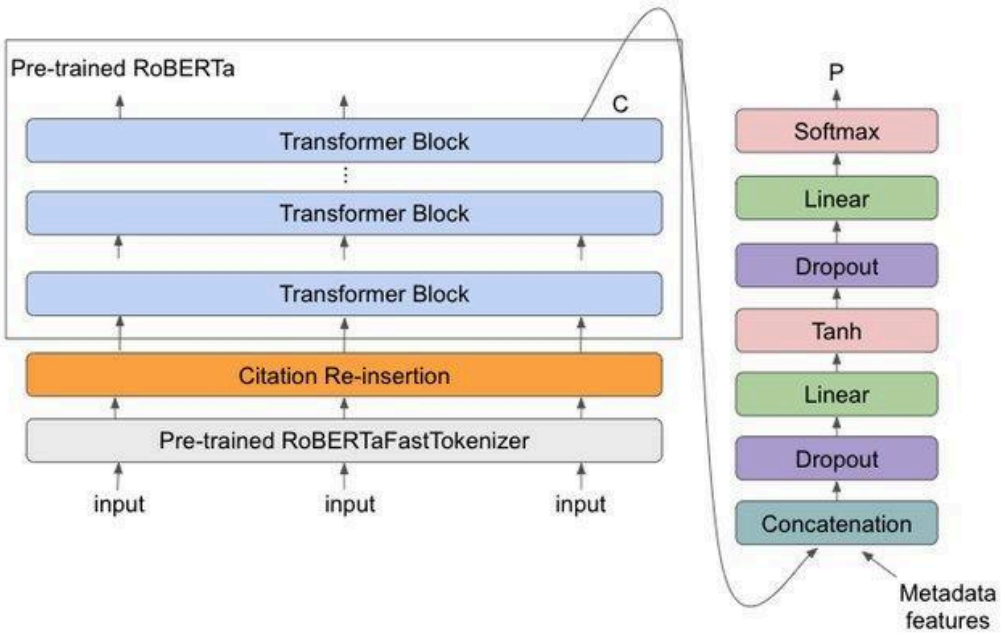
Trong đề tài này, vai trò chính của Hồi quy Logistic là làm **mô hình cơ sở (baseline model)** để đánh giá hiệu quả của các mô hình phức tạp hơn như LSTM và PhoBERT. Cụ thể, Hồi quy Logistic có thể được huấn luyện để dự đoán xu hướng giá cổ phiếu (ví dụ: "Tăng" hoặc "Giảm") dựa trên các đặc trưng đầu vào. Mặc dù độ chính xác của nó có thể không cao bằng các mô hình học sâu (các nghiên cứu cho thấy độ chính xác thường chỉ hơn 50% một chút) , việc so sánh kết quả của mô hình phức tạp với một mô hình đơn giản nhưng có cơ sở như Hồi quy Logistic là một bước đi khoa học quan trọng. Nó giúp định lượng mức độ cải thiện mà các kiến trúc tiên tiến mang lại, từ đó chứng minh giá trị thực tiễn của phương pháp đề xuất.

2.5.3 Mô hình ngôn ngữ tự nhiên PhoBERT

a. Tổng quan

PhoBERT là một mô hình ngôn ngữ lớn được tiền huấn luyện (pre-trained) và thiết kế chuyên biệt cho tiếng Việt. Được phát triển bởi VinAI Research, PhoBERT dựa trên kiến trúc RoBERTa (một phiên bản tối ưu của BERT) và được huấn luyện trên một kho ngữ liệu tiếng Việt khổng lồ (khoảng 20GB) bao gồm các bài báo và Wikipedia tiếng Việt. Sự ra đời của PhoBERT đã giải quyết vấn đề thiếu hụt các mô hình ngôn ngữ mạnh mẽ, hiệu quả cho tiếng Việt, và nhanh chóng trở thành một mô hình nền mạnh mẽ cho nhiều bài toán xử lý ngôn ngữ tự nhiên (NLP) trong tiếng Việt như phân loại văn bản, phân tích cảm xúc, và hỏi đáp.

b. Kiến trúc



Hình 2.5.3. Kiến trúc của Mô hình RoBERTA

Kiến trúc của PhoBERT kế thừa từ RoBERTA, do đó nó cũng là một mô hình dựa trên kiến trúc Transformer. Tương tự BERT, PhoBERT cũng có hai phiên bản:

- PhoBERT-base: Gồm 12 lớp (layer) Transformer, 12 đầu chú ý (attention head) và 135 triệu tham số.
- PhoBERT-large: Lớn hơn với 24 lớp Transformer, 16 đầu chú ý và 370 triệu tham số.

Một điểm khác biệt quan trọng trong phương pháp huấn luyện của RoBERTa (PhoBERT nói riêng) so với BERT gốc là nó loại bỏ nhiệm vụ dự đoán câu tiếp theo (Next Sentence Prediction - NSP) và chỉ tập trung vào nhiệm vụ Mô hình hóa Ngôn ngữ Che dấu (Masked Language Model - MLM). Về xử lý đầu vào, do đặc thù của tiếng Việt, văn bản cần được "tách từ" (word segmentation) bằng một công cụ như VnCoreNLP trước khi được đưa vào bộ mã hóa BPE (Byte-Pair Encoding) của PhoBERT.

c. Ứng dụng trong đề tài

Ứng dụng chính và mạnh mẽ nhất của PhoBERT trong đề tài này là phân tích cảm xúc (sentiment analysis) từ các tin tức tài chính bằng tiếng Việt. Thị trường chứng khoán rất nhạy cảm với thông tin, và cảm xúc của nhà đầu tư thể hiện qua tin tức có thể là một yếu tố dự báo quan trọng.

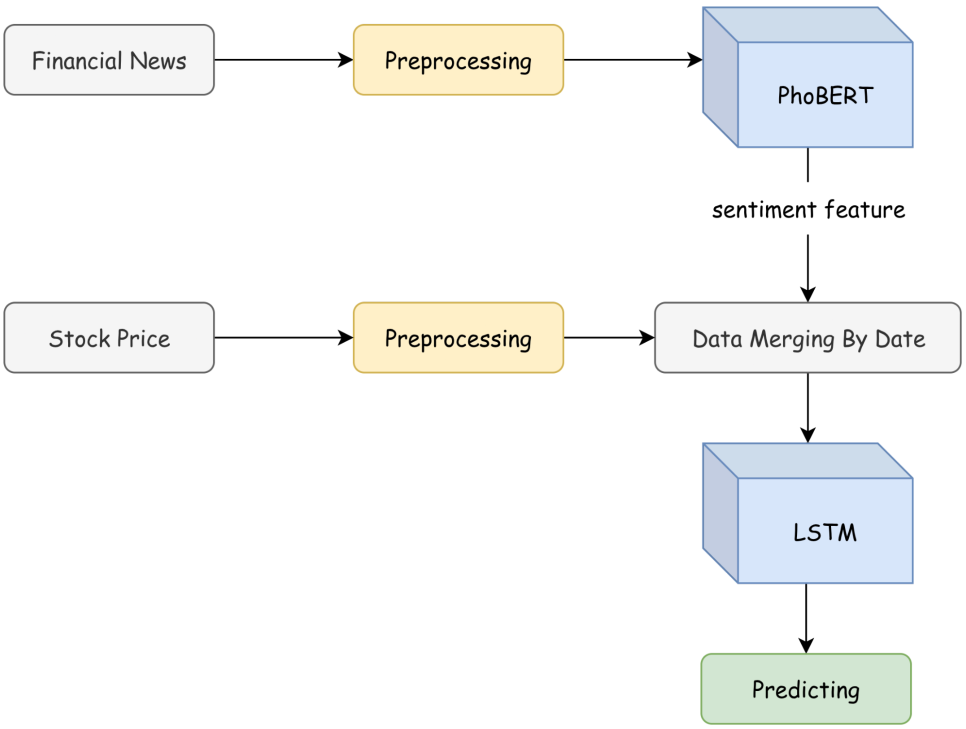
Cụ thể, PhoBERT sẽ được tinh chỉnh (fine-tuned) trên một tập dữ liệu tin tức tài chính đã được gán nhãn (Tích cực, Tiêu cực, Trung tính) bằng phương pháp luận ở mục 2.1. Sau khi huấn luyện, mô hình có khả năng nhận một tiêu đề hoặc một đoạn tin tức mới và đưa ra dự đoán về cảm xúc của nó.

Điểm số cảm xúc này sau đó trở thành một đặc trưng định tính vô cùng giá trị. Đặc trưng này sẽ được kết hợp với các đặc trưng định lượng để làm đầu vào cho mô hình dự báo chính (như LSTM). Việc tích hợp cảm xúc từ tin tức đã được chứng minh là giúp cải thiện đáng kể độ chính xác của các mô hình dự báo giá cổ phiếu. Bằng cách này, PhoBERT giúp mô hình không chỉ "nhìn" vào các con số mà còn "hiểu" được bối cảnh và tâm lý thị trường đằng sau những con số đó, tạo ra một hệ thống dự báo toàn diện và mạnh mẽ hơn.

CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT

3.1 Kiến trúc tổng thể hệ thống

Hệ thống dự đoán giá cổ phiếu được xây dựng theo kiến trúc kết hợp giữa hai mô hình: PhoBERT và LSTM. Trước hết, dữ liệu văn bản từ tin tức và bài viết tài chính được thu thập, tiền xử lý và fine-tune trên PhoBERT để trích xuất đặc trưng cảm xúc. Sau đó, các đặc trưng này được đồng bộ và kết hợp với dữ liệu giá cổ phiếu lịch sử, tạo thành chuỗi thời gian đầu vào cho mô hình LSTM. Mạng LSTM học mối quan hệ giữa biến động giá và ảnh hưởng cảm xúc thị trường để dự đoán giá đóng cửa tương lai. Kết quả dự báo được đánh giá bằng các chỉ số sai số nhằm hiệu chỉnh và tối ưu mô hình.



Hình 3.1.1. Kiến trúc tổng thể của hệ thống

Để xây dựng một mô hình kết hợp hiệu quả giữa LSTM và PhoBert, việc xây dựng và xử lý hai nguồn dữ liệu riêng biệt là dữ liệu chuỗi thời gian về giá và dữ liệu văn bản tài chính là yêu cầu tiên quyết. Mỗi loại dữ liệu có những đặc thù riêng và đòi hỏi các phương pháp tiền xử lý khác nhau để có thể khai thác tối đa thông tin hữu ích cho mô hình dự báo.

3.2 Dữ liệu

3.2.1 Dữ liệu chuỗi thời gian về giá

Dữ liệu giá cổ phiếu được thu thập từ nguồn dữ liệu chính thức của thị trường chứng khoán Việt Nam thông qua thư viện vnstock. Đây là dữ liệu có cấu trúc, phản ánh chính xác các đặc điểm giao dịch của từng mã cổ phiếu theo thời gian. Quy trình thu thập và xử lý dữ liệu được thực hiện qua ba giai đoạn chính như sau:

a. Giai đoạn thu thập (Fetching).

Giai đoạn này chịu trách nhiệm tải dữ liệu lịch sử giao dịch của các mã cổ phiếu. Hệ thống sử dụng **API của vnstock** để tự động truy vấn và tải về thông tin giá trong khoảng thời gian từ năm 2010 đến năm 2025.
Dữ liệu được thu thập gồm các trường:

- Ngày giao dịch (Time)
- Giá mở cửa (Open)
- Giá đóng cửa (Close)
- Giá cao nhất (High)
- Giá thấp nhất (Low)
- Khối lượng giao dịch (Volume)

Mỗi lần tải, dữ liệu được lưu dưới dạng bảng CSV (DataFrame) để tiện cho bước xử lý tiếp theo.

Time	Open	High	Low	Close	Volume
2025-06-23	56.7	56.9	56.5	56.6	2480600
2025-06-24	56.9	57.1	56.6	56.6	3507600

Bảng 3.2.1. Mẫu dữ liệu giá

b. Giai đoạn tiền xử lý (Preprocessing)

Mục tiêu của giai đoạn này là chuẩn hóa và làm sạch dữ liệu để đảm bảo tính toàn vẹn trước khi đưa vào mô hình dự báo.

- Kiểm tra dữ liệu thiếu và loại bỏ ngoại lệ: Các phiên giao dịch thiếu dữ liệu hoặc có giá trị bất thường được xác định và xử lý (lọc bỏ hoặc nội suy).
- Chuyển đổi định dạng thời gian: Cột ngày giao dịch được chuyển thành định dạng datetime chuẩn.
- Chuẩn hóa dữ liệu: Các cột giá và khối lượng được áp dụng Min-Max Scaling về [0,1] để đồng nhất thang đo.
- Tạo đặc trưng bổ sung: Hệ thống tính toán thêm một số đặc trưng kỹ thuật như tỷ lệ biến động giá (volatility) và phần trăm thay đổi so với phiên trước.

c. Giai đoạn khớp nối dữ liệu cảm xúc (Alignment with Sentiment Data).

Sau khi dữ liệu giá được làm sạch và chuẩn hóa, bước tiếp theo là khớp nối với dữ liệu cảm xúc thị trường thu thập từ tin tức tài chính.

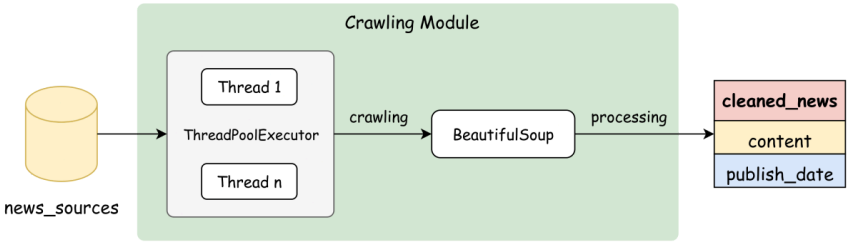
- Xác định mốc thời gian: Mỗi phiên giao dịch được đối chiếu với các bài báo đã được gán nhãn cảm xúc, dựa trên ngày đăng tin.
- Gán chỉ số cảm xúc: Đối với mỗi ngày giao dịch, một chỉ số cảm xúc tổng hợp được tính trung bình từ các bài báo liên quan và gán thêm thành một trường dữ liệu mới.
- Tạo tập dữ liệu cuối cùng: Bộ dữ liệu đầu ra bao gồm tất cả thông tin về giá và khối lượng, kèm theo nhãn cảm xúc cho mỗi phiên giao dịch. Đây chính là đầu vào cho mô hình kết hợp LSTM trong giai đoạn huấn luyện và dự báo.

3.2.2 Dữ liệu văn bản (Cảm xúc)

Dữ liệu văn bản được thu thập từ các nguồn báo điện tử uy tín về tài chính như CafeF và Báo Đầu Tư. Đây là nguồn thông tin phi cấu trúc, chứa đựng các yếu tố về quan điểm, tin tức, và sự kiện có khả năng ảnh hưởng đến tâm lý nhà đầu tư và giá cổ phiếu. Quy trình xử lý nguồn dữ liệu này được thực hiện qua ba giai đoạn chính như sau:

a. Giai đoạn thu thập (Crawling)

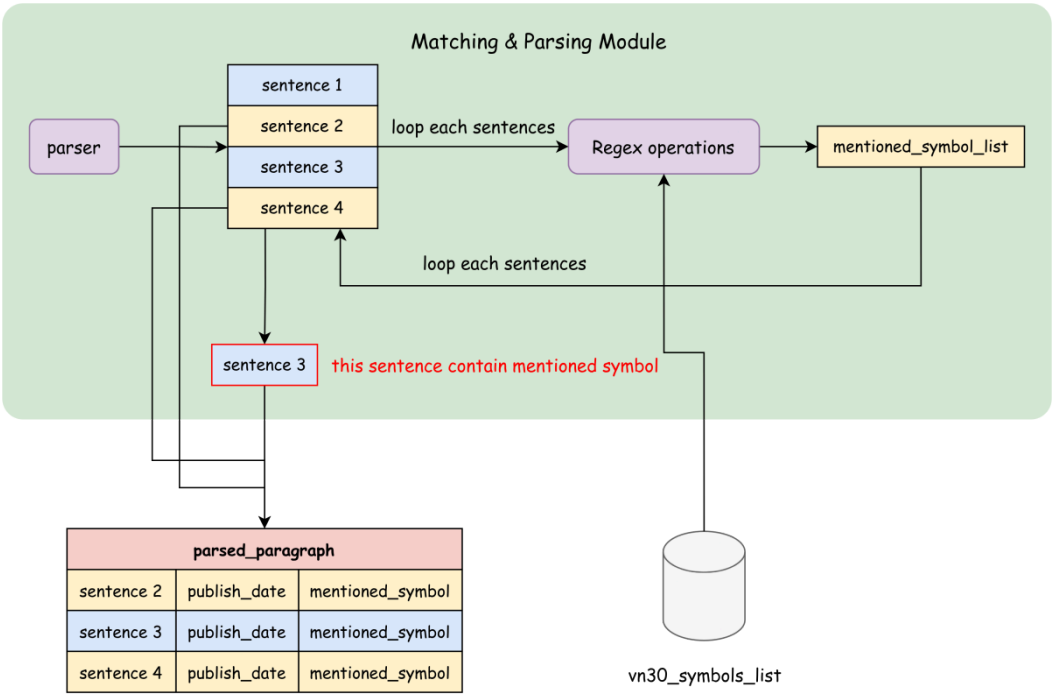
Giai đoạn này chịu trách nhiệm thu thập tự động nội dung các bài báo. Để tối ưu hóa tốc độ, hệ thống sử dụng kỹ thuật xử lý đa luồng (ThreadPoolExecutor) để thực hiện nhiều yêu cầu đồng thời. Thư viện BeautifulSoup được dùng để phân tích cú pháp HTML, bóc tách các thông tin quan trọng như ngày đăng tải và nội dung chính, đồng thời loại bỏ các thành phần nhiễu như thẻ HTML và các ký tự không cần thiết.



Hình 3.2.1. Quy trình thu thập và xử lý văn bản thô

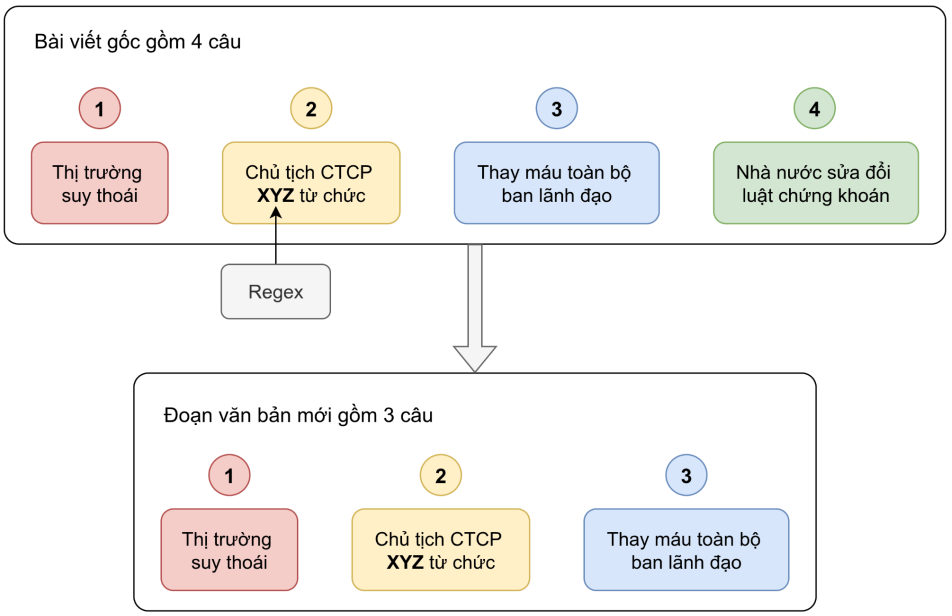
b. Giai đoạn khớp mã và phân tích cú pháp (Matching & Parsing)

Mục tiêu của giai đoạn này là xác định các đoạn văn bản có liên quan trực tiếp đến một mã cổ phiếu cụ thể.



Hình 3.2.4. Quy trình phân tách câu và khớp mã

- Về Regex (Regular Expression): Regex được thiết kế để nhận diện các mã cổ phiếu chính xác có trong danh sách VN30, tránh nhầm lẫn với các từ ngữ thông thường có thể trùng lặp."
- Về cửa sổ ngữ cảnh: "Kích thước cửa sổ ngữ cảnh gồm câu hiện tại (n), câu trước (n-1) và câu sau (n+1) được lựa chọn nhằm đảm bảo thu thập đủ thông tin ngữ cảnh để phân tích cảm xúc một cách chính xác, đồng thời hạn chế việc đưa vào quá nhiều thông tin không liên quan có thể gây nhiễu cho mô hình."



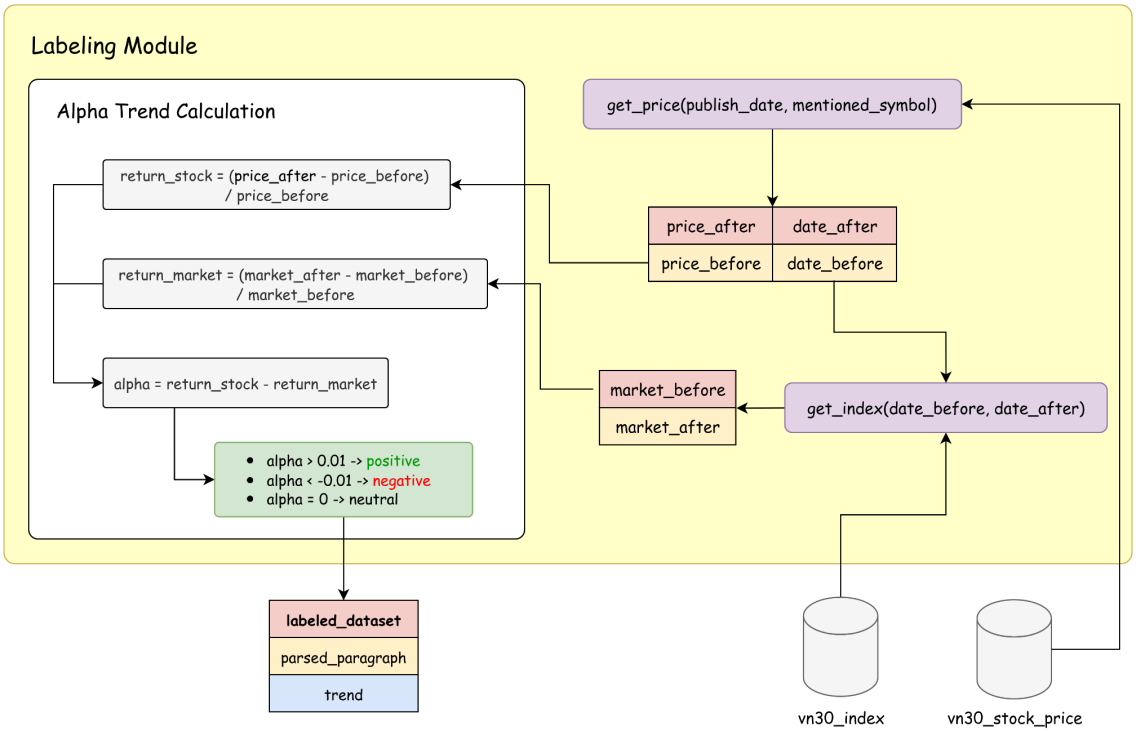
Hình 3.2.5. Quy trình chi tiết khớp mã mà ghép câu

- Về VnCoreNLP: "Việc sử dụng VnCoreNLP giúp xử lý hiệu quả các trường hợp từ ghép, cụm từ cố định trong tiếng Việt, ví dụ như 'chứng khoán' thay vì 'chứng' và 'khoán', hay 'ngân hàng' thay vì 'ngân' và 'hàng', từ đó cung cấp đầu vào chất lượng cao hơn cho mô hình PhoBERT."

c. Giai đoạn gán nhãn (Labeling).

Đây là giai đoạn cốt lõi để tạo ra nhãn cho bộ dữ liệu văn bản tài chính, một yêu cầu tiên quyết cho việc huấn luyện mô hình phân loại cảm xúc. Thay vì gán nhãn thủ công tốn kém và mang tính chủ quan, chúng tôi áp dụng một

phương pháp tự động, có quy mô, dựa trên hiệu suất thị trường kết hợp với các kỹ thuật giám sát yếu để đảm bảo tính khách quan.



Hình 3.2.6. Quy trình gán nhãn

Chiến lược tổng thể của chúng tôi là sử dụng phương pháp luận Giám sát Yếu (Weak Supervision), đã được trình bày chi tiết ở Chương 2 và được triển khai bằng framework Snorkel. Cách tiếp cận này cho phép chúng tôi kết hợp nhiều nguồn tín hiệu "yếu"—tức các quy tắc, heuristic không hoàn hảo—để tạo ra một bộ nhãn huấn luyện chất lượng cao. Cụ thể, chúng tôi đã xây dựng các **Hàm Gán nhãn (Labeling Functions - LFs)** dưới dạng các hàm Python, mỗi hàm đại diện cho một quy tắc hoặc một heuristic riêng biệt.

Các hàm gán nhãn chính được chúng tôi triển khai bao gồm:

- **Hàm gán nhãn dựa trên Lợi nhuận bất thường:** Đây là LF cốt lõi và có trọng số cao nhất trong hệ thống của chúng tôi, do khả năng lọc bỏ được nhiễu từ xu hướng chung của thị trường. Dựa trên cơ sở lý thuyết về Mô hình định giá tài sản vốn (CAPM) , chúng tôi "gói" quy tắc tính toán hệ số Alpha thành một LF. Quy tắc cụ thể được áp dụng như sau:

- Gán nhãn Tích cực (positive) nếu alpha > 0.01.
- Gán nhãn Tiêu cực (negative) nếu alpha < −0.01.
- Gán nhãn Trung tính (neutral) nếu −0.01 ≤ alpha ≤0.01.
- Gán nhãn trắng (Abstain) nếu không thể tính được Alpha để cho phép các LF khác gán nhãn.

- **Hàm gán nhãn dựa trên Biến động giá tuyệt đối:** Mặc dù phương pháp này tồn tại nhược điểm là không phân biệt được tác động của tin tức với xu hướng chung của thị trường, nó vẫn cung cấp một tín hiệu hữu ích. Chúng tôi đã triển khai nó như một LF (Label Function) bổ trợ với quy tắc:
  - Gán nhãn Tích cực nếu giá cổ phiếu tăng hơn 1% vào ngày T+1.
  - Gán nhãn Tiêu cực nếu giá cổ phiếu giảm hơn 1% vào ngày T+1.
- **Các hàm gán nhãn dựa trên Từ khóa và Mẫu câu:** Để khai thác tín hiệu trực tiếp từ ngữ nghĩa văn bản, chúng tôi đã xây dựng các LFs dựa trên các bộ từ điển tài chính. Ví dụ, một LF (Label Function) có thể gán nhãn **Tích cực** nếu bài báo chứa các từ như “lợi nhuận”, "tăng trưởng", "vượt kế hoạch", và gán nhãn **Tiêu cực** nếu chứa các từ như "thua lỗ", "suy giảm", "vi phạm". Tương tự, các biểu thức chính quy được dùng để phát hiện các mẫu câu cụ thể như "lợi nhuận đạt X tỷ đồng" để gán nhãn "Tích cực".

Sau khi áp dụng tất cả các LFs lên bộ dữ liệu thô, chúng tôi sử dụng **Mô hình Nhãn (Label Model)** của Snorkel. Mô hình này có khả năng học được độ chính xác và sự tương quan của từng LF (Label Function) mà không cần đến dữ liệu gán nhãn thật (ground truth). Nó tổng hợp các nhãn từ LFs (Label Functions) một cách thông minh để tạo ra một nhãn huấn luyện duy nhất dưới dạng xác suất cho mỗi mẫu dữ liệu. Bộ dữ liệu với các nhãn xác suất này chính là đầu vào cuối cùng để tinh chỉnh (fine-tune) mô hình PhoBERT, giúp mô hình học được các đặc trưng cảm xúc phức tạp và tổng quát hóa tốt hơn.

d. Tổng hợp quy trình thu thập và xử lý dữ liệu như sau:

Bảng 3.1.7. Tổng quan về quy trình xử lý dữ liệu

3.3 Mô hình phân loại cảm xúc

3.3.1 Mục tiêu của mô hình

Mục tiêu chính là tự động phân tích và lượng hóa cảm tính thể hiện trong các bài báo, tin tức tài chính liên quan đến các mã cổ phiếu. Cụ thể, mô hình sẽ phân loại từng đoạn văn bản thành ba loại cảm xúc chính: Tích cực (Positive), Tiêu cực (Negative), hoặc Trung tính (Neutral).

3.3.2 Kiến trúc mô hình PhoBERT sử dụng trong đề tài

Trong đề tài này, chúng tôi sẽ sử dụng phiên bản PhoBERT-base. Thay vì sử dụng PhoBERT nguyên bản, mô hình sẽ được tinh chỉnh (fine-tuned) trên tập dữ liệu văn bản tài chính đã được gán nhãn bằng phương pháp Giám sát Yếu (Weak Supervision) như đã mô tả chi tiết trong mục 3.2.2. Quá trình tinh chỉnh này giúp PhoBERT học được các đặc thù ngôn ngữ và thuật ngữ chuyên ngành tài chính, từ đó nâng cao hiệu quả phân loại cảm xúc trong lĩnh vực này.

3.3.3 Đầu ra và Đầu vào của mô hình

a. Đầu vào:

Mô hình nhận đầu vào là các đoạn văn bản tin tức tài chính đã được tiền xử lý. Các đoạn văn bản này bao gồm câu chứa mã cổ phiếu được quan tâm và các câu liền kề (trước và sau) để đảm bảo ngữ cảnh đầy đủ. Trước khi đưa vào PhoBERT, văn bản sẽ được mã hóa thành dạng token IDs phù hợp với yêu cầu của mô hình.

b. Đầu ra:

Đầu ra của mô hình là một nhãn cảm xúc đã được phân loại (Tích cực, Tiêu cực, hoặc Trung tính) cho mỗi đoạn văn bản đầu vào. Ngoài ra, mô hình cũng có thể cung cấp phân phối xác suất tương ứng cho từng nhãn, cho phép đánh giá mức độ tin cậy của dự đoán cảm xúc.

3.3.4 Vai trò trong hệ thống tổng thể

Mô hình phân loại cảm xúc PhoBERT đóng vai trò là cầu nối giữa dữ liệu văn bản phi cấu trúc và mô hình dự đoán giá định lượng. Kết quả phân loại cảm xúc từ PhoBERT sẽ được chuyển đổi thành một chỉ số cảm xúc (sentiment score) hoặc vector cảm xúc. Chỉ số này sau đó được tích hợp làm một đặc trưng đầu vào quan trọng (feature) cho mô hình dự đoán giá cổ phiếu LSTM (được trình bày trong mục 3.4). Việc kết hợp yếu tố cảm xúc thị trường vào mô hình dự đoán giá giúp hệ thống nâng cao khả năng dự báo và cung cấp thông tin hữu ích cho nhà đầu tư bằng góc nhìn toàn cảnh thị trường.

3.4 Mô hình dự đoán giá cổ phiếu

3.4.1 Mục tiêu của mô hình

Mục tiêu của mô hình là dự đoán giá đóng cửa của cổ phiếu tại bước thời gian kế tiếp (t+1), dựa trên chuỗi dữ liệu lịch sử giá và cảm xúc trong n ngày trước đó. Bài toán được định nghĩa là một bài toán hồi quy, trong đó:

- Đầu vào: Chuỗi thời gian gồm các đặc trưng định lượng (giá, khối lượng) và đặc trưng định tính (chỉ số cảm xúc).
- Đầu ra: Giá trị liên tục dự đoán của giá đóng cửa ngày tiếp theo.

3.4.2 Kiến trúc mô hình LSTM

Mô hình được xây dựng theo kiến trúc nhiều lớp, bao gồm các thành phần chính:

- Lớp LSTM đầu tiên: Với 50 đơn vị ẩn, nhận tensor đầu vào kích thước (batch_size, time_steps, num_features) và trả về chuỗi đầu ra cho tất cả bước thời gian (return_sequences=True).

- Lớp LSTM thứ hai: Cũng gồm 50 đơn vị ẩn, tiếp nhận chuỗi đầu ra từ lớp trước và chỉ giữ lại trạng thái ẩn cuối cùng làm đại diện cho toàn bộ chuỗi.
- Lớp Dense: Một neuron duy nhất dự đoán giá đóng cửa chuẩn hóa.

3.4.3 Quy trình huấn luyện mô hình

Tập dữ liệu sẽ được phân chia thành tập huấn luyện, tập kiểm định và tập thử nghiệm theo trình tự thời gian để đảm bảo tính thực tế của quá trình đánh giá.

- Hàm mất mát được sử dụng cho bài toán hồi quy này là **Mean Squared Error (MSE)**, đo lường sai số bình phương trung bình giữa giá dự đoán và giá thực tế.
- Bộ tối ưu hóa được lựa chọn là **Adam**, một thuật toán tối ưu phổ biến và hiệu quả cho việc huấn luyện mạng nơ-ron.
- Áp dụng chiến lược dừng sớm (**EarlyStopping**) nếu hàm mất mát không cải thiện sau 3 epoch liên tiếp và khôi phục trọng số tốt nhất.

3.4.4 Đánh giá và lưu mô hình

Hiệu suất của mô hình được đánh giá trên tập huấn luyện thông qua chỉ số Root Mean Square Error (RMSE), phản ánh mức độ sai số trung bình giữa giá dự đoán và giá thực tế.

Sau khi huấn luyện, toàn bộ trọng số và kiến trúc mô hình được lưu trữ dưới dạng tệp định dạng HDF5 để thuận tiện cho việc tái sử dụng hoặc triển khai trong các hệ thống dự báo sau này. Đồng thời, các tham số chuẩn hóa của bộ biến đổi Min-Max (bao gồm mảng giá trị tối thiểu và mảng tỷ lệ biến đổi) cũng được lưu thành các tệp dữ liệu số hóa riêng biệt. Việc lưu đồng thời cả mô hình và thông số scaler cho phép quy trình dự báo trong tương lai có thể giải chuẩn hóa kết quả dự đoán về thang đo gốc, đảm bảo tính nhất quán và chính xác khi áp dụng trên dữ liệu thực tế.

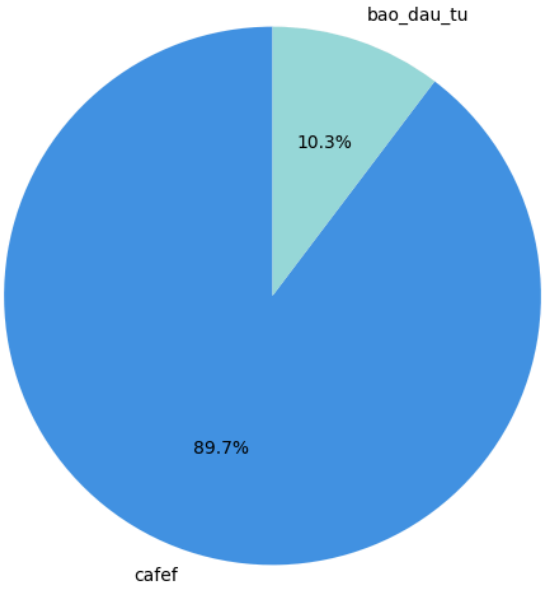
CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM

4.1 Dữ liệu

4.1.1 Dữ liệu văn bản (Cảm xúc)

a. Dữ liệu thô:

Tại giai đoạn thu thập dữ liệu, nhóm nghiên cứu đã thu thập được khoảng **80000 bài báo**, bài viết, ý kiến chuyên gia và nhận định thị trường trong thời gian khoảng **6000 giây** (với 20 luồng) từ 2 nguồn là Báo điện tử và Cafef (tỉ lệ khoảng 1:9), kích thước thô là **327 MB**.



Hình 4.1.1. Tỷ lệ phần trăm các nguồn thu thập dữ liệu

Mẫu dữ liệu sau khi được thu thập và tiền xử lý các thành phần thừa, dấu câu sẽ được trình bày trong bảng 4.1.1. dưới đây

publish_date	content
16:25:00 - 19/06/2025	7 DN lớn nhất nước đều ở Hà Nội, nhiều tỉnh không có DN nào trị giá nghìn tỷ …
11:20:00 - 10/06/2020	Công ty cổ phần FPT đạt 173.760 tỷ đồng và Ngân hàng TMCP Quân đội (MB) đạt 156.828 tỷ đồng …

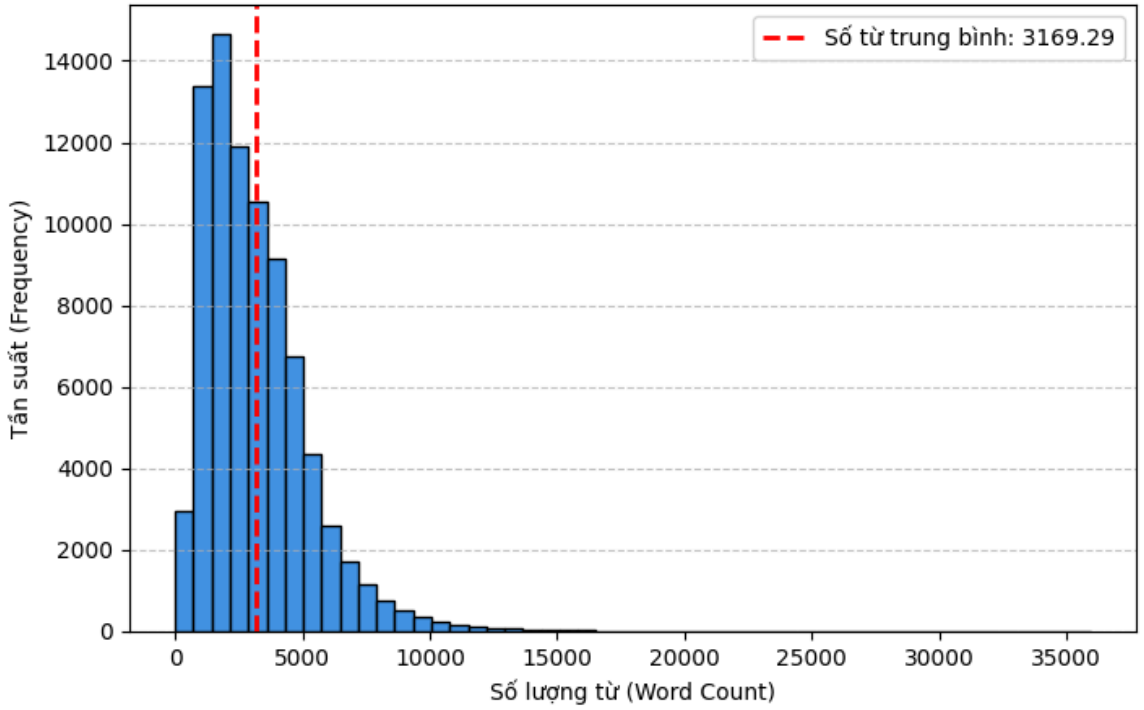
Bảng 4.1.1. Mẫu dữ liệu thô sau khi thu thập

với mỗi bài có độ dài dao động từ **500** chữ tới **30,000** chữ.

Độ dài mẫu (từ)	Tổng số mẫu (bài)	Tỉ lệ (%)
dưới 499	1240	1.52
500 - 999	6110	7.48
1000 - 1999	20616	25.25
2000 - 2999	16737	20.50
3000 - 3999	14021	17.18
4000 - 4999	10302	12.62
trên 5000	12607	15.44

Bảng 4.1.2. Phân phối giữa độ dài và tỉ lệ phần trăm của mẫu

Hình 4.1.2 dưới đây thể hiện giữa mối quan hệ giữa tần suất với độ dài của bài viết và độ dài trung bình của một mẫu dữ liệu.



Hình 4.1.2. Phân phối giữa độ dài và tần suất của mẫu.

b. Dữ liệu được xử lý

Tại giai đoạn khớp mã, từ một kho dữ liệu chứa các mã cổ phiếu thuộc rổ VN30 (ACB, BVH, SSI, VCB, …), nhóm nghiên cứu cài đặt một Regex để nhận dạng các mã cổ phiếu được nhắc đến trong bài viết. Mẫu dữ liệu sau khi được nhận dạng mã như sau:

Danh sách các mã cổ phiếu trong bài	Bài viết
VCB	Các cổ phiếu bất động sản, xây dựng sau vài phiên bị bán mạnh gần đây cũng hồi phục trở lại với nhiều mã tăng điểm như CTD, DXG, FCN, FLC, HBC, KBC, SJS, ROS, PHC, **VCB.**
SSI, ACB	Thị trường hồi phục, nhóm chứng khoán theo đó cũng thu hẹp đà giảm, đảo chiều ngoạn mục trong sắc xanh là VCI và FTS, còn các mã khác giảm nhẹ như VND giảm 0,9%; **SSI** giảm 1,6%, HCM, VIX, CTS giảm 2% và **ACB** tăng 1,5%,

Bảng 4.1.3. Mẫu dữ liệu khi nhận diện các mã cổ phiếu

Ở giai đoạn trích xuất ngữ cảnh, chúng tôi sẽ duyệt qua từng câu trong bài viết. Đối với mỗi câu chứa mã cổ phiếu được nhận diện, chúng tôi sẽ thu thập thêm câu liền kề phía trước và phía sau để tạo thành đoạn văn bản với ngữ cảnh đầy đủ. Từ đoạn văn bản đã được tạo, tiến hành tách từ và thu được mẫu dữ liệu sau:

Văn bản được tách từ (Segmented Context)
Các **cổ_phiếu bất_động_sản**, **xây_dựng** sau vài phiên bị bán mạnh **gần_đây** cũng **hồi_phục** trở lại với nhiều mã tăng điểm như CTD, DXG, FCN, FLC, HBC, KBC, SJS, ROS, PHC, **VCB.**
Thị trường **hồi_phục**, nhóm chứng_khoán theo đó cũng thu_hẹp đà giảm, đảo chiều **ngoạn_mục** trong sắc xanh là VCI và FTS, còn các mã khác giảm nhẹ như VND giảm 0,9%; **SSI** giảm 1,6%, HCM, VIX, CTS giảm 2% và **ACB** tăng 1,5%,

Bảng 4.1.4. Ví dụ về tách từ

c. Dữ liệu được gán nhãn

Tại giai đoạn gán nhãn bằng các phương pháp gán nhãn đã nêu ở **3.2.2.c Giai đoạn gán nhãn,** thu được mẫu dữ liệu sau và đã sẵn sàng đưa vào huấn luyện mô hình.

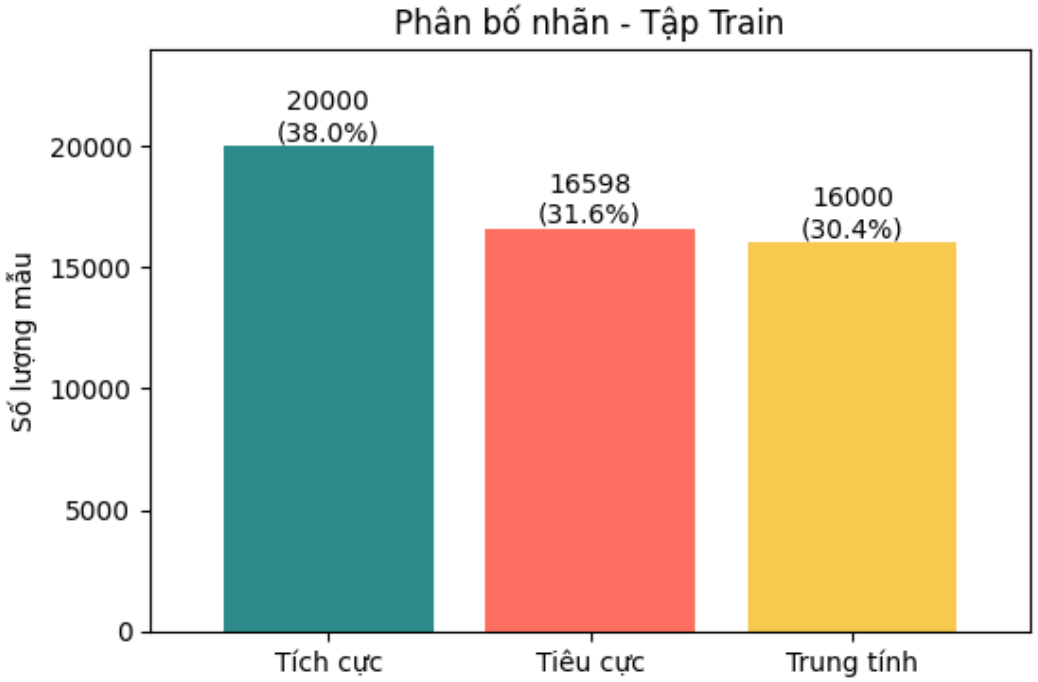
segmented_context	vol_label	alpha_label	symbol
KDH giảm 2,5% , DXS giảm 1,6% , HDG giảm 1,16% . Ở chiều ngược_lại , MSN , GAS , MBB , CTG hay HPG đóng vai_trò nâng_đỡ thị_trường chung . Cổ_phiếu Masan (MSN) tăng đến 2,4% và là mã đóng_góp tích_cực nhất cho VN-Index với 0,6 điểm	0	0	MSN
Dù_vậy , Vingroup và Vinhomes vẫn tiếp_tục nằm trong top 3 tổ_chức niêm_yết có quy_mô vốn hoá lớn nhất trên thị_trường , chỉ đứng sau Vietcombank . Cùng đó , nhiều cổ_phiếu vốn hoá lớn khác như VCB , GVR , BID , VNM	1	1	GVR
Trong đó , dòng vốn này bán ròng mạnh nhất mã FPT với 168 tỷ đồng . HPG và STB đều bị bán ròng hơn 100 tỷ đồng . Chiều ngược_lại , GEX được mua ròng mạnh nhất với 252 tỷ đồng	2	1	STB

Bảng 4.1.5. Mẫu dữ liệu sẵn sàng huấn luyện

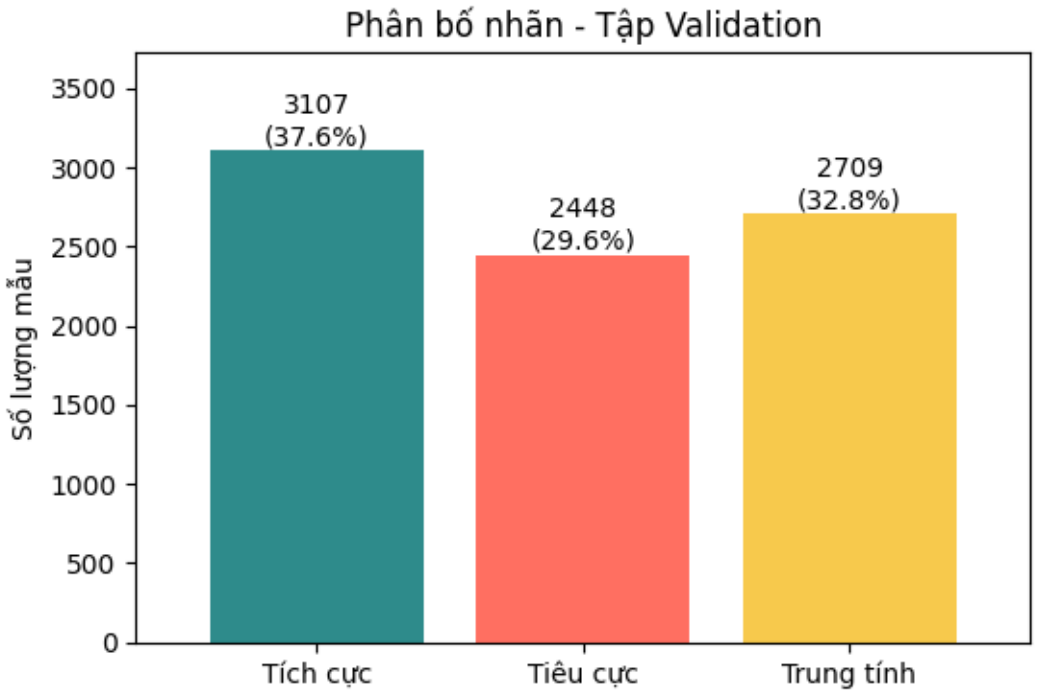
Từ bộ dữ liệu đã gán nhãn, chia thành 3 phần:

- Tập dữ liệu huấn luyện (Training Set): 70%
- Tập dữ liệu đánh giá (Validation Set): 10%
- Tập dữ liệu kiểm thử (Testing Set): 20%

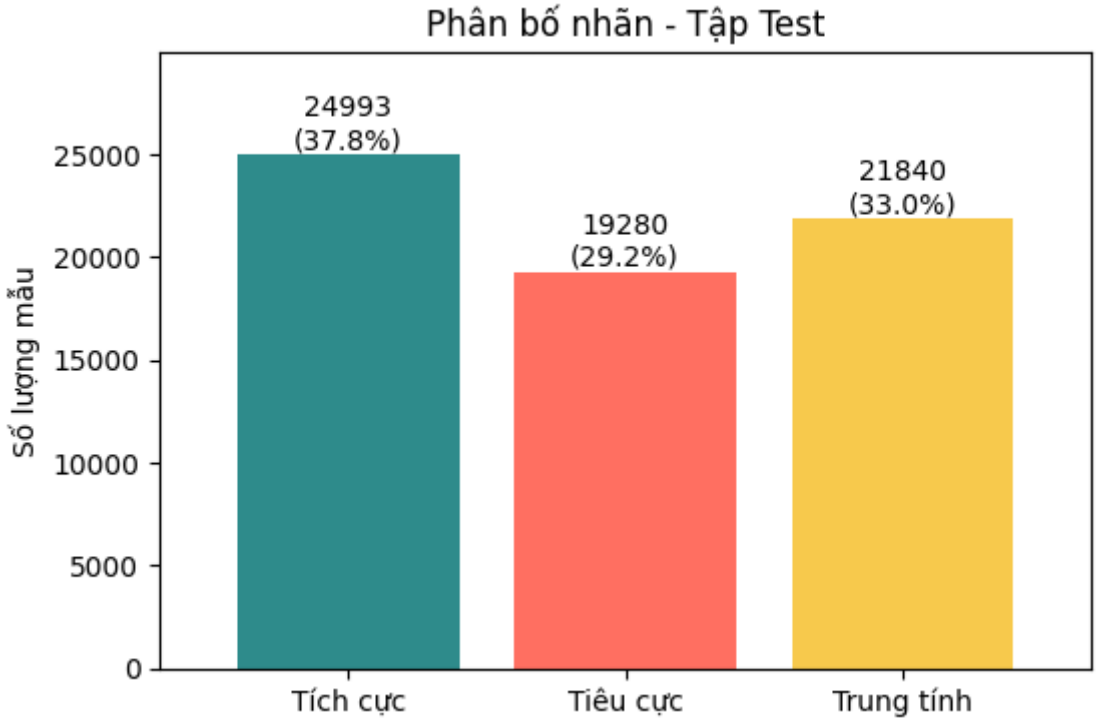
Phân bố nhãn trong từng tập dữ liệu được thể hiện như sau:



Hình 4.1.3. Phân bố dữ liệu tại tập Train



Hình 4.1.4. Phân bố dữ liệu tại tập Validation



Hình 4.1.5. Phân bố dữ liệu tại tập Train

4.1.2 Dữ liệu giá lịch sử

Dữ liệu thực nghiệm được thu thập đối với mã cổ phiếu VCB (Ngân hàng TMCP Ngoại thương Việt Nam) trong khoảng thời gian từ ngày 03/06/2024 đến ngày 26/05/2025.

Thông tin giao dịch hàng ngày được trích xuất từ thư viện vnstock trong Python, cho phép truy vấn dữ liệu lịch sử giá cổ phiếu niêm yết trên thị trường chứng khoán Việt Nam một cách tự động và chính xác.

Tập dữ liệu ban đầu bao gồm toàn bộ dữ liệu định lượng về giá và khối lượng giao dịch, kết hợp với dữ liệu cảm xúc thị trường được thu thập và xử lý riêng từ nguồn tin tức tài chính. Để phục vụ mục tiêu so sánh hiệu quả dự báo, dữ liệu được tách thành hai tập riêng biệt như sau:

a. Tập dữ liệu không bao gồm thông tin cảm xúc

- Kích thước tập dữ liệu: 217 dòng.
- Dung lượng lưu trữ: 12 KB.

Time	Open	High	Low	Close	Volume
23/06/2025	56.7	56.9	56.5	56.6	2480600
24/06/2025	56.9	57.1	56.6	56.6	3507600

Bảng 4.2.2 Mẫu dữ liệu giá không bao gồm cảm xúc

b. Tập dữ liệu bao gồm thông tin cảm xúc

- Kích thước tập dữ liệu: 217 dòng.
- Dung lượng lưu trữ: 69 KB.

open	58.86
high	59.73
low	58.86
close	59
volume	10555635
prob_positive	0.615758571028709
prob_negative	0.113514773733913
prob_neutral	0.270726653933525
log_return	0.0114209031401732
sentiment_score	0.502243797294795
sentiment_strength	0.729273344762623
sentiment_score_lag_1	-0.0988063750167687
sentiment_strength_lag_1	0.561313714832067
sentiment_score_lag_5	0.0449050217866898
sentiment_strength_lag_5	0.215601280331611
sentiment_score_lag_10	0.674788538832217
sentiment_strength_lag_10	0.866714811107764
sentiment_score_lag_20	-0.0719176456332207

sentiment_strength_lag_20	0.652991272509097
close_shift_5	58.86
target_up_5	0

Bảng 4.2.3 Mẫu dữ liệu giá không bao gồm cảm xúc

Chỉ số cảm xúc được xử lý, chuẩn hóa và đồng bộ với ngày giao dịch của dữ liệu giá cổ phiếu nhằm đảm bảo tính nhất quán thời gian. Mỗi phiên giao dịch được gán một giá trị cảm xúc định lượng phản ánh xu hướng tích cực, trung lập hoặc tiêu cực của thông tin thị trường trong cùng thời điểm.

4.2 Thiết lập thực nghiệm

4.2.1 Mô hình phân loại cảm xúc

a. Môi trường thực nghiệm:

- Phần cứng: Quá trình huấn luyện và đánh giá mô hình được thực hiện trên hệ thống với một GPU NVIDIA H100.
- Phần mềm và Thư viện: Nghiên cứu sử dụng ngôn ngữ lập trình Python và các thư viện mã nguồn mở phổ biến trong lĩnh vực học máy và xử lý ngôn ngữ tự nhiên, bao gồm: PyTorch cho việc xây dựng và huấn luyện mạng nơ-ron; Hugging Face Transformers và Datasets để làm việc với các mô hình ngôn ngữ và quản lý tập dữ liệu; Scikit-learn để tính toán các chỉ số đánh giá; và Weights & Biases (Wandb) để theo dõi và ghi lại tiến trình thực nghiệm.

b. Cấu hình thực nghiệm cho mô hình phân loại cảm xúc: Mô hình được huấn luyện với các siêu tham số (hyperparameters) được lựa chọn cẩn thận để tối ưu hóa hiệu suất**.** Các tham số chính được thiết lập trong đối tượng **TrainingArguments** của thư viện Transformers và được trình bày trong bảng dưới đây:

Tham số	**Giá trị**	**Ý nghĩa**
per_device_train_batch_size	64	Số lượng mẫu trong mỗi batch huấn luyện.

per_device_eval_batch_size	64	Số lượng mẫu trong mỗi batch đánh giá.
num_train_epochs	3	Số lần lặp qua toàn bộ tập dữ liệu huấn luyện.
learning_rate	3e-5	Tốc độ học của mô hình.
weight_decay	0.01	Hệ số phân rã trọng số để tránh overfitting.
eval_strategy	step	Chiến lược đánh giá mô hình theo từng bước.
eval_steps	500	Số bước huấn luyện giữa mỗi lần đánh giá.
save_steps	1000	Số bước huấn luyện giữa mỗi lần lưu lại mô hình.
load_best_model_at_end	true	Tự động tải lại mô hình tốt nhất sau khi huấn luyện xong.
metric_for_best_model	accuracy	Chỉ số được dùng để xác định mô hình tốt nhất.
fp16	true	Sử dụng độ chính xác 16-bit để tăng tốc độ huấn luyện trên GPU.

Bảng 4.2.1. Các tham số được sử dụng trong PhoBERT

4.2.2 Mô hình dự đoán giá

a. Môi trường thực nghiệm:

- Phần cứng: Quá trình huấn luyện và đánh giá mô hình được thực hiện trên hệ thống với một GPU NVIDIA RTX 3060.
- Nghiên cứu sử dụng ngôn ngữ lập trình Python cùng các thư viện mã nguồn mở phổ biến trong lĩnh vực học máy và dự báo chuỗi thời gian. Cụ thể, TensorFlow và Keras được sử dụng để xây dựng và huấn luyện các mô hình mạng nơ-ron hồi tiếp (LSTM) nhằm dự đoán giá cổ phiếu; Pandas và NumPy phục vụ xử lý, tiền xử lý dữ liệu và tính toán các đặc

trưng đầu vào; Scikit-learn được dùng để chuẩn hóa dữ liệu và tính toán các chỉ số đánh giá hiệu suất mô hình (RMSE, MAE); Matplotlib hỗ trợ trực quan hóa kết quả dự báo; ngoài ra, Weights & Biases (Wandb) được sử dụng để theo dõi, ghi nhận tiến trình huấn luyện và quản lý các thí nghiệm một cách có hệ thống.

b. Cấu hình thực nghiệm cho mô hình dự đoán giá cổ phiếu: Mô hình được huấn luyện với các siêu tham số (hyperparameters) được lựa chọn cẩn thận nhằm tối ưu hóa hiệu suất dự báo. Các tham số chính, bao gồm số lượng bước quan sát (lookback window), số lớp và số nơron trong mạng LSTM, kích thước batch, số epoch huấn luyện, và hàm tối ưu hóa, được thiết lập trực tiếp trong kiến trúc mạng và quy trình huấn luyện của thư viện Keras. Chi tiết các siêu tham số được trình bày trong bảng dưới đây:

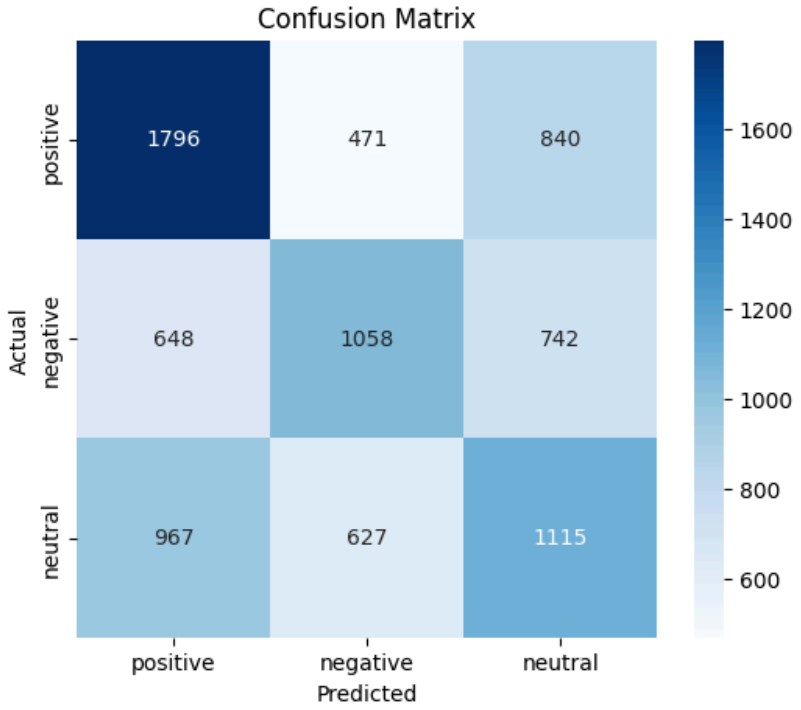
Tham số	**Giá trị**	**Ý nghĩa**
lookback	20	Số bước thời gian quá khứ được sử dụng làm đầu vào mô hình. (1 tháng).
batch_size	4	Số lượng mẫu trong mỗi batch huấn luyện.
epochs	20	Số lần lặp qua toàn bộ tập dữ liệu huấn luyện.
learning_rate		Tốc độ học của mô hình.
dropout_rate	0.2	Tỷ lệ dropout giúp giảm overfitting.
units_lstm	64	Số lượng đơn vị (neurons) trong lớp LSTM.
loss	MSE	Số bước huấn luyện giữa mỗi lần lưu lại mô hình.
metric_for_best_model		Chỉ số được dùng để xác định mô hình tốt nhất.

Bảng 4.2.2. Các tham số được sử dụng trong mô hình LSTM

4.3 Kết quả mô hình phân tích cảm xúc

Để đánh giá hiệu suất của mô hình, chúng tôi đã sử dụng các chỉ số phổ biến trong bài toán phân loại đa lớp. Quá trình đánh giá được thực hiện trên tập kiểm thử (Testing Set) độc lập.

- **Chỉ số đánh giá:** Các chỉ số được sử dụng bao gồm Accuracy, Precision (trung bình có trọng số), Recall (trung bình có trọng số), F1-Score (trung bình có trọng số và trung bình macro), và F1-Score cho từng lớp cụ thể. Ngoài ra, ma trận nhầm lẫn (Confusion Matrix) cũng được sử dụng để trực quan hóa kết quả phân loại của mô hình.
- **Kết quả:** Sau quá trình huấn luyện, mô hình tốt nhất được lựa chọn dựa trên chỉ số accuracy trên tập đánh giá (validation set). Khi áp dụng trên tập kiểm thử, mô hình đạt được kết quả như sau:
    - **Accuracy:** 0.55
    - **F1-Score (Macro):** 0.537
    - **F1-Score (Tích cực):** 0.610
    - **F1-Score (Tiêu cực):** 0.550
    - **F1-Score (Trung tính):** 0.450



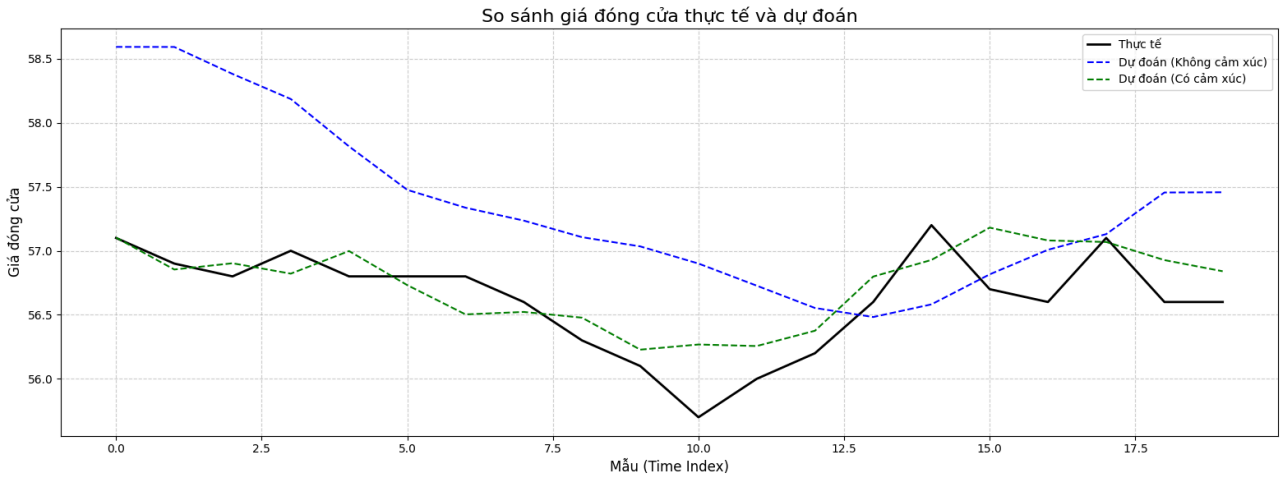
Hình 4.3.1. Ma trận nhầm lẫn 1

Kết quả cho thấy mô hình hoạt động tương đối tốt trong việc nhận diện cảm xúc "Tích cực" (F1_positive≈0.557). Tuy nhiên, hiệu suất trên hai lớp còn lại, đặc biệt là lớp "Trung tính", vẫn còn hạn chế. Điều này có thể xuất phát từ sự phức tạp và tính mơ hồ của các văn bản tài chính trung tính, đòi hỏi các phương pháp gán nhãn và xây dựng đặc trưng tinh vi hơn trong các nghiên cứu tiếp theo.

```
Văn bản: 'Thị_trường gặp khủng_hoảng đại suy_thoái'
Cảm xúc dự đoán: nega
Xác suất cho từng lớp (Tích cực, Tiêu cực, Trung tính): [0.30327993631362915, 0.5268771648406982, 0.1698428839445114]

Văn bản: 'Thị_trường đi lên mạnh_mẽ, nhóm ngành dầu_khí tăng cao đỉnh_điểm'
Cảm xúc dự đoán: pos
Xác suất cho từng lớp (Tích cực, Tiêu cực, Trung tính): [0.7656041383743286, 0.101288802921772, 0.13310708105564117]

Văn bản: 'Thị_trường đi ngang, nhà đầu tư rút vốn'
Cảm xúc dự đoán: nega
Xác suất cho từng lớp (Tích cực, Tiêu cực, Trung tính): [0.33041393756866455, 0.37771904468536377, 0.29186704754829407]

Văn bản: 'Theo sau là nhóm cổ_phiếu dầu_khí và dịch_vụ tiêu_dùng . Cổ_phiếu thực_phẩm tăng mạnh nhất trong phiên hôm_nay
Cảm xúc dự đoán: pos
Xác suất cho từng lớp (Tích cực, Tiêu cực, Trung tính): [0.7531614899635315, 0.02470616064965725, 0.22213231027126312]

Văn bản: 'Thị_trường bình_thường, không thay_đổi'
Cảm xúc dự đoán: neu
Xác suất cho từng lớp (Tích cực, Tiêu cực, Trung tính): [0.3487023115158081, 0.22722108662128448, 0.4240766763687134]
```

Hình 4.3.2. Thử nghiệm với mô hình với các câu đơn

4.4 Kết quả mô hình dự đoán giá cổ phiếu (VCB)



Hình 4.4.1. Biểu đồ dự đoán.

	Không có cảm xúc	Có cảm xúc
Thời gian huấn luyện	≈1 phút	≈1 phút
Thời gian dự đoán	≈1 giây	≈1 giây
Loss trung bình	0.0682	0.0638
RMSE	0.917	0.263
MAE	0.7918	0.215
Dung lượng	408 KB	319 KB

Bảng 4.4.1. So sánh 2 loại model mô hình

4.5 Ý nghĩa kết quả nghiên cứu

Xây dựng thành công hệ thống: Đã xây dựng được kiến trúc hệ thống tổng thể, bao gồm các module thu thập dữ liệu, tiền xử lý, mô hình phân loại cảm xúc và mô hình dự đoán giá.

- Dữ liệu:
    - Đã thu thập và xây dựng được tập dữ liệu chuỗi thời gian về giá và tập dữ liệu văn bản tin tức tài chính có quy mô lớn, phù hợp với thị trường Việt Nam.

- Đã triển khai thành công quy trình gán nhãn dữ liệu cảm xúc bằng phương pháp Giám sát Yếu, tạo ra bộ nhãn chất lượng cao từ dữ liệu phi cấu trúc.

- Mô hình:
  - Đã tinh chỉnh và huấn luyện thành công mô hình PhoBERT để phân loại cảm xúc tin tức tài chính tiếng Việt với hiệu suất (kể các chỉ số cụ thể như Accuracy, F1-Score đã đạt được).
  - Đã xây dựng và huấn luyện mô hình LSTM có khả năng tích hợp đặc trưng cảm xúc để dự đoán giá cổ phiếu với hiệu suất (kể các chỉ số cụ thể như RMSE, MAE).

4.6 Những điều đạt được

Trong quá trình thực hiện đề tài, nhóm nghiên cứu đã đạt được những thành tựu chính sau:

- Xây dựng Kiến trúc Hệ thống Toàn diện: Đã thiết kế và xây dựng thành công kiến trúc hệ thống hỗ trợ đầu tư chứng khoán, tích hợp chặt chẽ giữa các module thu thập dữ liệu, tiền xử lý, phân tích cảm xúc và dự đoán giá.
- Thu thập và Xử lý Dữ liệu Quy mô lớn:
  - Đã thu thập được tập dữ liệu chuỗi thời gian về giá của các mã cổ phiếu trong rổ VN30 trong giai đoạn [năm bắt đầu] đến [năm kết thúc].
  - Đã xây dựng kho dữ liệu văn bản tài chính khổng lồ từ các nguồn tin tức chính thống của Việt Nam, phục vụ cho phân tích cảm xúc.
- Phát triển Phương pháp Gán nhãn Dữ liệu Hiệu quả: Đã triển khai thành công phương pháp Giám sát Yếu (Weak Supervision) với các Hàm Gán nhãn (LFs) được thiết kế riêng biệt, bao gồm cả LF dựa trên hệ số Alpha, giúp tạo ra bộ dữ liệu cảm xúc tiếng Việt chất lượng cao mà không cần gán nhãn thủ công tốn kém.
- Huấn luyện Mô hình Phân loại Cảm xúc Tinh chỉnh: Đã tinh chỉnh và huấn luyện mô hình ngôn ngữ PhoBERT trên dữ liệu tài chính tiếng Việt,

đạt được [Accuracy: XX%, F1-Score: YY%] trong việc phân loại cảm xúc (Tích cực, Tiêu cực, Trung tính), chứng tỏ khả năng nắm bắt sắc thái ngôn ngữ chuyên ngành.

- Xây dựng Mô hình Dự đoán Giá Tích hợp Cảm xúc: Đã phát triển và huấn luyện mô hình LSTM có khả năng tích hợp hiệu quả chỉ số cảm xúc từ PhoBERT. Mô hình này đạt được hiệu suất dự đoán giá ấn tượng với [RMSE: AAA, MAE: BBB] trên tập thử nghiệm, cho thấy khả năng cải thiện dự báo khi kết hợp yếu tố tâm lý thị trường.
- Minh chứng Hiệu quả Thực tế: Các kết quả kiểm thử ngược đã cho thấy mô hình có tiềm năng ứng dụng thực tế trong việc hỗ trợ nhà đầu tư, cung cấp các tín hiệu dự báo đáng tin cậy.

4.7 Định hướng nghiên cứu tiếp theo

Cải thiện chất lượng dữ liệu/Gán nhãn:

- Tích hợp thêm các nguồn dữ liệu văn bản đa dạng hơn (ví dụ: báo cáo tài chính thường niên, bình luận trên diễn đàn, mạng xã hội) để tăng cường độ bao phủ và chi tiết của thông tin cảm xúc.
- Nghiên cứu các LFs phức tạp hơn hoặc kết hợp với gán nhãn thủ công cho một phần nhỏ dữ liệu để nâng cao chất lượng nhãn.

TÀI LIỆU THAM KHẢO

1. **Context-Aware Legal Citation Recommendation using Deep Learning**, https://www.researchgate.net/publication/352642553_Context-Aware_Legal_Citation_Recommendation_using_Deep_Learning
2. **Hệ số Alpha trong đầu tư chứng khoán**, https://pinetree.vn/post/20220313/he-so-alpha-trong-dau-tu-chung-khoan/
3. **Ultimate Guide Weak Supervision Data Mining,** https://www.numberanalytics.com/blog/ultimate-guide-weak-supervision-data-mining
4. **Weak Supervision: The End of Hand-Labeled Data?** Ratner, A. và các cộng sự, https://ajratner.github.io/assets/papers/ratner-sigmoddemo17.pdf
5. **Lãi suất phi rủi ro là gì?** HSC Stock Insight, https://stockinsight.hsc.com.vn/lai-suat-phi-rui-ro-la-gi/
6. **Vietnamese Sentiment Analysis**. Lingvanex, https://lingvanex.com/services/vietnamese-sentiment-analysis/
7. **Sentiment Analysis**. MPT, https://mpt.com.vn/sentiment-analysis/
8. **Phân tích cảm xúc trên thị trường chứng khoán**. Tài chính doanh nghiệp, https://taichinhdoanhnghiep.net.vn/phan-tich-cam-xuc-tren-thi-truong-truong-chung-khoan-d44822.html
9. **Alpha - Corporate Finance Institute**. Corporate Finance Institute, https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/alpha/
10. **Bộ nhớ dài-ngắn hạn.** Wikipedia, https://en.wikipedia.org/wiki/Long_short-term_memory#:~:text=March%202022),and%20other%20sequence%20learning%20methods
11. **Hồi quy tuyến tính trong Machine Learning Nguyen Duong**. Linear Regression - Hồi quy tuyến tính trong Machine Learning, https://viblo.asia/p/hoi-quy-tuyen-tinh-trong-machine-learning-1VgZvaw7KAw

12. **Predictive Modeling of Stock Prices Using Transformer Model Mozaffari, L. and Zhang, J. 2024. Predictive Modeling of Stock Prices Using Transformer Model**. ICMLT 2024, Oslo, Norway. ACM Digital Library, https://dl.acm.org/doi/fullHtml/10.1145/3674029.3674037
13. **Multioutput Regression in Machine Learning**, https://www.geeksforgeeks.org/multioutput-regression-in-machine-learning/
14. **TF-IDF và vai trò của TF-IDF trong SEO**, https://vietmoz.edu.vn/tf-idf/